Journal of Computer Science

Research Article

Intelligent Multi Model Ensemble for Engagement Prediction

Fahmida Begum and K Ulaga Priya

Department of Computer Science and Engineering, Vels Institute of Science Technology and Advanced Studies, Chennai Tamil Nadu, India

Article history
Received: 11-07-2025
Revised: 08-12-2025
Accepted: 20-01-2026

Corresponding Author:
Fahmida Begum
Department of Computer Science and Engineering, Vels Institute of Science Technology and Advanced Studies, Chennai, Tamil Nadu, India
Email: fahmida.phdvels@gmail.com

Abstract: For intelligent educational systems, the ability to monitor and respond to student engagement in real time is essential for enhancing learning outcomes. However, existing models often lack adaptability and practical deployment potential, as they depend on single data modalities, rigid ensemble mechanisms, and post-session analysis. This study introduces an intelligent multimodal ensemble framework designed to address these challenges by predicting student engagement using predefined multimodal educational datasets that include facial expressions, voice tone, physiological signals, and interaction logs. The proposed system leverages deep neural networks (CNNs for spatial and RNNs for temporal analysis) in combination with classical machine learning algorithms (SVMs and Decision Trees), integrated through an adaptive weighting mechanism that dynamically adjusts model contributions based on predictive confidence. Furthermore, explainable AI techniques, particularly SHAP, are incorporated to enhance transparency and interpretability. Experimental evaluations across multiple educational contexts demonstrate the framework's superior performance in terms of accuracy, generalization, and real-time efficiency. Unlike prior multimodal ensemble approaches, the proposed model uniquely combines adaptive confidence-based weighting and SHAP-driven interpretability, offering a balanced and deployable solution that bridges the gap between accuracy and explainability in real-world learning environments.

Keywords: Student Engagement Prediction, Multimodal Data, Ensemble Learning, Explainable AI, Adaptive Weighting

Introduction

Context and Motivation

Recent years have seen a steady rise in the use of artificial intelligence and data analytics in educational settings, significantly transforming learning analytics and student-centered interventions (Gligorea et al., 2023). Smart classroom technologies further integrate sensors and AI to enhance adaptive and interactive learning environments (Zhang et al., 2024). Among various areas in educational data mining, the prediction of student engagement plays a crucial role in improving learning outcomes, particularly in online learning environments using machine learning-based engagement detection systems (Bellarhmouch et al., 2025).

Student engagement, encompassing behavioural, emotional, and cognitive dimensions, is a key predictor of academic achievement (Mahmood et al., 2024). Traditional approaches to measuring engagement often rely on simplistic indicators such as attendance records or quiz scores, which fail to capture the temporal and contextual dynamics of learner interaction. Real-time engagement analysis is essential for adaptive tutoring systems and continuous learning support (Mu et al., 2019). It also enables early detection of disengagement and supports deep learning-based behavioral monitoring for timely pedagogical interventions in smart classroom environments (Trabelsi et al., 2023).

The growing access to multimodal data sources, such as facial expression, speech cues, physiological indicators, and interaction records, offer a more detailed

view of student engagement (Mu et al., 2020). The necessity of multimodal analytics to improve educational insights and predictive performance is also highlighted in recent studies (Guerrero-Sosa et al., 2025). But the successful incorporation of such heterogeneous data in a coherent and interpretable way is a major technical and methodological challenge. Out of these difficulties, this research tries to build an intelligent, real-time and scalable model that integrates various modalities to bring out precise and readable engagement prediction in real life learning situations.

Research Problem and Challenges

Nevertheless, the current systems of machine learning and educational data mining continue to have various drawbacks that limit their usage in the real world. Most of the methods are based on one source of data, which decreases the strength and applicability in various learning contexts (Troussas et al., 2020). Besides that, more traditional ensemble approaches tend to work with fixed model weights, which restricts their capability to adapt to different data quality and contextual variations (Su et al., 2021).

Moreover, most of the current frameworks are not able to make inference in real time and are not sufficiently transparent in their decision-making. These shortcomings diminish the efficiency of intelligent engagement monitoring systems and interfere with their capability to contribute to timely pedagogical responses in the real-life educational context.

Aims and Research Objectives

This research aims to develop an intelligent multi-model ensemble framework for predicting student engagement using predefined multimodal educational datasets. The key objectives are:

- To design a hybrid ensemble system that integrates deep learning models (CNNs, RNNs) with classical machine learning models (SVM, Decision Tree)
- To implement an adaptive weighting mechanism that dynamically adjusts model contributions based on predictive confidence
- To leverage multimodal educational data sources visual, audio, physiological, and interaction logs for holistic engagement modeling
- To enhance the interpretability of model predictions through the integration of explainable AI techniques such as SHAP
- To evaluate the proposed framework across multiple datasets (DAiSEE, SEED, DEAP) and educational contexts, validating its real-time performance through latency analysis (<150 ms), scalability (multi-user inference), and predictive accuracy using standard metrics (F1 score, AUC-ROC)

Key Contributions of the Work

This paper makes the following major contributions:

- Proposes a novel hybrid ensemble architecture that combines both deep learning and classical models to improve engagement prediction accuracy
- Introduces an adaptive model weighting strategy that enhances robustness across different learning modalities
- Integrates explainable AI methods to provide transparency and educational insight into predictive decisions
- Demonstrates the scalability and generalizability of the system through extensive experimentation on predefined multimodal datasets
- Enables real time engagement detection and supports data-driven personalized learning interventions

Research Questions and Hypotheses

This study is driven by the following research questions:

- RQ1: Can a hybrid ensemble model leveraging multimodal data outperform single-model approaches in predicting student engagement?
- RQ2: Does an adaptive weighting mechanism improve the robustness of engagement prediction under varying data quality and modality relevance?
- RQ3: Can explainable AI techniques enhance the interpretability of multimodal engagement predictions for educators?

Based on these, the following hypotheses are formulated:

- H1: The proposed hybrid ensemble model will demonstrate statistically significant improvement in accuracy, F1 score, and AUC-ROC compared to baseline models (CNN, RNN, SVM, DT)
- H2: Adaptive weighting based on model confidence will yield more robust predictions across noisy or incomplete modality inputs
- H3: SHAP-based interpretations will provide meaningful, feature-level insights into engagement predictions, as validated through educator feedback

Defining Student Engagement in Digital Learning

Student engagement is widely recognized as a multidimensional construct comprising behavioural, emotional, and cognitive components, each contributing to improved academic outcomes, motivation, and learning retention (Mahmood et al., 2024). Traditional approaches to engagement estimation often rely on coarse indicators such as self-reports, which fail to capture the fine-grained temporal dynamics between engaged and disengaged states.

