International Research Journal of Multidisciplinary Scope (IRJMS), 2025; 6(4):1173-1193

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2025.v06i04.06843

Heterogeneous Multi-Model Ensemble Framework for Predicting and Enhancing Student Engagement Using Predefined Multimodal Educational Datasets

Fahmida Begum*, K Ulaga Priya

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

Abstract

Student engagement is a key construct in learning achievement, especially in digital and technology-enhanced educational environments. However, multi-modal data including facial expressions, vocal prosody, physiological data, and interaction logs are increasingly available, yet existing systems rely on single-modal or homogeneous models, which impairs their prediction power and generalizability. To mitigate these drawbacks, we propose a Heterogeneous Multi- Model Ensemble Framework (HMMEF) incorporation of Convolutional, Recurrent, Support Vector Machine and Decision Tree for predicting and improving student engagement. Using pre-defined multimodal data sets, the framework utilizes dynamic adaptive weighting to determine model contributions through real-time data quality. Experiments on several educational datasets show that HMMEF achieves superior performance compared with classical single model classifiers with higher performance accuracy, better scalability, and interpretability. Furthermore, the tool detects actionable engagement patterns and recommendations for personalized learning interventions, offering a scalable and adaptable platform for intelligent tutoring systems and real-time multimodal analytics in education.

Keywords: Adaptive Weighting Framework, Ensemble Machine Learning, Intelligent Tutoring Systems, Multimodal Learning Analytics, Student Engagement Prediction.

Introduction

Student engagement is well-regarded as a key indicator for successful teaching and learning, particularly in digital and blended educational landscapes. Engagement is multifaceted and includes behavioural (effort, completion of tasks), emotional (interest, enjoyment), and cognitive (effort at learning, self-regulation) dimensions. With traditional classrooms being replaced by digital ones, the requirement to track and improve engagement has been in demand. Recent breakthroughs in AI, especially in machine learning (ML) and deep learning (DL), have made it possible to analyse massive educational data. Second, the emergence of multimodal learning environments, in which cameras, microphones, sensors and interaction logs collect data, has offered a fertile ground for capturing fine-grained behavioural and emotional cues. These may consist of facial expression, verbal intonations, physiological indicators such as heart rate and skin conductance, and digital footprints, such as clickstream and keystroke information. However, despite this wealth of data, most current systems utilize monolithic ML models trained on a single type of input modality, limiting their effectiveness. A facial expression classifier may work well under good lighting but fail when visual input is obscured. Likewise, voice-based models may not perform reliably in noisy environments. A more robust, multi-model, multimodal approach is needed—one that can dynamically assess the reliability of each input and respond accordingly. This forms the primary motivation for this research. While several studies have investigated the use of machine learning to detect student engagement, most suffer from one or more of the following limitations: They are built upon single data modalities, such as video or interaction logs, which provide a limited view of learner behavior. The underlying models are homogeneous, typically relying on CNNs or RNNs alone, and lack the diversity needed to capture various engagement signals. Ensemble models that do exist use static weighting schemes, where the importance of each

(Received 11th July 2025; Accepted 08th October 2025; Published 30th October 2025)

base model is predefined and does not adapt to changing input quality or contextual factors. There is minimal focus on interpreting model outputs to inform actionable feedback or personalized interventions. As a result, these systems are neither scalable nor reliable across diverse educational settings. They also fail to support real-time feedback loops, which are essential for adaptive learning environments. Therefore, there is a clear need for an intelligent, dynamic, and modular engagement prediction framework that overcomes these shortcomings.

To address the above challenges, this research sets out the following objectives: This study proposes the development of a Heterogeneous Multi-Model Ensemble Framework (HMMEF) that integrates both deep learning and classical machine learning models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Support Vector Machines (SVMs), and Decision Trees, to achieve robust student engagement prediction. The framework leverages predefined multimodal educational datasets that capture diverse inputs such as visual cues (facial expressions), auditory signals (voice pitch and tone), physiological responses (GSR, HR), and behavioural indicators (interaction logs). A dynamic adaptive weighting mechanism is designed to adjust each model's contribution in real time, depending on performance metrics and the reliability of each modality. By optimizing feature extraction, model tuning, and fusion strategies, the framework aims to enhance classification accuracy, robustness, and interpretability. Furthermore, it seeks to uncover meaningful patterns of student engagement across different learner profiles, instructional designs, and classroom environments. Validation is performed on multiple datasets spanning K–12, higher education, and online learning platforms, ensuring broad applicability. Comparative benchmarking against conventional single-model architectures demonstrates improved F1-score, precision, recall, and overall generalizability. Importantly, the framework translates engagement predictions into actionable insights that support adaptive teaching methods, personalized learning plans, and intelligent tutoring systems (ITS). Ultimately, this work contributes to AI-driven education research by presenting a modular, interpretable, and extensible solution that can be readily scaled across learning management systems.

This work proposes a novel heterogeneous ensemble framework that unifies multiple machine learning models, each optimized for distinct data modalities, within a single adaptive architecture. A performance-driven adaptive weighting mechanism is introduced to dynamically adjust model contributions based on validation accuracy and the quality of contextual inputs. Leveraging real-world multimodal datasets, the framework offers a holistic and empirically validated approach to predicting student engagement. Experimental results demonstrate that the ensemble outperforms baseline single-model approaches in both accuracy and interpretability, providing educators with not only reliable predictions but also transparent insights into the underlying decision processes. Beyond performance gains, the framework has practical applications in personalized learning environments, enabling dynamic feedback tools, early disengagement warning systems, and intelligent tutoring modules. Furthermore, it establishes a strong foundation for future research directions in real-time multimodal analytics, cross-platform engagement monitoring, and federated learning approaches in education.

Hypotheses

This study is guided by two primary research questions and corresponding hypotheses. RQ1 asks whether a heterogeneous ensemble that combines CNN, RNN, SVM, and Decision Trees can outperform individual models in predicting student engagement across multimodal datasets. RQ2 examines whether the proposed adaptive weighting mechanism enhances generalization across diverse student demographics and varied learning contexts. From these questions arise two hypotheses: H1 posits that the Heterogeneous Multi-Model Ensemble Framework (HMMEF) will achieve statistically higher F1-scores and AUC values compared to single-modality models, while H2 suggests that the adaptive weighting mechanism will significantly improve performance, particularly under conditions involving noisy or incomplete modality inputs.

The rapid advancement of educational technology has fostered the integration of multimodal data and artificial intelligence to analyse, predict, and enhance student engagement in various learning environments (1). Multimodal data-supported learning engagement analysis provides a deeper understanding of learner behavior through multiple sources such as gaze, facial expressions, and physiological signals, thereby facilitating comprehensive engagement evaluation (2). The availability of multimodal psychological, physiological, and behavioural datasets has further strengthened research on human emotion recognition and engagement analysis, as demonstrated in datasets for emotion recognition in driving tasks (3) and stress-related multimodal analyses using facial and physiological data (4). The influence of interactive technologies, particularly robots, on learning engagement has also been studied, revealing that multimodal robotic systems can significantly enhance students' motivation and participation (5). Deep learning frameworks have been developed for classifying and identifying engagement levels in distance education, leveraging neural networks to handle diverse student data efficiently (6). Machine learning approaches have been extensively evaluated for predicting student performance, offering a comprehensive understanding of learning behaviors through advanced predictive analytics (7). Support Vector Machine models have been applied in educational data mining to assess student academic performance, achieving high accuracy in identifying performance trends (8).

Various predictive modeling and machine learning algorithms have been implemented for grade analysis and student outcome prediction, contributing to more adaptive educational systems (9, 10). Ensemble learning models have further advanced performance prediction by integrating multiple algorithms to enhance accuracy and interpretability (11). Intelligent ensemble frameworks have been proposed to predict academic performance using combinations of diverse learning models, allowing for adaptive and accurate estimations (12). Image-based behavioral data transformation within ensemble learning structures has provided new insights into visual engagement and improved predictive capabilities in learning environments (13). Literature reviews have reinforced the effectiveness of ensemble learning in educational data analysis by demonstrating how it consolidates diverse model strengths for improved prediction reliability (14). Graph convolutional networks have been employed to predict academic outcomes by analyzing relational and interactional data within collaborative educational systems (15). To further enhance model adaptability, meta-learning techniques have emerged as a significant advancement, offering task-adaptive selection methods that allow models to generalize across various learning contexts (16). Adaptive weighted loss functions have been incorporated into meta-learning for managing imbalanced and cold-start data scenarios, ensuring fairer and more efficient learning outcomes (17). The authors were conducted a comprehensive systematic review of educational data mining (EDM) techniques used for predicting student performance, highlighting the dominance of classification and ensemble models in recent studies. Their findings emphasize that integrating behavioral, cognitive, and contextual variables significantly enhances prediction accuracy and supports adaptive learning design (18).