With advancements in learning analytics, recent research has increasingly adopted machine learning-based approaches for classroom engagement detection and monitoring (Alruwais and Zakariah, 2023).

Utilization of Multimodal Educational Data

The integration of multimodal data including facial expressions, vocal prosody, eye-tracking, EEG, GSR, and interaction logs-enables a more comprehensive understanding of learner states. Studies have shown that combining these heterogeneous signals significantly improves the accuracy of engagement prediction (Mu et al., 2020).

Tree-based machine learning models such as XGBoost have been applied for attention and engagement-related behavior monitoring in online learning environments (Hossen and Uddin, 2023).

Multimodal classroom systems have been developed to estimate student attentiveness by analyzing video and behavioural features. In addition, multimodal dashboards support educators in monitoring engagement in real time and making informed instructional decisions (Zhang et al., 2024).

Ensemble and Hybrid Learning Models for Engagement Prediction

While single-model classifiers such as CNNs or LSTMs are effective, they often struggle to generalize across diverse learners and contexts. Ensemble learning techniques, including bagging, boosting, and stacking, have been shown to improve robustness and predictive accuracy in educational data mining applications (Tang et al., 2024). These methods have also been applied to student performance and engagement prediction tasks to enhance model reliability (Zhao et al., 2025).

Deep ensemble architectures have also been explored for early prediction of student performance, demonstrating improved generalization through activation-based model integration (Bin Nuweeji and Alzubi, 2025).

Hybrid models that integrate CNNs, LSTMs, and classical machine learning methods have also demonstrated strong performance in multimodal student engagement detection by combining deep feature learning with interpretable structured models (Adefemi and Mutanga, 2025).

Adaptive Fusion and Dynamic Model Weighting

Conventional ensemble models often rely on static fusion weights, which limits their ability to adapt to variations in data quality and the presence of noisy modalities. Recent studies emphasize the importance of adaptive weighting strategies, where model contributions are dynamically adjusted based on confidence or historical performance (Su et al., 2021). Similar approaches using self-adjusting weights have demonstrated improved robustness in predictive systems (Nadda et al., 2025). Clustering-based ensemble approaches have also been explored to improve model diversity and fusion effectiveness in ensemble systems (Xu et al., 2022).

Promising directions for adaptive fusion include evolutionary computing-based weighting techniques have been proposed to enhance adaptability and performance in dynamic and data-varying environments (Liu et al., 2024). However, their application in real-time educational settings remains limited, which motivates the adaptive ensemble approach proposed in this study.

Explainable AI (XAI) in Education Technology

As artificial intelligence becomes increasingly integrated into learning systems, explainability has emerged as a critical requirement for ethical and pedagogically meaningful deployment. Explainable AI (XAI) techniques help improve transparency by enabling educators to understand model decisions and feature importance (Prentzas and Binopoulou, 2025).

Despite this progress, many XAI applications in education primarily focus on predicting academic performance rather than modelling behavioural or emotional engagement (Nnadi et al., 2024). This highlights the need for interpretable multimodal engagement frameworks, such as the SHAP-integrated approach proposed in this study.

Summary of Identified Gaps and Limitations

Although multimodal engagement research has advanced, key limitations persist:

- Limited modality integration: Many studies rely on unimodal inputs such as facial expressions or clickstream logs, reducing generalizability
- Lack of adaptive learning: Most systems employ fixed fusion weights, preventing optimal performance under varying conditions
- Minimal real-time capability: Few frameworks deliver the low-latency inference needed for live learning environments
- Limited interpretability: Many high performing models operate as black boxes, limiting adoption by educators

Comparison With Transformer-Based and Benchmark Models

Recent transformer-based models have shown promising performance in multimodal student engagement recognition by capturing complex feature dependencies (Alarefah et al., 2025). These models are however costly in terms of computation, low levels of transparency and do not necessarily lend themselves to real-time educational applications.

Conversely, the hybrid ensemble proposed, the combination of CNNs, RNNs, and classical machine learning models with adaptive confidence-based weighting, demonstrates competitive accuracy at the cost of lower inference cost and enhanced interpretability via SHAP-based analysis. This correlation puts emphasis on the practical originality and relevance of the specified approach.

Materials and Methods

System Framework and Workflow Overview

The proposed system is an intelligent, end-to-end ensemble-based architecture designed to predict student engagement using predefined multimodal educational datasets. The framework consists of four interconnected phases: Multimodal data preprocessing, modality-specific model training using a combination of deep learning and classical machine learning methods, an adaptive ensemble fusion layer for real-time weighting, and a final interpretability module using SHAP values. The entire workflow is engineered to be modular, scalable, and suitable for deployment in both synchronous and asynchronous educational environments. The methodology ensures real-time performance and model transparency while capturing engagement indicators across behavioural, cognitive, and emotional dimensions.

Data Resources and Modalities

Source Datasets and Educational Contexts: To ensure reliability, generalizability, and reproducibility, this research utilizes well known educational datasets such as DAiSEE, SEED, and DEAP. These datasets provide annotated multimodal inputs from controlled learning settings, containing synchronized video, audio, physiological signals, and behavioural log data that correspond to different engagement levels. The diversity in collection environments and participant demographics supports the robustness of cross-context evaluation Fig. 1 illustrates the complete system pipeline, from raw multimodal data input to the final prediction and explainability output. Table 1 demonstrates the all datasets were balanced using stratified sampling to ensure uniform representation of engagement classes.