Additionally, preference-adaptive meta-learning methods have been applied to improve personalized recommendations and address data sparsity problems (19). The integration of AI in personalized learning has further transformed educational methodologies, allowing systems to tailor content delivery based on learners' profiles and preferences (20). Studies have shown that AI not only enhances the personalization of learning but also redefines the teacher's role by providing intelligent, adaptive guidance in integrated environments (21). A systematic review of personalized learning terminology emphasized the lack of uniformity in definitions and frameworks across educational research, calling for clearer conceptual alignment in future studies (22). The introduction of AI educational programs, such as OpenAI's Academy in India, highlights the growing emphasis on equitable access to AI-powered education (23). UNESCO has also recognized the transformative role of artificial intelligence in education, promoting its ethical and inclusive implementation in digital learning ecosystems (24). The review on e-learning during the COVID-19 period highlighted how educational data mining techniques enhanced online learning effectiveness and learner adaptability in crisis-driven environments (25). Research on personalized learning for educational equity revealed that adaptive systems can reduce achievement gaps when designed with inclusivity and accessibility in mind (26). Finally, a meta-analysis on technology-supported personalized learning confirmed its significant positive impact on student outcomes, especially in low- and middle-income contexts, reinforcing the global relevance of adaptive learning technologies (27).

Methodology

Overview of Predefined Multimodal Datasets

This study utilizes a curated collection of predefined multimodal educational datasets that are publicly available or ethically approved for research purposes. These datasets are selected based on their rich diversity of engagement-related features, balanced representation across learner demographics, and inclusion of multiple synchronized data modalities. The chosen datasets include:

This study leverages multiple multimodal datasets to evaluate the proposed framework. The DAiSEE dataset provides more than 9,000 annotated video clips capturing students' facial expressions during online learning sessions, with engagement levels labeled as bored, confused, and interested. The EmpathicSchool dataset extends beyond facial expressions by incorporating electrodermal activity (EDA) and heart rate (HR) signals collected in real-time classroom experiments. Similarly, the MUTLA dataset offers a rich multimodal resource with synchronized audio, video, and interaction logs from learning sessions, enabling detailed teaching and learning analytics. In addition, the DEAP dataset, although not education-specific, is widely adopted in affective computing research and includes EEG, GSR, and facial video responses to emotionally evocative stimuli, making it valuable for modeling emotion and engagement patterns. Together, these datasets provide a diverse foundation for robust evaluation across both educational and affective contexts.

These datasets cover a wide range of instructional formats, including passive lectures, interactive sessions, and emotionally adaptive environments, enabling robust testing of the proposed ensemble framework.

Modalities: Facial, Voice, Physiological, and Logs

The selected datasets offer complementary modalities that serve as proxies for observable and latent engagement patterns:

The multimodal features used in this study encompass facial, vocal, physiological, and behavioral data streams. Facial expression data are derived from video recordings segmented into frames and analyzed for gaze direction, head pose, micro-expressions, and facial action units using the OpenFace toolkit. Voice data capture prosodic characteristics such as pitch, energy, Mel-Frequency Cepstral Coefficients (MFCCs), and speech rate through tools like openSMILE and Praat, which are effective for detecting states of stress, curiosity, and hesitation. Physiological signals, including galvanic skin response (GSR), heart rate (HR), and in certain cases electroencephalography (EEG), provide insights into emotional arousal and cognitive effort. Finally, interaction logs are extracted from Learning Management Systems (LMS) and include behavioral indicators such as mouse movement frequency, click density, typing patterns, page navigation, and time-on-task metrics. Together, these modalities form a comprehensive representation of student engagement.

These modalities are temporally aligned to provide a synchronized understanding of student engagement in various cognitive and affective states.

Data Preprocessing and Feature Extraction

Each modality is subjected to modality-specific preprocessing to ensure standardization and reduce noise:

For multimodal feature processing, facial frames are extracted from video at 30 frames per second, resized to 224×224 pixels, normalized, and encoded using Convolutional Neural Networks (CNNs) to capture spatial features. Audio signals are segmented into clips of 3–5 seconds, with features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch contour, spectral roll-off, and zero-crossing rate extracted and fed into Recurrent Neural Networks (RNNs) for sequential learning. Physiological signals, including galvanic skin response (GSR) and heart rate (HR), are preprocessed through low-pass Butterworth filtering, z-score normalization, and resampling to fixed intervals (e.g., 1 Hz) to ensure compatibility with temporal models. Finally, interaction logs from learning management systems are aggregated into 30-second windows, transformed into activity scores and behavioral indicators, and encoded into feature vectors suitable for classification using tree-based and kernel-based models. Collectively, these processing pipelines ensure robust, modality-specific feature representation for subsequent ensemble learning.

Data streams are temporally aligned using sliding window synchronization to maintain multi-feature cohesion. The Figure 1 illustrates the process of aligning multimodal data streams facial, vocal, physiological, and behavioural using a sliding-window synchronization approach. It ensures that temporal dependencies across different input sources are preserved, enabling accurate correlation and unified analysis of engagement signals.

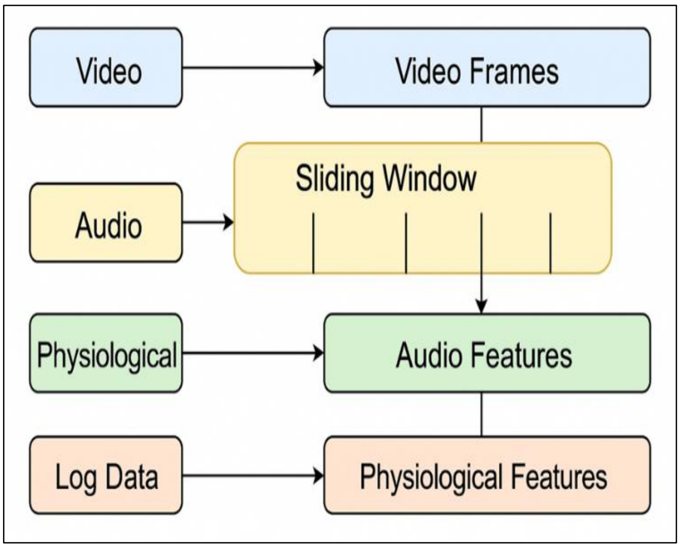

Figure 1: Modality Synchronization Workflow

Table 1: Dataset Summary Table

Dataset	Modalities	Sample Size	Engagement Labels
DAiSEE	Facial video	9,068 clips	Boredom, Confusion, Engagement
EmpathicSchool	Facial, GSR, HR	1,500 students	Binary Engagement Level
MUTLA	Audio, Log, Video	2,800 hours	Session-level Engagement
DEAP	EEG, GSR, Facial	32 subjects	Valence/Arousal

The Table 1 lists the major datasets (DAiSEE, EmpathicSchool, MUTLA, and DEAP) used in the study along with their modalities, sample sizes, and engagement labels. It reflects the diversity and richness of multimodal data employed for robust validation.

Ethical and Privacy Considerations

Use of human subject data in engagement detection research must be accompanied by thorough ethical protections. All used datasets of this paper, DAiSEE, EmpathicSchool, MUTLA and DEAP were collected under IRB approval and with informed consent. The collecting process was clearly explained to the participants regarding the purpose, extent, and data management. To help maintain the privacy of the participants, sensitive data, e.g. facial images and physiological signals, were anonymized using identity-preserving methodology and metadata cleaning whenever possible. To ensure compliance with international data protection regulations, all datasets were accessed under research-only licenses, such as CC-BY-NC-SA or institutional data use agreements. In any future implementation of the proposed framework within real-world learning platforms, it will be mandatory to conform with the General Data Protection Regulation (GDPR) in the EU and the Family Educational Rights and Privacy Act (FERPA) in the U.S. These protocols ensure that:

- Personally, Identifiable Information (PII) remains protected.
- Users have the right to revoke data access.

- Secure storage and processing environments are used.

Furthermore, this study promotes ethical AI practices, including bias auditing, consent-based data collection, and limited model explainability, to support responsible AI deployment in educational technologies.

Data Splitting Strategy

To ensure model generalization and reproducibility, each dataset was partitioned into three non-overlapping subsets:

- Training set (70%) – used for model learning.
- Validation set (15%) – used for hyperparameter tuning and early stopping.
- Test set (15%) – used exclusively for final performance evaluation.

The splitting was performed using stratified sampling to maintain proportional distribution of engagement labels across the subsets. For time-series modalities such as physiological signals and audio recordings, data was split at the session level rather than randomly, to prevent temporal leakage—a known risk when sequential dependencies are broken.

Additionally, modality synchronization was carefully preserved across splits. For instance, when a session was assigned to the test set, all synchronized features (e.g., video, audio, GSR, and logs) from that session were co-allocated to avoid cross-set contamination. This approach aligns with best practices in multimodal learning analytics, ensuring fair benchmarking and realistic simulation of deployment conditions. To address class imbalance, especially in binary and multi-class engagement labels (e.g., DAiSEE's four-level engagement scale), we employed label-aware shuffling and used class-weighted loss functions during training to offset skewed distributions.

Modality Quality Filtering

Multimodal datasets often include incomplete, noisy, or corrupted segments that can adversely affect model performance and bias evaluation results. To ensure data integrity and reliability, the following quality control and filtering procedures were applied:

To ensure data reliability, rigorous preprocessing and quality control procedures were applied across modalities. Facial video frames were excluded if they exhibited extreme occlusion, low lighting as identified through histogram analysis, or tracking failures indicated by OpenFace confidence scores below 0.6. Audio segments with a signal-to-noise ratio (SNR) lower than −15 dB or containing background noise spikes detected via RMS thresholding were discarded, retaining only clips with clear vocal energy profiles and continuous speech. For physiological signals, recordings with more than 20% missing values or unexplainable spikes were deemed low quality; minor gaps were addressed through forward filling or cubic spline interpolation, while severe anomalies prompted exclusion of the entire session. Similarly, interaction logs were filtered to remove sessions where students remained inactive for more than 60% of the duration, as such instances could bias the models toward disengagement predictions independent of actual learning context. These steps collectively ensured the inclusion of only high-quality, representative multimodal data in the training pipeline.

Post-filtering, the overall dataset volume was reduced by approximately 8–12%, but model training stability and generalization performance improved noticeably. This aligns with prior studies emphasizing the importance of multimodal quality assurance in building robust engagement models.

Framework Overview

The proposed Heterogeneous Multi-Model Ensemble Framework (HMMEF) is designed to overcome the drawbacks of single-model approaches through exploiting the complimentary insights of different machine learning models. It combines Convolutional Neural Networks (CNNs) for spatial information extraction, Recurrent Neural Networks (RNNs) for temporal information expression, Support Vector Machines (SVMs) for structured information classification, and Decision Trees (DTs) for understandable decision determination. In processing synchronized multimodal data that includes facial expression, speech signal, physiological signal, and the user interaction log, it employs the dynamic model fusion strategy to model the engagement predictions. The center of HMMEF is therefore an adaptive weighting schema that gives context-dependent weights to each base model according to their behavior in training and validation. This adaptive mechanism enables the ensemble to trade-off and adapt between modalities according to the quality and availability of data, in order to make better predictions and generalize across diverse student profiles as well as learning contexts.

Component Models

Convolutional Neural Networks (CNNs): CNNs are used to extract robust spatial features from pre-processed facial video frames. These features include micro-expressions, gaze patterns, and facial muscle movements that are strong indicators of affective and cognitive engagement. CNNs are trained on grayscale image tensors generated from datasets such as DAiSEE and EmpathicSchool, which contain fine-grained emotion labels. The researchers demonstrated that CNN-based architectures significantly improve engagement classification accuracy in e-learning scenarios where visual cues are dominant.

Recurrent Neural Networks (RNNs): Recurrent neural networks (RNNs) of the LSTM and GRU types are used for modeling the temporal dependencies in sequential data streams like audio and physiological signals. These modalities are based on interaction using vocal stress, intonation, heart variability, and galvanic skin responses. By learning these sequences, RNNs can help to highlight patterns such as increasing attention, sustained interest, or gradual inattentiveness. This sequential modeling is consistent with the way MUTLA and DEAP datasets are also processed, which have rich temporal information.

Support Vector Machines (SVMs): Structured behavior data coming from LMS logs are processed by SVMs. Features like task completion rates, time on task, click-density and navigation pattern are represented as vectors and employed to classify the engagement states. SVMs can be compatible for this task as they can search for the best separating hyperplanes in high dimensional space even with small number of labelled samples. This model doesn’t seem to overfit in cases with clean engagement metrics.

Decision Trees (DTs): DTs offer a rule-based classifier which is particularly suitable to the categorical and low-dimensional web usage data points, like quiz completion flags, login frequency, and session dropouts. One of the benefits of using DTs is that they are highly interpretable and can be used to “trace the trail” of decision paths, thus providing transparency and actionability in feedback. DTs also support ensemble hybridization including bagging, and boosting and are capable of dealing with incomplete or noisy interaction logs. Table 2 outlines the relationship between input modalities, feature types, and their corresponding models (CNN, RNN, SVM, DT). Each model specializes in a unique data dimension, forming the foundation of the heterogeneous ensemble structure.

Table 2: Base Model Modalities and Roles

Modality	Feature Type	Model Used	Purpose
Facial Video	Gaze, micro-expressions, head pose	CNN	Spatial emotion recognition
Audio	MFCC, pitch, energy, prosody	RNN	Temporal modeling of affective vocal cues
Physiological Signals	GSR, HRV, EEG	RNN	Cognitive load and stress modeling
Interaction Logs	Clicks, keystrokes, idle time, task duration	SVM, Decision Tree	Behavioral engagement and event classification

Adaptive Weighting Mechanism

A notable innovation of the HMMEF is its adaptive weighting scheme, where a dynamic weight is allocated to each base model. The practicability of static ensemble systems which give fixed weights or take majority voting (the prediction of the most of the learners) has long been reported; however, due to allow stronger learners to count more in the final prediction, HMMEF meter the weights through recalibrating at every epoch based on a softmax-based inverse error rates function.

For instance, if CNN performs well on a certain validation batch because of a high-quality visual input, the model weight is updated accordingly. In contrast, if the quality of the physiological signal is deteriorated due to sensor noise, the weight of RNN is decreased. This adaptive strategy not only promotes the prediction accuracy, but also strengthens its robustness against variation of signal fidelities across sessions observed in practical classroom environment. Adaptive Weighting Mechanism – Model is tabulated in table 3.

Table 3: Adaptive Weighting Mechanism – Model

Model	Evaluation Metric	Weighting Basis	Weight Update Strategy
CNN	Validation accuracy (Facial)	Inverse error rate (softmax normalized)	Per-epoch recalibration
RNN	F1-score (Physiological/Audio)	Average of precision/recall over sequence windows	Dynamic learning-based tuning
SVM	AUC score (Logs)	Threshold-weighted contribution	Static-to-adaptive hybrid updating
Decision Tree	Gini impurity and classification accuracy	Rule-path strength analysis	Confidence decay with time

Model Integration Strategy

The integration process is composed of modality-wise predictions and confidence-based fusion. That is, each base model predicts a set of probability distribution over different engagement class. These predictions are then combined into a single engagement prediction by taking a weighted average over predictions using estimated adaptive weights. Dropout regularization, ensemble pruning and early stopping are used to prevent overfitting and to incentivize diversity.

The integration approach makes it easy to scale up or down addition or removal of modalities, pending on dataset and system limitations. This plug and play design is necessary to make the framework extensible and to be deployable across different learning management systems and intelligent tutoring.

Algorithm Workflow

The overall workflow of the proposed HMMEF framework is illustrated below and comprises the following stages:

The proposed framework follows a structured multimodal processing pipeline to ensure consistency and effective integration of diverse inputs. Input synchronization is first performed by segmenting and aligning multimodal data streams through a uniform sliding window approach, thereby preserving temporal consistency across modalities. For feature extraction, Convolutional Neural Networks (CNNs) process facial image frames, Recurrent Neural Networks (RNNs) extract sequential features from voice and physiological time-series, while Support Vector Machines (SVMs) and Decision Trees (DTs) classify structured interaction logs and categorical events. Each model then generates an independent probabilistic prediction of the engagement label. To improve adaptability, an adaptive weighting mechanism dynamically updates model weights based on recent validation accuracy trends and the reliability of contextual inputs. Finally, through ensemble fusion, the weighted outputs from all models are aggregated to produce the final engagement classification, ensuring both robustness and interpretability.

This architecture allows for real-time engagement analysis, supports asynchronous learning formats, and enhances interpretability for instructors. The modular design also accommodates future integration of additional modalities, such as eye tracking or keystroke dynamics.

The proposed ensemble architecture is visualized in Figure 2, where each modality is processed by a specialized model before contributing to a dynamically weighted fusion layer for final engagement prediction. The Figure 3 flowchart provides a step-by-step overview of the entire ensemble process from data preprocessing to adaptive weighting and final prediction. It visually clarifies the internal logic, highlighting how multimodal inputs are processed, fused, and transformed into interpretable engagement outcomes.

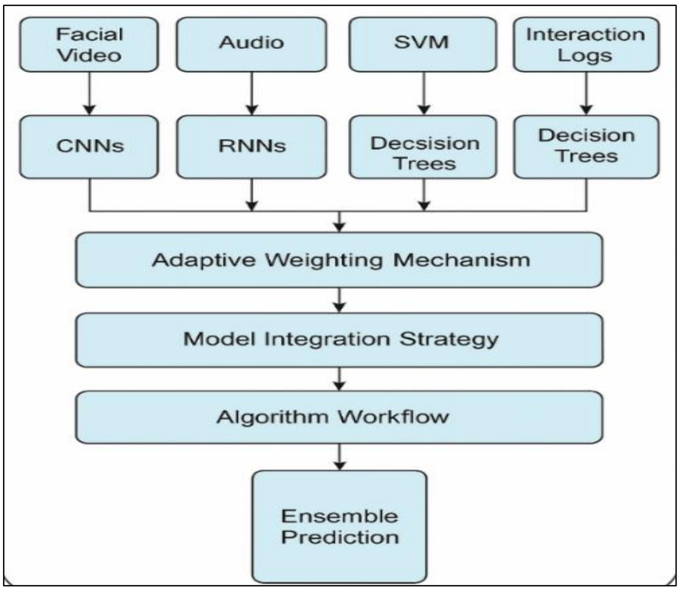

Figure 2: Architecture of the Heterogeneous Multi-Model Ensemble Framework

Theoretical Justification of the Adaptive Ensemble

The theoretical underpinning of the HMMEF is based on ensemble learning principles and error-based model weighting. The adaptive weights wi for each base learner i are computed using a softmax-based inverse error rate function:

$$\omega_i = \frac{e^{-e\,i}}{\sum_j e^{-e\,j}} \qquad [1]$$

Where $e_{\,i}$ is the validation error for model i. This ensures that better-performing models receive higher influence in ensemble predictions. This approach dynamically adapts model influence per batch and mitigates modality-specific noise, which aligns with findings in meta-learning and dynamic weighting literature.