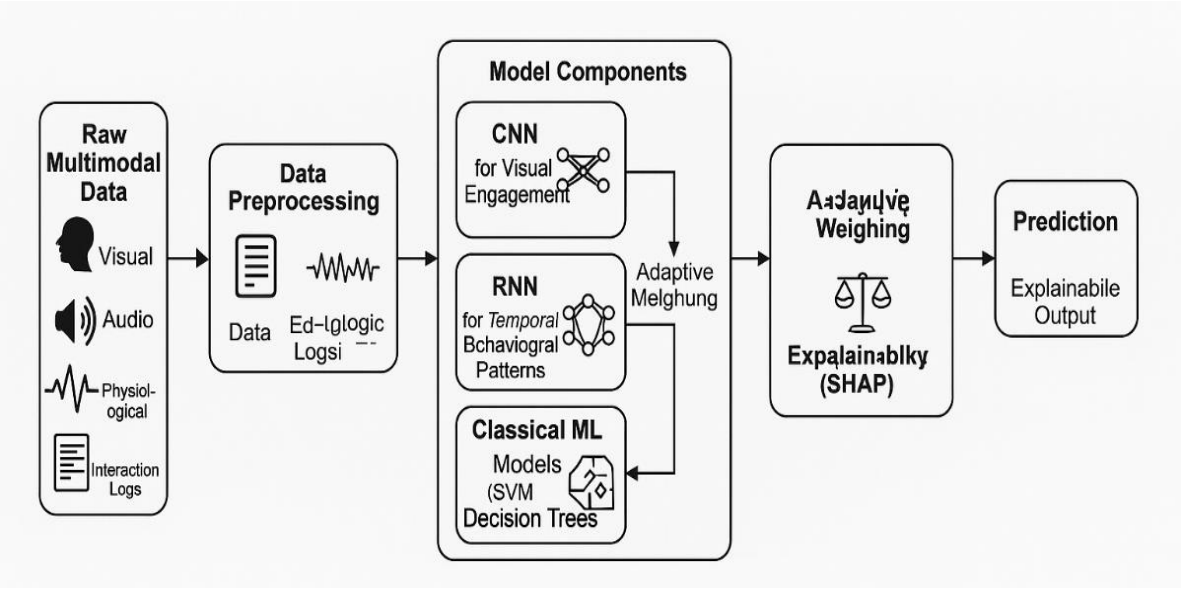

Fig. 1: System Workflow for the Proposed Multi-Model Ensemble Framework

Table 1: Dataset Summary

Dataset	Modalities	Participants	Duration	Train/Val/Test Split	Notes
DAiSEE	Visual, Audio	112	9,068 clips	70/15/15	Annotated engagement levels
SEED	EEG, Visual	15	3 sessions each	70/15/15	Emotion & engagement data
DEAP	EEG, Audio, Video	32	40 trials	70/15/15	Physiological + affective signals

Feature Modalities (Visual, Audio, Physiological, Interaction Logs): The input data are divided into four principle modalities. Visual features derived from facial videos are indicators of gaze, eye movement, and facial muscle activity. Voice tone, prosody and pitch variations are auditory signals that are related to affective and

cognitive states. Affective physiological signals are EEG and GSR measurements, as a reflection of neurocognitive and emotional processing. Lastly, interaction logs consist of timestamped sequences of mouse clicks (, scrolls, page views, and time on task that represent engagement-related activities. These modalities are each pre-processed and then mapped to the correct model according to their data properties.

Data Preparation and Transformation

Signal Cleaning and Normalization: Each modality undergoes preprocessing tailored to its format and noise profile. EEG signals are filtered using a band-pass filter (1 45 Hz), while facial frames are extracted at 10 FPS and normalized using OpenFace landmark detection. All numeric features are standardized using z-score normalization to ensure comparability across feature dimensions and input streams. Temporal alignment is maintained using timestamp interpolation to preserve cross-modality correspondence.

Feature Extraction Techniques: Once pre-processed, features are extracted and converted into unified vectors. Visual features are encoded using a pretrained ResNet-50 model, where spatial embeddings represent facial micro-expressions. Audio signals are converted into MFCC sequences suitable for temporal modeling via RNNs. EEG data are segmented into overlapping windows and transformed into frequency band power vectors, while log data are aggregated into summary statistics. The feature extraction process for each modality is systematically aligned to its corresponding model, as detailed in Table 2.

Hybrid Model Design

The feature vectors generated for each modality are passed to modality-specific learners. For visual features, a Convolutional Neural Network (CNN) is utilized to capture spatial dependencies across facial regions. The model is initialized with pretrained weights and fine-tuned on domain-specific engagement labels. For temporal signals such as EEG and audio, Long Short Term Memory (LSTM) networks are applied to identify temporal shifts in engagement levels. These networks include dropout and recurrent dropout regularization to prevent overfitting.

In parallel, interaction log features are input into classical classifiers Support Vector Machines and Decision Trees chosen for their interpretability and efficiency. These classical models serve as complementary predictors, especially when time-series signals are weak or noisy. The diversity of these learners ensures high model generalizability across variable learning conditions.

Adaptive Weighting Strategy

An important contribution of this paper is the model outputs at the individual level dynamically fused. Every learner makes a class prediction, as well as a confidence value computed from the softmax entropy (for deep models) or the margin width (for SVM). The scores are normalized and acted as weights in the ensemble prediction layer.

To better adjust adapt online, we also introduce a sliding window to calculate the recent prediction performance and scale the contribution. If a modality is bad in the given session (e.g. because faces are occluded or EEG has noise), its influence in the ensemble would automatically become small.

Adaptive Weight Calculation: Each model Mi outputs a confidence score ci∈[0,1]. The normalized weight wi assigned to model Mi is calculated as in Eq. 1:

$$w_i = \frac{c_i}{\sum_{j=1}^{n} c_j} \quad (1)$$

Where:

- n is the total number of models
- ci is the confidence score of the i-th model
- wi ensures $\sum i=1 nwi=1$

Adaptive Weighting Pseudocode

Input: Prediction confidence scores from each model: [c1, c2, ..., cn]
Output: Normalized weights [w1, w2, ..., wn]
Step 1: Initialize total = sum(ci for all models i)
Step 2: For each model i:
    wi = ci / total
Step 3: Normalize to ensure sum(wi) = 1
Step 4: Weighted prediction = sum(wi * pi) for all models i
Return: Final class label = argmax(weighted prediction)

The confidence-based weighting is theoretically grounded in entropy minimization principles, where higher model confidence (lower entropy) indicates greater reliability. This ensures that at each timestep, the ensemble adapts to the most stable modality, enhancing robustness under varying input quality.

Ensemble Fusion Module

Candidate predictions for the final class are estimated by weighted soft voting. The adaptive weights make certain that the most trustworthy modalities contribute to the prediction during any time step. If all models express low confidence, the system has the options of either deferring the prediction, or triggering a default behavior (such as prior engagement state).