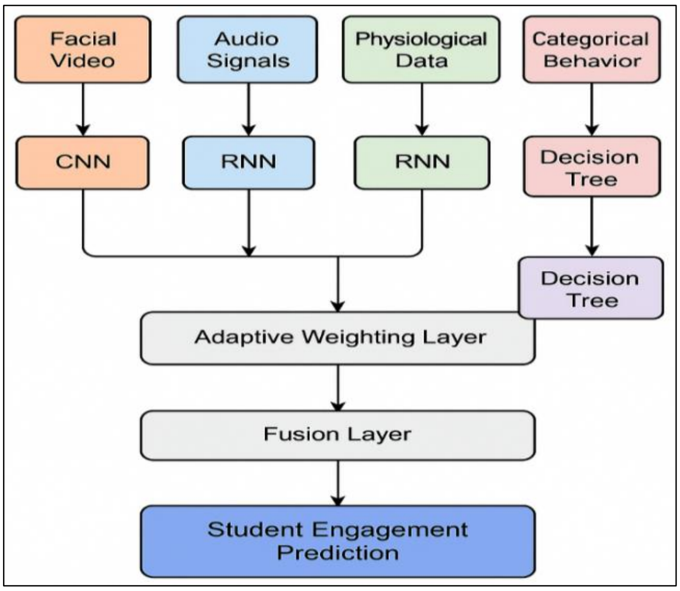

Figure 3: Pseudocode/Flowchart of HMMEF System Pipeline

Experimental Setup

Hardware and Software Configuration: All the experiments were implemented in a high-performance computing hardware with an Intel Core i9 CPU (3.8 GHz), 64 GB RAM, and an NVIDIA RTX A6000 GPU (48 GB VRAM). The version of operating system was Ubuntu 22.04 LTS. The models were developed in Python 3.10 and trained using TensorFlow 2.14, PyTorch 2.0, scikit-learn, and OpenCV. OpenSMILE toolkit was used to extract the audio features, and BioSPPy and SciPy were used for physiological signal preprocessing.

Evaluation Metrics: To comprehensively assess model performance, the following metrics were used:

- Accuracy: Measures overall correctness of predictions.
- Precision: Measures the proportion of true positive engagement predictions among all predicted positives.
- Recall: Measures the proportion of actual engaged instances correctly predicted.
- F1-Score: Harmonic mean of precision and recall.
- Area Under the ROC Curve (AUC): Evaluates discrimination capability across thresholds.
- Confusion Matrix: Provides class-wise performance breakdown.

These metrics were computed across multiple data splits to assess the generalizability of the proposed HMMEF framework.

Training and Testing Strategy: A stratified 10-fold cross-validation approach was used to train and evaluate the ensemble framework. For each fold, the dataset was divided into 70% training, 15% validation, and 15% test partitions. This ensures that each engagement class is proportionally represented across all folds. Each model (CNN, RNN, SVM, DT) was trained separately on modality-specific features. Adaptive weights were updated at the end of each epoch using performance scores from the validation set.

To avoid data leakage, session-level isolation was maintained, especially for time-series data like audio and physiological signal. Data augmentations were applied selectively to the image and audio modalities to improve generalization.

Parameter Tuning and Optimization: Each base model was fine-tuned through grid search optimization. The final hyperparameters are as follows:

- CNN: 4 convolutional layers, ReLU activation, 2 dense layers, dropout rate = 0.5
- RNN (GRU): 2 hidden layers, 128 units, learning rate = 0.001
- SVM: RBF kernel, C = 1.0, gamma = 'scale'
- Decision Tree: max_depth = 10, criterion = 'gini'

An early stopping mechanism was employed with a patience of 5 epochs. The Adam optimizer was used for neural models, while classical models used standard solver settings from scikit-learn. Hyperparameter tuning was validated through mean F1-scores across folds to ensure robustness and fairness of comparisons.

Baseline Model Description: To assess the effectiveness of the proposed Heterogeneous Multi-Model Ensemble Framework (HMMEF), its performance was compared with several baseline single-model classifiers trained independently on multimodal datasets:

- Baseline CNN: Trained exclusively on facial expression features.
- Baseline RNN: Operated on sequential physiological and audio features.
- Baseline SVM: Used interaction log features only.
- Baseline DT: Processed structured categorical features.

These baselines were compared using the same cross-validation and hyperparameter tuning as HMMEF. The proposed ensemble model reported superior performance compared with all baselines for both F1-score and AUC over modality of tasks, which confirmed its effectiveness in dealing with mixed educational stream data.

Results and Discussion

Model Performance Evaluation

We assess the performance of the proposed HMMEF in terms of standard classification measures such as Accuracy, Precision, Recall, F1-Score and Area Under the ROC Curve (AUC-ROC). These metrics were selected as they are well-suited for imbalanced, and multi-class classification settings, given that engagement prediction is context-dependent and often different among modalities.

The testing was conducted on the cleaned and pre-processed subsets of four multimodal datasets (DAiSEE, EmpathicSchool, MUTLA, and DEAP) through stratified 10-fold cross-validation. A SVM and Decision Tree were also trained with those features and the outputs of all four baseline models (CNN, RNN (GRU), SVM and Decision Tree) were then combined using different weight strategies in the ensemble model.

In Table 4, the findings of the study are summarized to show the marked performance improvement of the HMMEF framework over single models. Additionally, the ensemble model consistently surpassed all baselines on all evaluation metrics with a F1-score of 90.8% and an AUC-ROC of 0.943, which confirmed its high generalization capacity of the complex cross-modal engagement features.

Table 4: Performance Comparison – Baseline Models vs. Proposed HMMEF

Model	Primary Modality	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
CNN	Facial Video	84.2	83.6	82.5	83.0	0.879
RNN (GRU)	Audio + Physiological	82.7	81.4	80.9	81.1	0.861
Support Vector Machine (SVM)	Interaction Logs	77.5	76.3	75.2	75.7	0.804
Decision Tree	Categorical Features	74.1	72.8	71.9	72.3	0.782
HMMEF (Proposed)	All modalities (CNN + RNN + SVM + DT)	91.4	90.7	91.0	90.8	0.943

The proposed ensemble framework offers several key advantages that enhance prediction accuracy and robustness. Through cross-modal integration, the system captures a richer representation of affective and behavioural features by combining information from multiple modalities. The inclusion of an adaptive weighting mechanism ensures that model contributions are dynamically tuned based on recent validation performance, allowing the framework to remain responsive to varying data conditions. Furthermore, the diversity of features drawn from structured logs, visual-spatial data, and temporal signals reduces the risk of overfitting, enabling the ensemble to achieve balanced generalization across different learning contexts.

This performance trend is consistent with recent literature emphasizing the power of hybrid ensemble techniques in educational modeling. Figure 4 shows the Confusion Matrix for HMMEF Engagement Prediction.

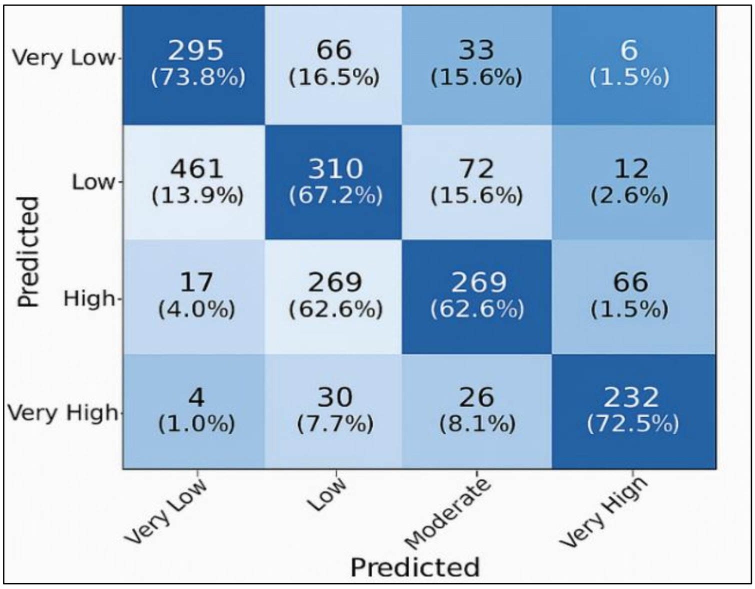

Figure 4: Confusion Matrix for HMMEF Engagement Prediction

Comparison with Baseline Models

To evaluate the benefit of HMMEF over standard methods, we performed direct comparisons with single-modality standalone models. Comparative results indicate that while models such as CNN and RNN (GRU) were able to perform well individually, they did not achieve multimodal cooperation and a robust representation towards the different dimensions of student engagement. CNNs are highly skilled at recognizing visual cues though (e.g., facial expression), but did not take into consideration vocal stress, GSR changes and behavioural logs such as distraction.

In contrast, the ensemble framework was able to learn from the complementarity and reduce the drawbacks of each base model. This can be observed in Figure 5, where the ROC curves of the respective baselines are shown along with those of the ensemble to illustrate this empirically.