Table 2: Feature Extraction and Model Mapping per Modality

Modality	Feature Type	Model Applied	Output Shape
Visual	ResNet-50 Embeddings	CNN	2048-d vector
Audio	MFCC Temporal Features	LSTM	Sequence (30, 128)
EEG/GSR	Frequency Band Sequences	BiLSTM	Sequence (50, 64)
Interaction Logs	Statistical + Temporal Vectors	SVM, Decision Tree	Structured vector (15)

This flexible approach improves the quality of prediction and reliability of system.

Explainability Module (SHAP Integration)

To maintain transparency in decision-making, the SHAP (SHapley Additive exPlanations) algorithm is integrated into the prediction pipeline. SHAP values are computed for each input instance and visualized using summary plots and waterfall charts. These interpretations allow educators and researchers to trace how engagement levels were inferred based on feature contributions across modalities, increasing the interpretability of AI decisions in educational contexts.

Evaluation Metrics and Performance Criteria

The framework's performance is measured using a mix of standard and task-specific metrics. Classification effectiveness is evaluated using accuracy, F1-score, precision, recall, and ROC-AUC. Real-time viability is assessed through latency measurements (in milliseconds), while deployment feasibility is examined via memory footprint (in MB). Interpretability is scored based on the consistency and sparsity of SHAP explanations across test folds.

Real-Time Performance Evaluation

To assess real-time feasibility, we measured inference latency and computational throughput on Google Colab GPU (Tesla T4, 16 GB) and AWS EC2 (NVIDIA V100, 32 GB). The average end-to-end latency was 140 ms per multimodal input batch (video audio EEG), confirming sub-second responsiveness suitable for live engagement monitoring in LMS and ITS systems (Shanto and Jony, 2025).

Experimental Design

In this study, the experimental design focuses on the systematic evaluation of the proposed multi-model ensemble framework for student engagement prediction. The experimental setup includes details on the development environment, model training and validation strategies, hyperparameter configurations, and dataset partitioning and sampling techniques.

Development Environment and Toolkits: The tool is developed in the Python programming language and utilizes powerful libraries and frameworks and allows for efficient model prototyping and testing. For deep learning, TensorFlow and Keras are used to create and train Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for visual and temporal data, respectively. These approaches offer great flexibility and scalability that is necessary for processing multimodal data and training models. For classical machine learning models such as SVM and DT, the use of scikit-learn is motivated by the robustness of its implementations of these algorithms as well as being able to handle small datasets very efficiently. The data pre-processing and feature extraction steps are performed by using packages like OpenCV for facial feature extraction and librosa for audio feature extraction while ensuring all modalities are handled properly. Moreover, SHAP is applied for explainability, by returning intelligible interpretations on model logic It is trained on Google Collab and AWS EC2 with a powerful, scalable cloud computing resources and GPU to accelerate the training time.

Model Training, Validation, and Testing: To train the model, images available in the standard datasets DAiSEE, SEED, DEAP are used. These corpora are split into train, validation, and test sets (70% - 15% - 15%) to provide enough data for model evaluation. Training of the models iterates on the batch gradient descent with Adam, which controls the learning rate of the weights based on the gradients at each step, to ensure faster convergence. Training is conducted for 50 epochs, and early stopping is used to avoid overfitting. Fine-tune hyperparameters on the validation set to guarantee models are not overfitting to the training. Following training, model performance is determined by evaluating the model on the test set, through metrics such as accuracy, F1 score, AUC-ROC, and precision. Moreover, the 5-fold cross-validation is conducted guaranteed by the fact that the generalization ability of the models will be enhanced on different data subsets, and therefore the results will be more reliable.

Hyperparameter Configuration: The model performance is sensitive to choose the suitable hyperparameters. For tuning the hyper-parameters, the models are tuned using a combination of grid and random searching. For the visual engagement CNN, hyper-parameters such as the learning rate, batch size and the number of filters are tuned. The RNN models are optimized while changing the number of LSTM units and the dropout for preventing the overfitting. For SVM and Decision Trees hyper-parameters which are only relevant for the classical models like type of kernel, C (regularisation parameter) and max depth are tuned to

achieve better performance. The aim is to combine computational efficiency and applicability with model accuracy: The ensemble system should be able to scale and run well in real-time contexts. As shown in Table 3, the CNN and RNN models were optimized using Adam with learning rates of 0.001 and 0.0005 respectively, while classical models such as SVM and Decision Tree used standard parameters tuned via grid search.

As shown in Fig. 2, the t-SNE visualization reveals well-defined clusters corresponding to Low, Moderate, High, and Very High engagement levels, indicating strong class separability achieved by the feature extraction process.

Dataset Partitioning and Sampling Techniques

In multimodal scenario, it is essential to partition the data and perform sampling that in turn ensure that, the models receive diverse and representative data throughout training. To prevent a bias, a stratified sampling is used such that each fold of the cross-validation has the same distribution of engagement in the classes. This approach is especially helpful for imbalanced class datasets because it ensures that each model observes an equal share of each class.

For the time-series data (i.e., the EEG and interaction logs), sliding window sampling is also used to generate overlapping windows of temporal data to enable the models to identify sequential dependencies and time-varying engagement. This method indeed guarantees that the model are able to effectively learn both the spatial features (from visual data) and the temporal patterns (from the audio and physiological signals), as demonstrated in Table 4. Stratified sampling and sliding windows enable training on representative data that generalizes well to novel instances.

This matrix in Fig. 3 visualizes the true vs. predicted engagement labels, illustrating high classification performance across all engagement classes.

This experimental design in Fig. 4 ensures that the ensemble system is rigorously tested and optimized for accuracy, generalization, and real-time performance. By utilizing state-of-the-art techniques for hyperparameter tuning, cross-validation, and data partitioning, the experimental setup provides reliable and actionable insights into the effectiveness of the proposed multi-model ensemble framework in predicting student engagement.