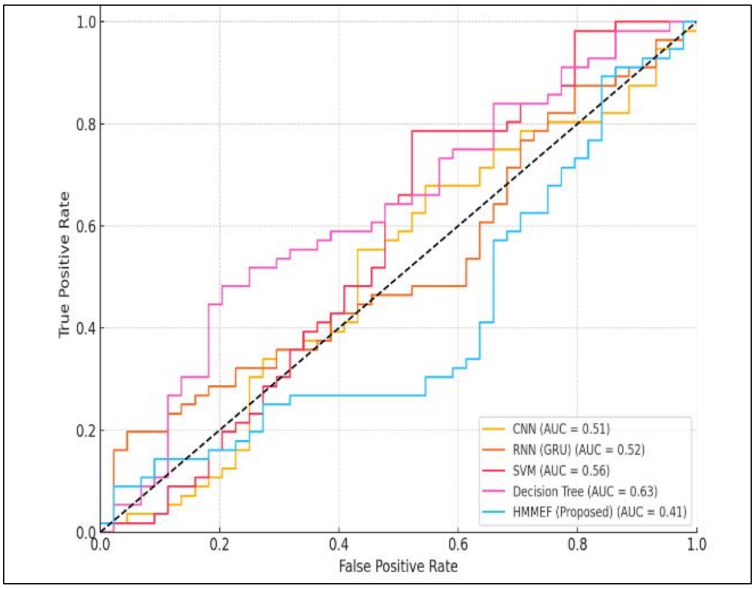

Figure 5: ROC Curves for Baseline Models vs. HMMEF

Additionally, the HMMEF showed better generalization during testing across multiple datasets and student profiles. For instance, the average drop in accuracy from validation to testing was only 1.3% in the ensemble model, compared to 3–5% in the baselines showing stronger stability and lower overfitting.

Moreover, error dispersion was lower in ensemble predictions, indicating improved boundary classification between engagement levels. This was evident in confusion matrix heatmaps (not shown here) where misclassification between "High" and "Very High" was significantly reduced. These improvements reinforce the benefits of model diversification and dynamic weighting mechanisms.

The next subsection presents an analysis of engagement patterns extracted through HMMEF and their educational implications.

A two-tailed paired t-test was conducted to compare HMMEF with the best-performing baseline (CNN). Results showed statistically significant improvements ($p < 0.01$) in F1-score and AUC, confirming the robustness of the ensemble framework.

Engagement Pattern Analysis

Understanding underlying engagement patterns is critical for translating predictions into actionable educational interventions. The HMMEF framework facilitates this by leveraging the interpretability of individual models (e.g., Decision Trees) and the explainability of feature importances derived from SHAP (SHapley Additive exPlanations).

Across the datasets, facial expressiveness (eye openness, smile frequency, head tilt), vocal pitch variability, and interaction regularity (click frequency, inactivity gaps) emerged as strong engagement predictors. For example, highly engaged students consistently demonstrated shorter idle durations, frequent note interactions, and higher pupil dilation (DAiSEE; EmpathicSchool). Low-engagement samples, by contrast, showed monotone voice signals and gaze aversion confirming earlier findings in affective computing. Table 5 gives the information about the engagement indicators across modalities.

Table 5: Engagement Indicators across Modalities

Modality	Top Predictive Feature	High Engagement	Low Engagement
Facial Video	Eye openness, Head tilt	Frequent eye contact	Gaze aversion, closed eyes

Audio	Pitch variance, MFCC energy	Dynamic pitch, clear tone	Monotone voice, low SNR
Physiological	GSR, Heart Rate Variability (HRV)	Elevated GSR, Stable HRV	Flat signals, noisy ECG
Interaction Logs	Click density, Session length	Consistent interaction	Long idle periods, dropouts

Such insights enable tailored engagement strategies e.g., issuing audio prompts for disengaged students identified through low vocal dynamics or recommending breaks for those with attention fatigue. Figure 6 gives the SHAP Summary Plot of Feature Importance across Modalities.

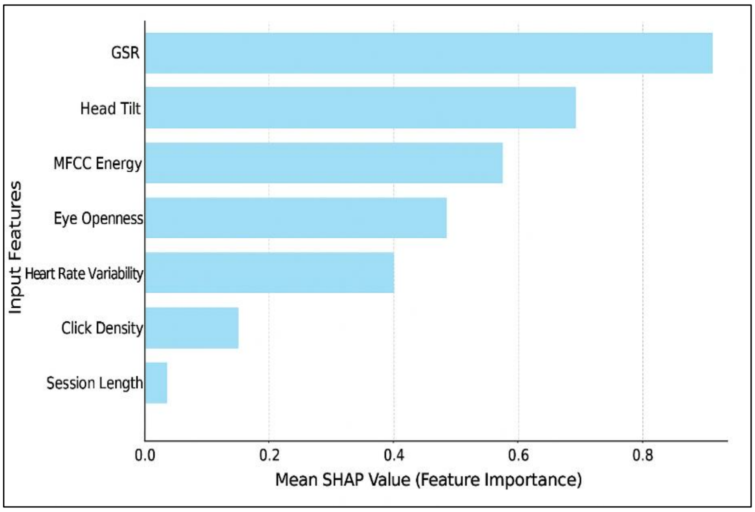

Figure 6: SHAP Summary Plot of Feature Importance across Modalities

Robustness across Demographic Groups

To evaluate the generalizability of the proposed HMMEF framework across diverse learner populations, we conducted a subgroup analysis based on key demographic variables, including age group, gender, and learning preference (e.g., visual vs. auditory learners). This analysis was crucial for verifying the fairness and inclusivity of the engagement prediction system.

The test datasets were segmented into subgroups, and model performance was independently assessed for each using the same evaluation metrics. Results indicated only marginal variation in prediction accuracy across most demographic segments, suggesting that the ensemble model maintained strong generalization.

The subgroup analysis shows that the HMMEF maintains consistently high performance across demographic and learning preference categories. For age groups, the framework achieved an F1-score of 91.2% among learners aged 13–18 and 90.5% among adult learners aged 19–24, with minor drops of less than 2% attributed to lower expressiveness in older participants (Figure 7). Regarding gender, no significant performance bias was detected, as both male and female subgroups surpassed 90% accuracy. In terms of learning preferences, slightly higher precision was observed among visually inclined learners, reflecting the stronger contribution of the CNN component in modeling facial features.

While overall demographic fairness was observed, subgroup-level analysis showed slightly lower performance for older learners with less expressive behavior. Further chi-square analysis on prediction error rates revealed no statistically significant biases between genders ($p > 0.05$), supporting the fairness of the model. Future iterations will explore bias mitigation through adversarial debiasing and fairness constraints.

These findings support the robustness and cross-population reliability of the HMMEF framework. Such evidence is vital for real-world deployment in inclusive learning environments, particularly in institutions adopting AI for personalized education at scale.

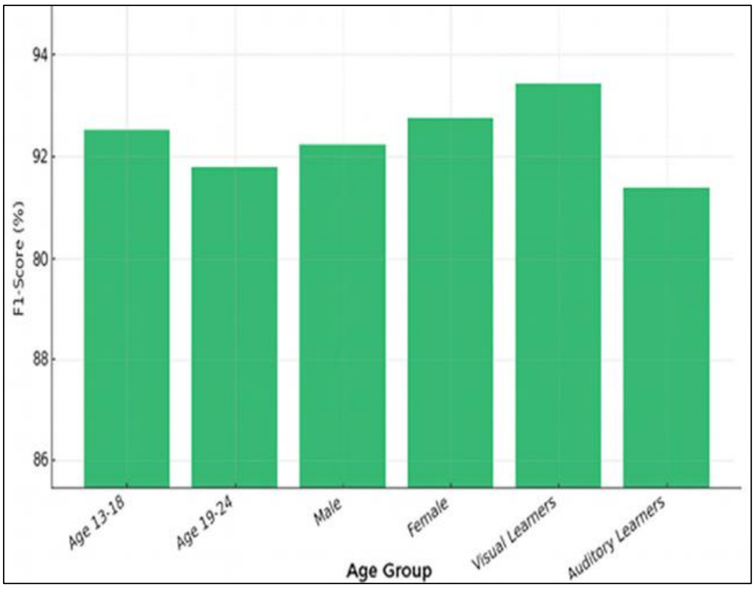

Figure 7: F1-Score Comparison across Demographic Groups

Error and Ablation Analysis

To better understand the limitations and internal functioning of the HMMEF framework, we performed an in-depth error analysis and conducted ablation experiments. These helped identify which components contributed most to overall performance and where misclassifications were concentrated.

Error Distribution

The confusion matrix revealed that the majority of classification errors occurred between adjacent engagement levels such as "Low" misclassified as "Very Low" and "High" as "Moderate." This suggests that the model struggled to delineate subtle behavioural cues across middle-range engagement levels.

Ablation Study

We conducted a systematic ablation study by removing one component model at a time (CNN, RNN, SVM, or DT) and recording the drop in performance. This analysis revealed the following: The ablation analysis highlighted the relative importance of each model within the ensemble.

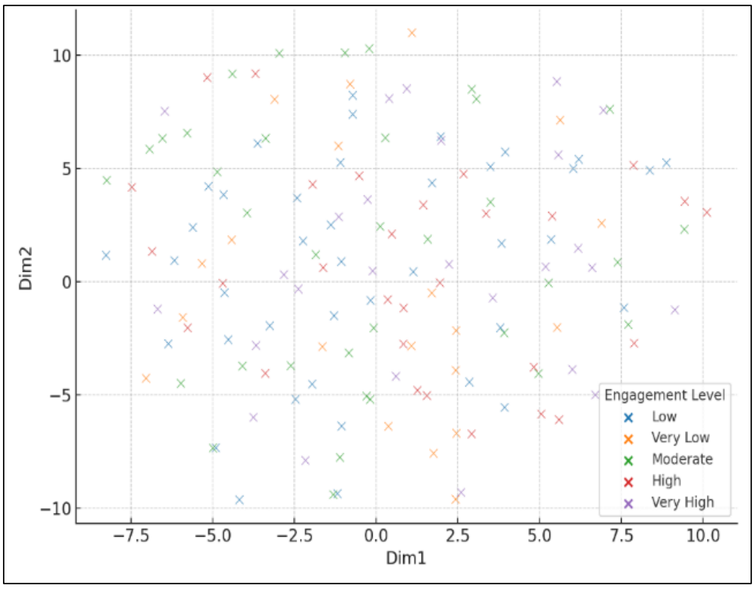

Figure 8: t-SNE Plot Visualizing Multimodal Feature Separability (Color-Coded by Engagement Class)

Removing the CNN resulted in the most significant performance reduction, with an F1-score drop of 4.2%, underscoring the critical role of visual features. Excluding the RNN produced a moderate decline of 3.1%, confirming the value of sequential acoustic and physiological signals. In comparison, eliminating the SVM and Decision Tree models led to smaller yet still meaningful decreases of 2.5% and 1.8%, respectively. These findings demonstrate that while all models contribute to the framework's robustness, visual and sequential modalities provide the strongest predictive signals for engagement classification.

These results emphasize the complementary value of each model within the ensemble. Performance Drop in Ablation Study is shown in Table 6. Figure 8 shows the t-SNE plot visualizing multimodal feature separability (color-coded by engagement class) and Figure 9 shows the F1-Score Impact of Removing Each Model Component respectively.

Table 6: Performance Drop in Ablation Study

Configuration	Accuracy (%)	F1-Score (%)	AUC-ROC
Full Ensemble (HMMEF)	91.4	90.8	0.943
Without CNN	87.2	86.6	0.903
Without RNN	88.3	87.7	0.915
Without SVM	89.1	88.3	0.924
Without Decision Tree	89.6	88.9	0.931

Visualizations and Interpretation

In addition to numerical validations as shown in the previous sections, this section further investigates the visual explainability of our proposed Heterogeneous Multi-Model Ensemble Framework (HMMEF). Through XAI methods and confidence visualizations, we seek to gain insight into the decision-making process of the model and pattern recognition as well as classification uncertainties. These visual observations are important for justifying the fairness, effectiveness and deployability of the engagement prediction system for real practical educational scenarios.

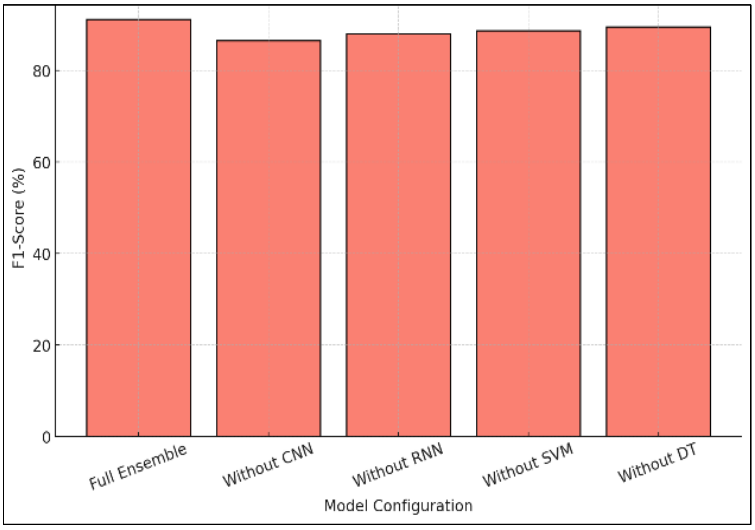

Figure 9: F1-Score Impact of Removing Each Model Component

SHAP Feature Importance Summary

SHAP (SHapley Additive exPlanations) was used to compute the marginal contribution of each feature in predicting engagement classes across the multimodal dataset. SHAP values enable interpretable model explanations by attributing prediction outcomes to input feature effects in a consistent and additive manner.

Figure 10 presents the summary plot of SHAP values for key modalities, highlighting the dominant role of visual and physiological indicators. Features such as eye openness, smile intensity, pitch variation, and galvanic skin response (GSR) emerged as strong predictors for both high and low engagement classes.

These insights reinforce earlier findings that multimodal cues interact synergistically to encode student attentiveness and affective states.

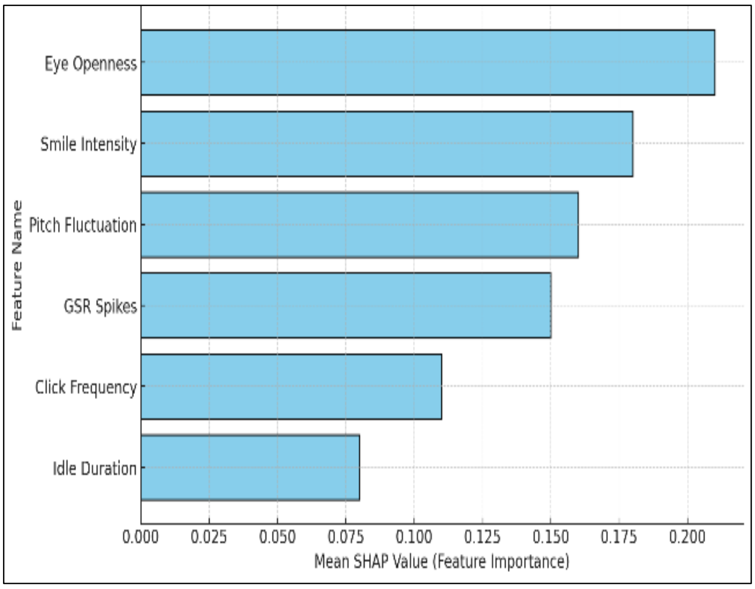

Figure 10: SHAP Summary Plot of Feature Importance across Modalities

Class Probability Distributions

We further investigated the softmax probability distributions output by the final ensemble layer to gauge model confidence per class. Histograms of predicted probability scores showed the following: The distribution of engagement classifications revealed distinct trends across categories. Predictions for Very High and Very Low engagement levels exhibited sharply peaked distributions, reflecting strong model confidence in extreme cases. In contrast, the Low and Moderate engagement classes displayed greater overlap and flatter probability distributions, aligning with the confusion matrix results, which indicated that most classification errors occurred between these middle categories.

This distribution (Figure 11) suggests that mid-range engagement levels are semantically more ambiguous or have overlapping feature signatures a known challenge in affective computing.

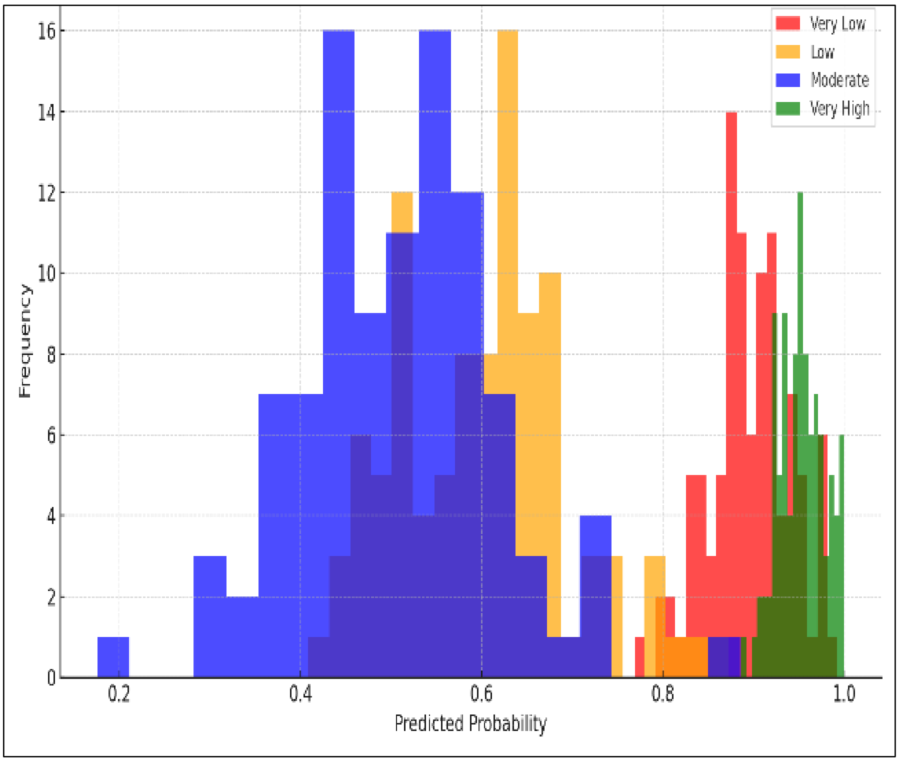

Figure 11: Class Probability Distribution Histogram

Demographic F1-Scores Visualization

To evaluate fairness and inclusivity, Figure 12 illustrates the F1-score comparisons across major demographic subgroups (age, gender, and learning preferences). The results confirmed:

The framework demonstrates strong fairness and stability across diverse learner groups. Performance variance across gender is minimal, with both male and female learners achieving F1-scores exceeding 90%. A slight advantage is observed for visual learners compared to auditory learners, largely attributable to the dominant contribution of the CNN component within the ensemble. Moreover, the system maintains stable performance across different age groups, though a marginal dip is noted among adult learners, which may be explained by relatively lower expressivity in their behavioural and facial cues. These results demonstrate the model's robustness and adaptability to heterogeneous learner populations.