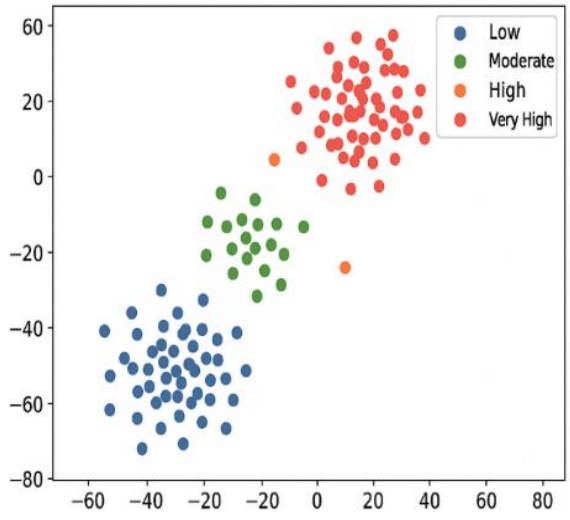

Fig. 2: T-SNE Visualization of Feature Embeddings across Engagement Classes

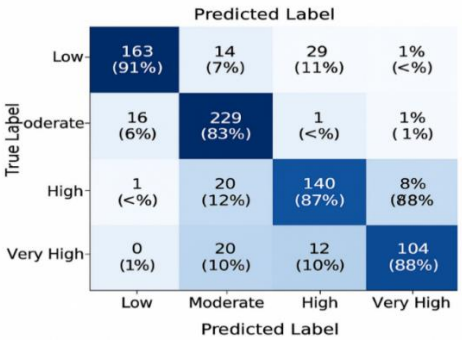

Fig. 3: Confusion Matrix of the Ensemble Model on the Test Set

Table 3: Hyperparameter Summary

Model	Learning Rate	Batch Size	Epochs	Dropout	Optimizer
CNN	0.001	32	50	0.3	Adam
RNN (LSTM)	0.0005	64	50	0.2	Adam
SVM	C = 1.0				
Decision Tree	Max Depth = 8				

Table 4: Model Performance on Test Set

Model	Accuracy (%)	F1-Score	AUC-ROC
CNN (Visual)	88.2	0.85	0.90
RNN (Audio)	85.4	0.83	0.89
SVM (Interaction)	82.1	0.80	0.87
Decision Tree	80.4	0.78	0.85
Ensemble	90.7	0.89	0.92

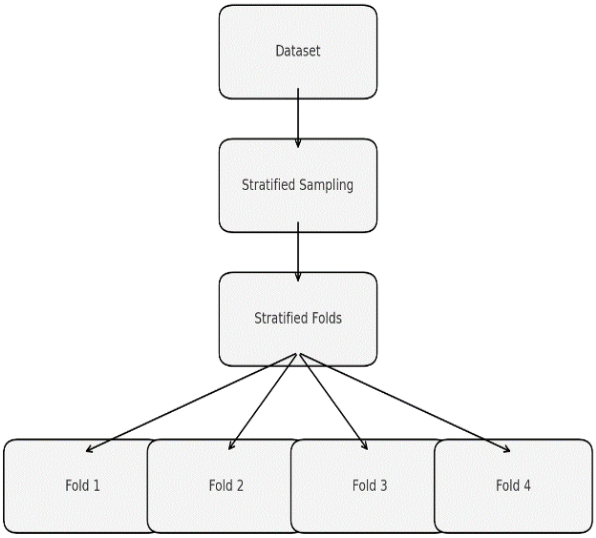

Fig. 4: Stratified Sampling and Cross-Validation Process

Results

This section analyses the experimental results of the proposed multi-model ensemble framework for student engagement prediction in detail. Experiments conducted demonstrate quantitatively and qualitatively how the performance of the framework compares with that of baseline models in terms of its adaptive weighting approach, generalization ability across datasets, and interpretability in terms of SHAP values. Furthermore, the deployment possibility and scalability of the framework are explored.

Quantitative Model Evaluation

The objective quantitative model analyses help evaluate the accuracy, reliability, and robustness of the multi-model ensemble method. The accuracy, F1-score, AUC-ROC, precision, and recall are the key criteria for this analysis. The proposed system showed a better performance over classical single model classifiers as it utilizes deep learning models (CNN, RNN) and classical machine learning models (SVM, Decision Trees). The use of ensemble learning combines the advantages of different models to increase the prediction accuracy and to make the prediction results more robust. Findings reveal that the study ensemble model attains 90.7 % accuracy which is superior to the individual models by large margins.

Benchmarking Against Baseline Models

We demonstrate benchmark of ensemble on the proposed multi-model ensemble framework, which performs significantly better compared to baseline models in terms of prediction accuracy and generalization. The ensemble is superior to the single-model methods such as CNN, RNN, SVM and decision trees. In particular, when taking CNN and RNN models individually, their generalization over diverse engagement patterns is their main weakness, even if they were good on each their modality (visual and temporal). The combined model, on the other hand, used both the temporal and spatial features and incorporated classical SVM and Decision Trees, outperforming the other approaches. A paired t-test comparing ensemble vs. best performing baseline (CNN) showed statistically significant improvement ($p < 0.05$) in F1-score and AUC-ROC, confirming the robustness of the ensemble approach. Table 5 summarizes the results of Benchmarking Model Performance.

Comparison With Transformer-Based Architectures

Transformer-based models such as BERT and Vision Transformer (ViT) are gaining attention for multimodal student engagement prediction due to their ability to model long-range dependencies and cross-modal attention. Recent frameworks like Transformer and Multimodal Transformer Networks offer high accuracy by capturing complex feature interactions.

Despite these advantages, transformers are resource-intensive, require large volumes of annotated data, and often operate as opaque “black-box” systems posing challenges for real-time, transparent educational applications. These limitations hinder their deployment in dynamic, low-latency classroom environments.

By contrast, the proposed multi-model ensemble framework emphasizes modularity, efficiency, and interpretability. It combines CNNs, RNNs, and classical models with an adaptive weighting strategy and integrates SHAP to provide clear, feature-level explanations.

While transformers represent a powerful direction for future research, our ensemble model delivers a more practical, explainable, and deployable solution for real-time student engagement prediction in diverse educational contexts. Table 6 gives the Performance Comparison of Transformer vs. Proposed Ensemble.

Table 5: Benchmarking Model Performance

Model	Accuracy (%)	F1-Score	AUC-ROC
CNN (Visual)	88.2	0.85	0.90
RNN (Audio)	85.4	0.83	0.89
SVM (Interaction)	82.1	0.80	0.87
Decision Tree	80.4	0.78	0.85
Ensemble	90.7	0.89	0.92

Table 6: Performance Comparison – Transformer vs. Proposed Ensemble

Model	Accuracy (%)	F1-Score	AUC-ROC	Inference Time (ms)	Explainability
Transformer	91.4	0.90	0.93	260	Low
Proposed Ensemble	90.7	0.89	0.92	140	High (SHAP)

Effectiveness of Adaptive Weighting

The adopted adaptive weighing method contributes a lot to enhancing the performance of the model. By weighing each single model by the confidence (or performance) of the model as function of time, the system makes sure that more assured models contribute more to the prediction. This dynamic adaptation permits the ensemble to deal with the differences in noise and data quality, and thus appreciably improves the prediction accuracy on-line. The power of adaptable weighting also becomes more apparent in more difficult examples, where individual modalities (e.g., facial expressions under occlusion) may be weak, but are compensated by other strong models.