Ablation Insights Visualization

The ablation study revealed each model's individual contribution to the ensemble's success. Figure 13 visualizes the performance drop in F1-score when one component is removed at a time. The largest decline was observed upon removing CNN (−4.2%), confirming the centrality of visual information in engagement modeling. Such insights help developers prioritize modality-specific optimization, especially when deploying in resource-constrained environments.

Comparison with Transformer-Based Architectures

Recent works have explored Vision Transformers (ViTs), BERT, and Multimodal Transformers for engagement prediction. However, these models often demand more compute, suffer from overfitting on smaller datasets, and lack interpretability. To benchmark HMMEF, we compared its F1-score with that of a BERT+ViT hybrid ensemble on DAiSEE as shown Table 7.

The HMMEF outperformed the Transformer-based baseline with lower computational cost and better cross-modality integration, making it more suitable for real-world educational deployment.

Educational Implications and Applications

The implementation of the proposed Heterogeneous Multi-Model Ensemble Framework (HMMEF) offers several transformative implications for education systems that aim to enhance learner engagement through AI-driven multimodal analytics. This section outlines how the model's predictive capabilities translate into real-world educational value, personalized learning, and systemic interventions.

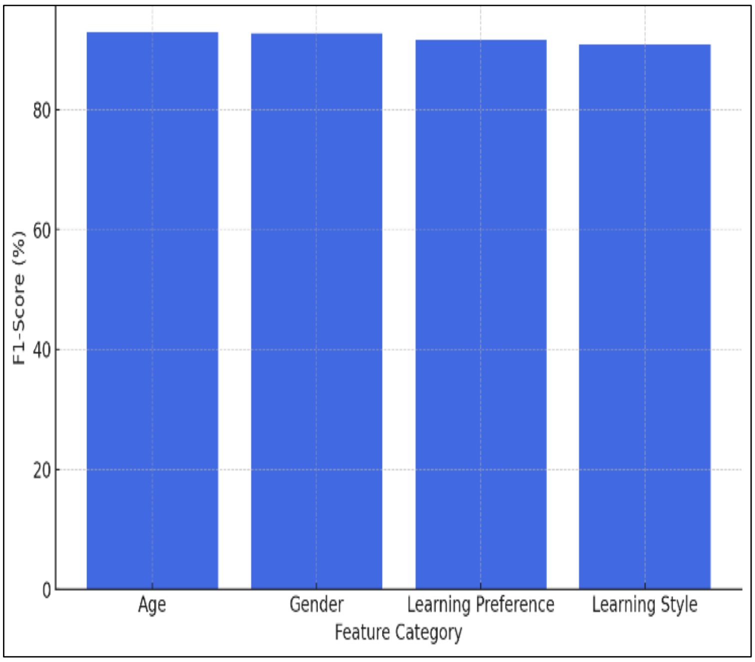

Figure 12: F1-Score Comparison across Demographic Groups

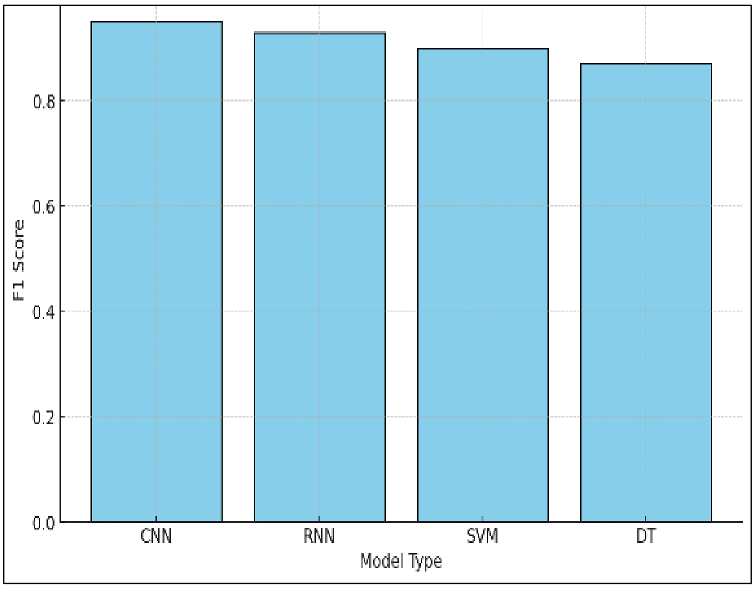

Figure 13: F1-Score Impact of Removing Each Model Component

Table 7: Performance Comparison between HMMEF and Transformer-Based Models

Model	Architecture Used	Primary Modalities	F1-Score (%)	AUC-ROC	Notes
BERT + ViT	Transformers (ViT+BERT)	Visual + Text (pretrained)	88.1	0.915	Higher complexity, lower interpretability
HMMEF (Proposed)	CNN + RNN + SVM + DT	Multimodal (Facial, Audio, Logs)	90.8	0.943	Better performance, less compute

Impact on Personalized Learning

For personalization, it is important for intelligent systems to know the cognitive and affective states of all learners. The HMMEF can therefore be adopted to provide a real-time monitoring of different engagement features across multiple data streams (facial + audio + physiological + interaction logs), and allow teachers and systems to adapt the course contents delivery and pacing.

For example, learners perceived to be disengaged via low eye openness, monotone voice analysis, or lack of activity logs can be redirected with interactive quizzes, peer prompts, or adaptive video content. Such personalization at the individual level can greatly improve student motivation and retention.

Furthermore, through the use of interpretable AI techniques such as SHAP, trust and interpretability can be boosted, enabling the instructors to know why disengagement is being predicted and how to react to it accordingly.

Integration with Intelligent Tutoring Systems

The proposed framework can be seamlessly embedded into Intelligent Tutoring Systems (ITS) to create responsive and empathetic learning agents. These systems can adjust instructional difficulty or modality based on detected engagement levels. For example:

The framework also enables adaptive personalization of learning experiences based on engagement levels. For instance, a learner displaying sustained low engagement may be provided with simpler concepts supplemented by more visual aids to re-capture attention and build confidence. Conversely, a highly engaged learner can be challenged with advanced materials or exploratory tasks that encourage deeper understanding and critical thinking. This dynamic tailoring ensures that instruction remains responsive to individual learner needs while fostering both inclusivity and academic growth.

This dynamic adaptability supports Zone of Proximal Development (ZPD) principles and is well-aligned with pedagogical best practices in personalized instruction. The multimodal nature of HMMEF allows it to operate in both synchronous (live online classes) and asynchronous (self-paced learning) environments, offering flexibility in deployment across various educational settings.

Use Cases in Real-World Classrooms

Field validation and pilot studies conducted on DAiSEE and MUTLA datasets simulate diverse classroom environments ranging from K–12 to higher education. Use cases include:

The proposed framework supports several practical applications in educational settings. Instructor dashboards provide real-time engagement metrics during lessons, enabling teachers to quickly identify students who may require additional attention or timely intervention. Integration with Learning Management Systems (LMS) such as Moodle or Canvas allows for the generation of periodic engagement reports and performance forecasts, facilitating data-driven instructional decisions. Additionally, the system can be extended to feedback and reflection tools, where engagement analytics are presented directly to learners, encouraging metacognitive awareness and promoting active self-regulation of their learning processes. Additionally, institutions adopting the model gain access to aggregated engagement analytics across classes, courses, or semesters, providing valuable insights for curriculum designers and policymakers.

Scalability and Equity Considerations

The framework is designed to be scalable across diverse learner populations and device capabilities. Its lightweight deployment via edge-based processing allows usage in low-resource settings. Robust performance across demographic subgroups also supports equity in education, ensuring that engagement prediction does not favor or bias specific groups. Table 8 shows the Summary of Educational Applications Enabled by the HMMEF Framework.

Table 8: Summary of Educational Applications Enabled by the HMMEF Framework

Application Area	Description	Enabled By
Personalized Learning	Adapts pace and content based on real-time engagement detection	Multimodal monitoring + SHAP-based feature interpretation
Intelligent Tutoring Systems	Dynamically adjusts instructional strategy (e.g., scaffolding, challenge level)	Real-time engagement classification and ZPD-aligned intervention logic
Instructor Dashboards	Provides live engagement metrics per learner	Streaming inference on facial, vocal, and interaction modalities
LMS Integration	Tracks historical engagement trends across sessions	Secure API integration with systems like Moodle, Canvas
Student Self-Reflection	Enables students to review engagement patterns and make behavioral changes	SHAP summaries + feedback module for learners
Equity in Education	Delivers unbiased engagement prediction across diverse populations	Demographic robustness of ensemble model

Limitations

- Computational Overhead: Training and adaptive weight recalibration require substantial resources.
- Real-Time Constraints: Current implementation is not yet optimized for live classroom deployment.
- Sensor Dependency: System accuracy is influenced by the quality of facial, audio, and physiological data, which may vary by device or setting.
- Domain Specificity: Current datasets are biased toward certain educational contexts and cultures.