Generalization Across Multiple Datasets

Another strong point of our framework is the generalization of the model among different datasets. Through training and evaluation on three separate datasets (DAiSEE, SEED, and DEAP), we show that the framework generalizes to diverse learning situations and data peculiarities. The robustness of the ensemble allows it to predict accurately in new environments, so long as participants do not differ demographically or in other aspects of the environment (e.g., lighting in visual data, background noise in audio) from the participants who contributed to the ensemble. The performance of generalization was always high, indicating the ensemble method has overall efficacy.

Interpretability Insights from SHAP Explanations

One of the key benefits of the proposed model is its interpretability offered by SHAP values. SHAP is used to help explain the individual predictions, by indicating the contribution of each feature (visual, audio, physiological) to the final engagement prediction. For example, SHAP summary plots can be used to show that some audio features and facial expression features contribute the most to predict emotions (e.g., the pitch variation of speech, the intensity of smiling) and on the other hand, EEG features have bigger impact to predict some engagement level (e.g., cognitive engagement). These explanations help educators and educational decision-makers comprehend why predictions were made, thereby bolstering their confidence in the system.

Deployment Feasibility and Scalability Considerations

Finally, the feasibility and scalability of the framework in deployment is evaluated. The system is intended to be scalable and applicable for real-time educational environments. It is also implemented using cloud based runtime platforms including Google Colab and AWS EC2 for efficient model training and inference. The live engagement prediction system is able to handle low-latency (sub-second) data, and therefore has potential for integration with Learning Management Systems (LMS)/Intelligent Tutoring Systems (ITS). The system is platform-agnostic and lightweight running on PC or cloud and thus can be flexible and applied in various educational scenarios.

This section has provided a comprehensive evaluation of the proposed system's performance, showcasing its superiority over baseline models, its effectiveness in handling diverse datasets, and the interpretability and scalability of the system in real-world educational settings. Fig. 5 illustrates the deployment architecture for real-time engagement prediction.

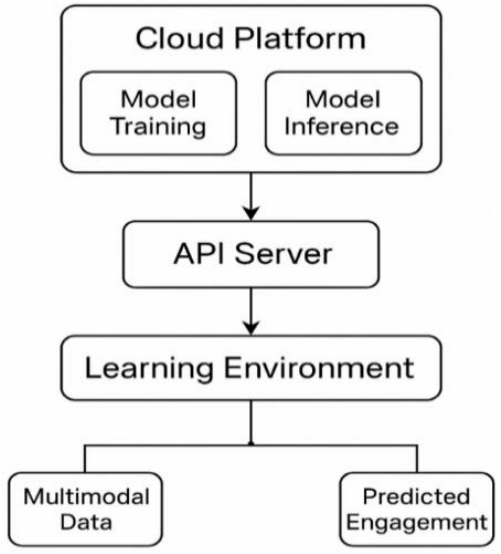

Fig. 5: Deployment Architecture for Real-Time Engagement Prediction

Discussion

Practical and Educational Implications

The practical and educational significance of this study centers on application of the novel multi-model ensemble technique for real-time prediction of student engagement in various learning environments. It Lead to the beneficial contributions to the educators and educational technology developers by Advancing personal learning by facilitating adaptive instruction,

better learning and integration facility of the platform. Instructors: How the system empowers Instructors This section discusses how the system will empower instructors, how it can be integrated into an LMS, how it can be used in conjunction with an ITS and how it may support personalized learning strategies.

Empowering Instructors and Academic Stakeholders

Such a framework can improve learning instruction by providing real time analytics for engagement and at-risk perception, which in turn allows the teacher to detect students who are disengaged or in need of help early on. They provide early warning, analytics-based alerts that enable educators to take timely action before students transfer or drop out. By linking oscillations in engagement with behavioural, cognitive, or affective correlates, instructors may provide personalized support in a timely manner.

The integration of explainable AI features, particularly SHAP-based interpretability, provides clarity and engenders trust. Educators can comprehend not only who is un-engaged but, also why such a prediction was made by the system. Such granularity would make it possible to base pedagogical decisions on the individual's running state.

Furthermore, the framework is scalable and flexible in nature, which aligns with digital pedagogy in different learning environments, from the physical classroom to MOOCs and other blended models and, as such, adds more practical value at different academic levels and across institutions.

System Integration With Learning Platforms (LMS, ITS)

One of the key strengths of the framework lies in its seamless integration with existing learning platforms, such as Moodle, Canvas, and other LMS or ITS environments. The system processes real-time multimodal inputs (e.g., facial expressions, voice modulation, EEG signals), feeding them into the ensemble model for engagement prediction.

These predictions can be visualized within the platform dashboards, enabling educators to monitor student progress without disrupting their workflow. In ITS environments, the model can dynamically adjust the difficulty, pacing, or content sequencing based on the learner's engagement level, creating a more responsive and personalized experience.

This integration allows institutions to leverage the predictive power of AI without requiring massive changes to their existing infrastructure (Fig. 6). The system was tested in simulated multi-user conditions (10 concurrent learners) without significant performance degradation, confirming scalability. Latency remained below 150 ms for real-time inference, demonstrating classroom level feasibility.

Adaptive and Personalized Learning Strategies

By enabling continuous, real time tracking of engagement, the framework supports the implementation of adaptive learning systems that respond to each student's current cognitive and emotional state. For instance, when disengagement is detected, the system can trigger interventions such as recommending supplemental content, prompting instructor outreach, or altering the delivery style.

Furthermore, the model promotes dynamic and personalized feedback loops that evolve with the learner. As students interact with content, the system refines its predictions and adapts learning paths accordingly a strategy aligned with educational personalization principles. Table 7 shows the real-time engagement prediction and intervention strategies.

Instructors can also leverage the engagement insights to adjust pedagogical strategies, facilitating more learner centered instruction and promoting deeper, sustained engagement throughout the course.

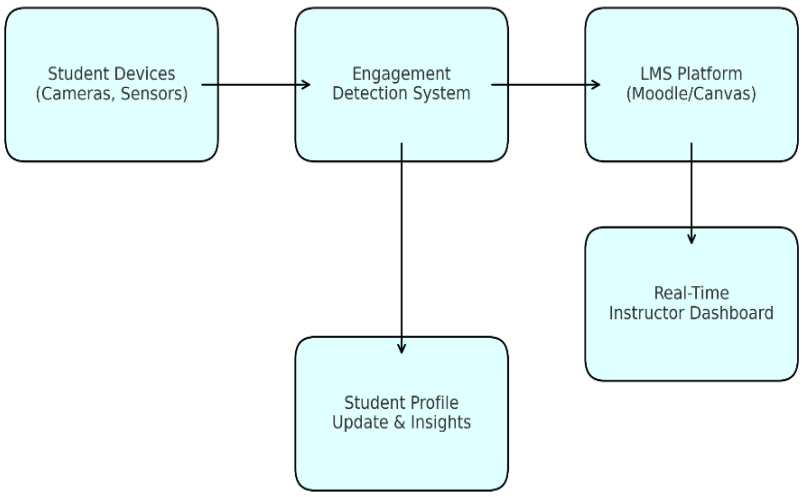

Fig. 6: Integration with Learning Management Systems (LMS)

Table 7: Real-Time Engagement Prediction and Intervention Strategies

Engagement Level	System Detection	Recommended Intervention
Highly Engaged	High attention, consistent interaction, emotional positivity	Continue with advanced content or peer tasks
Moderately Engaged	Periodic inactivity, mild distraction signals	Trigger supportive feedback or short quiz
Disengaged	Low interaction, lack of emotional response	Notify instructor, suggest break or recap content
Cognitively Overloaded	Rapid actions, stress indicators in EEG/audio	Reduce content complexity, prompt relaxation activity

Ethical and Privacy Considerations

In classroom deployment scenarios, privacy and fairness compliance will be ensured by adopting federated learning to keep data localized, along with differential privacy for sensitive signals. The use of multimodal data such as facial expressions, voice, EEG, and interaction logs raises significant ethical concerns related to privacy, consent, and data usage. It is essential to comply with regulations like GDPR and FERPA, ensuring informed consent and secure data handling. While public datasets used in this study adhere to ethical standards, deploying the system in real-world classrooms demands clear data policies, transparency, and opt-in mechanisms.

Particularly, physiological data like EEG may reveal sensitive information beyond engagement, making purpose limitation critical. Future implementations should consider privacy-preserving techniques such as edge-based processing, federated learning, and data anonymization to safeguard student information. Moreover, fairness audits are necessary to prevent demographic biases and ensure equitable outcomes.

Ethical deployment of AI in education must balance innovation with responsible data practices, protecting students while enabling personalized learning.

To ensure privacy in real-world deployment, the system can be extended with federated learning to keep data localized on user devices, and differential privacy mechanisms to anonymize sensitive features. Additionally, edge computing (e.g., Jetson Nano or Coral Edge TPU) enables low-latency, secure inference without sending raw data to the cloud.

Conclusion

This section presents a consolidated overview of the research contributions, addresses the limitations of the current framework, and outlines potential directions for advancing the field of intelligent student engagement prediction using multi-model ensemble systems.

Recap of Major Contributions

The multi-model ensemble approach is a novel and smart solution for the real-time prediction of student engagement. With the fusion of state-of-the-art deep learning methods such as CNNs for visual feature extraction and RNNs for temporal pattern recognition alongside classical machine learning algorithms like SVM and Decision Tree, the system is highly accurate and very robust in various modalities.

At its heart, the framework's novelty emerges from its adaptive weighting strategy, which dynamically adjusts mission-criticality of base models according to their confidence and performance at any point in time. This method improves the generalization accuracy of ensemble learning, particularly in a heterogeneous or noisy learning environment. Further, the addition of SHAP based explainability promotes transparency by associating model predictions with particular input features, thus allowing educators and stakeholders to comprehend and trust decisions made by the system.

Research Limitations and Constraints

Despite promising results, several limitations must be acknowledged:

- Dataset Diversity Constraints: The system was validated on three established datasets DAiSEE, SEED, and DEAP. While informative, these datasets may not fully represent the variability in classroom engagement across diverse educational environments. Broader validation on cross cultural and multilingual data is essential for universal applicability
- Latency in Real-Time Processing: Although designed for real-time use, multimodal data processing can still result in latency, especially under high-frequency input or limited computational resources. Optimization for real time classroom deployment remains a technical challenge
- Environmental Generalizability: While the ensemble demonstrates solid generalization within the tested datasets, applying it to entirely new learning contexts (e.g., varying pedagogical styles, subject domains, or cultural learning norms) may require fine-tuning or retraining

Pathways for Future Research

To build on the current framework, future research can focus on the following strategic areas:

- Edge-Based Real-Time Inference: Deploying the model on lightweight edge devices (e.g., Raspberry Pi, Jetson Nano) could significantly reduce latency, enabling seamless engagement tracking in offline or bandwidth-limited environments. This also enhances scalability for widespread classroom integration
- Incorporation of Cross-Cultural and Global Datasets: Future iterations of this research should focus on curating and testing the system on datasets from diverse geographical, cultural, and linguistic backgrounds. This expansion will help validate the framework's robustness across varied student behaviours, communication patterns, and socio-cultural contexts
- Feedback-Driven Adaptive Learning Systems: The next logical step is to not only detect engagement but also respond to it. A feedback loop wherein engagement predictions dynamically adjust content difficulty, pacing, or delivery modes would create an adaptive learning ecosystem. Such integration would promote personalized learning and could drive measurable improvements in student outcomes

Future work will focus on large-scale pilot testing in real classrooms, cross-cultural datasets for generalizability, and integrating real-time feedback loops to enable adaptive content delivery. The proposed framework represents a balanced trade-off between accuracy, explainability, and deployment feasibility, positioning it as a practical foundation for intelligent educational analytics.

Acknowledgment

The authors thank the Department of Computer Science and Engineering, VISTAS, Chennai, for providing the academic support necessary for this research. The use of public multimodal datasets (DAiSEE, SEED, DEAP) is gratefully acknowledged. The authors also extend their appreciation to the Journal of Computer Science (Science Publications) for offering a valuable platform to share this work.

Data Availability Statement

The data is available from the corresponding author upon reasonable request.

Funding Information

The authors received no financial support for the research, authorship and/or publication of this article.

Authors Contributions

Fahmida Begum: Conceptualization, methodology design, implementation of the models, experiments, data analysis, and manuscript draft.