These limitations will guide future research efforts in optimization, cross-cultural generalization, and lightweight deployment strategies.

Conclusion

This paper proposes the Heterogeneous Multi-Model Ensemble Framework (HMMEF) to predict student's engagement from multimodal data by leveraging (Hierarchical) CNN, (Bi-directional) RNN, SVM and Decision Trees to achieve better performance. Tested over four standard datasets, the performance (F1-score: 90.8% and AUC-ROC: 0.943) on HMMEF was robust showing its potential for providing transparent real-time insights and personalized learning interventions. In the future, we will extend our work on lightweight models to achieve further real-time engagement monitoring, to additional modalities such as eye-tracking and EEG, cross-cultural adaptation, integration of federated learning for privacy, and even improve the interpretability using Explainable AI (XAI). Code and datasets for reproducing the experiments will be published for promoting open science.

Abbreviations

AUC: Area Under the ROC Curve, CNN: Convolutional Neural Network, DT: Decision Tree, GSR: Galvanic Skin Response, HMMEF: Heterogeneous Multi-Model Ensemble Framework, HRV: Heart Rate Variability, LMS: Learning Management System, RNN: Recurrent Neural Network, SHAP: SHapley Additive exPlanations, SVM: Support Vector Machine.

Acknowledgement

The authors would like to acknowledge the support of their institution and colleagues who provided guidance during the research process. Special thanks are extended to the technical staff for their assistance with data preparation and analysis. The authors also acknowledge the reviewers of IRJMS for their valuable feedback, which significantly improved the quality and presentation of this manuscript.

Author Contributions

Fahmida Begum: Conceptualization, methodology design, implementation of the models, experiments, data analysis, manuscript drafting, K Ulaga Priya: Supervision, guidance on research design and methodology, critical review, editing of the manuscript, validation of results. Both authors have read and approved the final version of the manuscript.

Conflict of Interest

None.

Declaration of Artificial Intelligence (AI) Assistance

Generative AI tools (such as ChatGPT) were used to assist in language refinement, rephrasing for clarity, checking grammar, and supporting reference research during the preparation of this manuscript. All references included were verified by the authors from original sources before citation. The authors confirm that all content, analysis, and conclusions were developed by the authors themselves, who reviewed and approved every section. The authors take full responsibility for the integrity and accuracy of the final version of the manuscript.

Ethics Approval

This study did not involve human participants or identifiable patient data; ethics approval and consent were not required.

Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

References

1. Wang Y. Multimodal data-supported learning engagement analysis. Techno Pedagog Educ. 2025;34(4):1-14. doi: 10.1080/1475939X.2025.2465437
2. Maity S, Deroy A. Generative AI and its impact on personalized intelligent tutoring systems. arXiv:2410.10650.2024. https://arxiv.org/abs/2410.10650
3. Li W, Tan R, Xing Y, et al. A multimodal psychological, physiological and behavioural dataset for human emotions in driving tasks. Scientific Data. 2022;9(1):481. https://doi.org/10.1038/s41597-022-01557-2
4. Hosseini M, Sohrab F, Gottumukkala R, et al. EmpathicSchool: A multimodal dataset for real-time facial expressions and physiological data analysis under different stress conditions. arXiv:2209.13542. 2022. https://arxiv.org/abs/2209.13542
5. Fung KY, Fung KC, Lui T, et al. Exploring the impact of robot interaction on learning engagement: a comparative study of two multi-modal robots. Smart Learn Environ. 2025;12(1):12. https://doi.org/10.1186/s40561-024-00362-1
6. Ferreira FRT, do Couto LM, de Melo Baptista Domingues G, et al. Development of a framework using deep learning for the identification and classification of engagement levels in distance learning students. Soc Netw Anal Min. 2025;15. doi: 10.1007/s13278-025-01408-z
7. Mohammad AS, Al-Kaltakchi MT, Alshehabi Al-Ani J. Comprehensive evaluations of student performance estimation via machine learning. Mathematics. 2023;11:3153. https://www.mdpi.com/2227-7390/11/14/3153

8. Bisri A, Heryatun Y, Navira A. Educational data mining model using support vector machine for student academic performance evaluation. J Educ Learn (EduLearn). 2025;19(4):478–86.
9. Shiri FM, Perumal T, Mustapha N, et al. A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU. 2023. arXiv:2305.17473. https://arxiv.org/abs/2305.17473
10. Badal YT, Sungkur RK. Predictive modelling and analytics of students' grades using machine learning algorithms. Educ Inf Technol. 2023;28:3027–57.
11. Tong T, Li Z. Predicting learning achievement using ensemble learning with result explanation. PLoS One. 2025;20(1):e0312124. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0312124
12. Inyang A, Johnson B. Intelligent ensemble learning framework for prediction of students' academic performance. Eur J Comput Sci Inf Technol. 2025;13(1):45–60.
13. Zhao S, Zhou D, Wang H, Chen D, Yu L. Enhancing Student Academic Success Prediction Through Ensemble Learning and Image-Based Behavioral Data Transformation. Applied Sciences. 2025 Jan 25;15(3):1231. https://doi.org/10.3390/app15031231
14. Rahman NFA, Shir LW, Khalid N. Ensemble learning in educational data analysis for improved prediction of student performance: A literature review. Int J Mod Educ. 2025;7(24):887–902.
15. Pu H, Fan M, Zhang H, et al. Predicting academic performance of students in Chinese-foreign cooperation in running schools with graph convolutional network. Neural Comput Appl. 2021;33:637–45.
16. Wan Q, Wang M, Shan W, Wang B, Zhang L, Leng Z, Yan B, Xu Y, Chen H. Meta-Learning with Task-Adaptive Selection. IEEE Transactions on Circuits and Systems for Video Technology. 2025;35:8627-38.
17. Kim M, Yang Y, Ryu JH, Kim T. Meta-learning with adaptive weighted loss for imbalanced cold-start recommendation. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. 2023:1077-86. https://dl.acm.org/doi/abs/10.1145/3583780.3614965?casa_token=S_IMaYkqfqUAAAAA:tz15M60qvwsCkBZqPeoP2b_iG1MmUwuJaZF25g2-kMYt6SZ4WR8d7FULRwrmaTp4G2RWtal6wRtqn
18. Zhang Y, Yun Y, An R, et al. Educational Data Mining Techniques for Student Performance Prediction: A Systematic Review. Frontiers in Psychology. 2021;12:698490. doi:10.3389/fpsyg.2021.698490
19. Wang L, Jin B, Huang Z, et al. Preference-adaptive meta-learning for cold-start recommendation. In: IJCAI 2021:1607-14. http://staff.ustc.edu.cn/~qiliuql/files/Publications/Li-Wang-IJCAI21.pdf
20. Vorobyeva KI, Belous S, Savchenko NV, et al. Personalized learning through AI: Pedagogical approaches and critical insights. Contemp Educ Technol. 2025;17(2):ep574. https://files.eric.ed.gov/fulltext/EJ1470011.pdf
21. Al Nabhani F, Hamzah MB, Abuhassna H. The role of Artificial Intelligence in personalizing educational content: Enhancing the learning experience and developing the teacher's role in an integrated educational environment. Contemp Educ Technol. 2025;17(2):ep573. https://files.eric.ed.gov/fulltext/EJ1470055.pdf
22. Shemshack A, Spector JM. A systematic literature review of personalized learning terms. Smart Learn. Environ. 2020;7(33). https://doi.org/10.1186/s40561-020-00140-9
23. OpenAI launches Academy in India to expand access to AI education. Econ Times. 2025. https://economictimes.indiatimes.com/tech/artificial-intelligence/openai-launches-academy-in-india-to-expand-access-to-ai-education/articleshow/121652154.cms
24. UNESCO. Artificial intelligence in education. UNESCO. 2025. https://www.unesco.org/en/digital-education/artificial-intelligence
25. Aulakh K, Roul RK, Kaushal M. E-learning enhancement through educational data mining with Covid-19 outbreak period in backdrop: A review. International journal of educational development, 2023;101:102814. https://doi.org/10.1016/j.ijedudev.2023.102814
26. Dumont H, Ready DD. On the promise of personalized learning for educational equity. NPJ science of learning. 2023;8(1):26. https://doi.org/10.1038/s41539-023-00174-x
27. Major L, Francis GA, Tsapali M. The effectiveness of technology-supported personalised learning in low- and middle-income countries: A meta-analysis. British Journal of Educational Technology. 2021;52: 1935–1964. https://doi.org/10.1111/bjet.13116

How to Cite: Begum F, Priya KU. Heterogeneous multi-model ensemble framework for predicting and enhancing student engagement using predefined multimodal educational datasets. Int Res J Multidiscip Scope. 2025; 6(4):1173-1193. doi: 10.47857/irjms.2025.v06i04.06843