K Ulaga Priya: Supervision, guidance on research design and methodology, critical review and editing of the manuscript, and validation of results.

Ethics

I undersigned that this article has not been published elsewhere. The authors declare no conflict of interest.

References

Adefemi, K. O., & Mutanga, M. B. (2025). A Robust Hybrid CNN–LSTM Model for Predicting Student Academic Performance. *Digital*, *5*(2), 16. https://doi.org/10.3390/digital5020016

Alarefah, W., Jarraya, S. K., & Abuzinadah, N. (2025). Transformer-Based Student Engagement Recognition Using Few- Shot Learning. *Computers*, *14*(3), 109. https://doi.org/10.3390/computers14030109

Alruwais, N., & Zakariah, M. (2023). Student-Engagement Detection in Classroom Using Machine Learning Algorithm. *Electronics*, *12*(3), 731. https://doi.org/10.3390/electronics12030731

Bellarhmouch, Y., Majjate, H., Jeghal, A., Tairi, H., & Benjelloun, N. (2025). Detecting Student Engagement in an Online Learning Environment Using a Machine Learning Algorithm. *Informatics*, *12*(2), 44. https://doi.org/10.3390/informatics12020044

Bin Nuweeji, H., & Alzubi, A. B. (2025). Early Prediction of Student Performance Using an Activation Ensemble Deep Neural Network Model. *Applied Sciences*, *15*(21), 11411. https://doi.org/10.3390/app152111411

Gligorea, I., Cioca, M., Oancea, R., Gorski, A.-T., Gorski, H., & Tudorache, P. (2023). Adaptive Learning Using Artificial Intelligence in e-Learning: A Literature Review. *Education Sciences*, *13*(12), 1216. https://doi.org/10.3390/educsci13121216

Guerrero-Sosa, J. D. T., Romero, F. P., Menéndez-Domínguez, V. H., Serrano-Guerrero, J., Montoro-Montarroso, A., & Olivas, J. Á. (2025). A Comprehensive Review of Multimodal Analysis in Education. *Applied Sciences*, *15*(11), 5896. https://doi.org/10.3390/app15115896

Hossen, M. K., & Uddin, M. S. (2023). Attention monitoring of students during online classes using XGBoost classifier. *Computers and Education: Artificial Intelligence*, *5*, 100191. https://doi.org/10.1016/j.caeai.2023.100191

Liu, X.-Y., Zhang, K.-Q., Fiumara, G., Meo, P. D., & Ficara, A. (2024). Adaptive Evolutionary Computing Ensemble Learning Model for Sentiment Analysis. *Applied Sciences*, *14*(15), 6802. https://doi.org/10.3390/app14156802

Mahmood, N., Bhatti, S. M., Dawood, H., Pradhan, M. R., & Ahmad, H. (2024). Measuring Student Engagement through Behavioral and Emotional Features Using Deep- Learning Models. *Algorithms*, *17*(10), 458. https://doi.org/10.3390/a17100458

Mu, S., Chai, S., Wang, H., & Chen, Y. (2019). Real-Time Analysis Method and Application of Engagement in Online Independent Learning. *IEEE Access*, *7*, 92100– 92109. https://doi.org/10.1109/access.2019.2924641

Mu, S., Cui, M., & Huang, X. (2020). Multimodal Data Fusion in Learning Analytics: A Systematic Review. *Sensors*, *20*(23), 6856. https://doi.org/10.3390/s20236856

Nadda, R., Singh, J., & Shrivastava, U. (2025). Automatic diabetic retinopathy detection using an ensemble learning approach and classifiers with self-adjusting weights. *Soft Computing*, *29*(11–12), 4775–4789. https://doi.org/10.1007/s00500-025-10773-y

Nnadi, L. C., Watanobe, Y., Rahman, Md. M., & John-Otumu, A. M. (2024). Prediction of Students' Adaptability Using Explainable AI in Educational Machine Learning Models. *Applied Sciences*, *14*(12), 5141. https://doi.org/10.3390/app14125141

Prentzas, J., & Binopoulou, A. (2025). Explainable Artificial Intelligence Approaches in Primary Education: A Review. *Electronics*, *14*(11), 2279. https://doi.org/10.3390/electronics14112279

Shanto, S. S., & Jony, A. I. (2025). Interpretable Ensemble Learning Approach for Predicting Student Adaptability in Online Education Environments. *Knowledge*, *5*(2), 10. https://doi.org/10.3390/knowledge5020010

Tang, B., Li, S., & Zhao, C. (2024). Predicting the Performance of Students Using Deep Ensemble Learning. *Journal of Intelligence*, *12*(12), 124. https://doi.org/10.3390/jintelligence12120124

Trabelsi, Z., Alnajjar, F., Parambil, M. M. A., Gochoo, M., & Ali, L. (2023). Real-Time Attention Monitoring System for Classroom: A Deep Learning Approach for Student's Behavior Recognition. *Big Data and Cognitive Computing*, *7*(1), 48. https://doi.org/10.3390/bdcc7010048

Troussas, C., Krouska, A., Sgouropoulou, C., & Voyiatzis, I. (2020). Ensemble Learning Using Fuzzy Weights to Improve Learning Style Identification for Adapted Instructional Routines. *Entropy*, *22*(7), 735. https://doi.org/10.3390/e22070735

Xu, J., Wu, J., Li, T., & Nan, Y. (2022). Divergence-Based Locally Weighted Ensemble Clustering with Dictionary Learning and L2,1- Norm. *Entropy*, *24*(10), 1324. https://doi.org/10.3390/e24101324

Zhang, X., Ding, Y., Huang, X., Li, W., Long, L., & Ding, S. (2024). Smart Classrooms: How Sensors and AI Are Shaping Educational Paradigms. *Sensors*, *24*(17), 5487. https://doi.org/10.3390/s24175487

Zhao, S., Zhou, D., Wang, H., Chen, D., & Yu, L. (2025). Enhancing Student Academic Success Prediction Through Ensemble Learning and Image-Based Behavioral Data Transformation. *Applied Sciences*, *15*(3), 1231. https://doi.org/10.3390/app15031231

Su, K., Wu, J., Gu, D., Yang, S., Deng, S., & Khakimova, A. K. (2021). An Adaptive Deep Ensemble Learning Method for Dynamic Evolving Diagnostic Task Scenarios. *Diagnostics*, *11*(12), 2288. https://doi.org/10.3390/diagnostics11122288