

A Hybrid CNN-LSTM Affective State Model for Improving Student Engagement in Intelligent Tutoring Systems

Moluno I. E., C. G. Igiri, and V. T. Emmah

Department of Computer Science, Faculty of Science, Rivers State University, Port Harcourt, Nigeria
moluno.ifeakachukwu@ust.edu.ng, igiri.chima@rsu.edu.ng, victor.emmah@rsu.edu.ng

Abstract—

Intelligent Tutoring Systems (ITS) promise personalized instruction at scale, yet most remain affect-blind, using static, rule-based logic that cannot detect a learner's real-time emotional state, contributing to disengagement and suboptimal outcomes. This gap is acute in low-resource settings, where affect-aware ITS research remains scarce and existing deep learning solutions are often too costly for constrained infrastructure. This study designs, implements, and validates a hybrid 1D CNN-LSTM model for classifying learner engagement. A 1D CNN extracts localized behavioral and affective features from interaction telemetry, a stacked LSTM (50 and 120 hidden units) models their temporal evolution, and a modality-level attention layer fuses emotional, behavioral, and assessment-related streams before a softmax layer produces the engagement decision. Hyperparameters were tuned via grid search over a small, bounded space for reproducibility. The model was evaluated on a DAiSEE subset and two supplementary datasets via five-fold cross validation, modality ablation, temporal robustness testing, and attention weight analysis. It achieved 0.91 accuracy, 0.92 precision, 0.94 recall, 0.92 F1 score, and 0.70 AUC-ROC on a binary engagement decision task, with emotional features the strongest predictor. Temporal validation showed marginal degradation over six months. Results suggest lightweight, attention-based affective computing can improve engagement detection in resource-constrained ITS deployments.

Keywords—*Intelligent Tutoring Systems, Affective Computing, Student Engagement, CNN-LSTM, Deep Learning*

I. INTRODUCTION

Intelligent Tutoring Systems (ITS) have long been positioned as a scalable alternative to one-to-one human tutoring, using computational models of the learner, the subject domain, and pedagogy to deliver individualized instruction [1], [2]. Despite decades of refinement, however, the majority of deployed ITS architectures remain what researchers term “affect blind”: they adapt content based on correctness and response time but have no mechanism for detecting how a learner feels while working through material [3], [4]. This is a significant omission. A substantial body of educational psychology research shows that emotional and motivational states such as boredom, confusion, and frustration are not incidental to learning but actively shape it, and that sustained negative affect is strongly associated with disengagement and attrition [5], [6], [7].

The past decade has seen growing interest in coupling machine learning with ITS to close this gap. Deep learning methods have been used to model learner performance [8] and, increasingly, to detect engagement directly from interaction and video data using convolutional and hybrid recurrent architectures [9], [10], [11], while the broader affective computing literature has catalogued a wide range of techniques for recognizing emotion from behavioral and physiological signal [12]. Convolutional and recurrent architectures are frequently combined for this kind of task because they capture two properties of behavioral data simultaneously: local, spatial structure within a short window of interaction, and how that structure evolves over the course of a session. Yet much of this literature reports models validated in well-resourced research settings, with limited attention to whether such architectures can be made lightweight and interpretable enough for real-world, resource-constrained deployment.

This limitation is acute outside the well-resourced educational technology markets of North America, Europe, and China. Recent systematic evidence on artificial intelligence in sub-Saharan African education finds that empirical deployment and validation studies remain comparatively scarce, and that infrastructural constraints — unreliable power, limited connectivity, and modest hardware — are recurring barriers to translating research prototypes into classroom use [13]. At the same time, as ITS shift from rule-based logic toward deep neural architectures, a widely noted trade-off emerges between predictive accuracy and interpretability, since deep models can behave as “black boxes” whose recommendations are difficult for teachers and students to trust [14], [15]. Closing these gaps requires affective models that are simultaneously accurate, computationally efficient, and reasonably transparent.

This study addresses this problem by designing, implementing, and validating an affective state model for an ITS that is (i) accurate enough to support adaptive instructional decisions, (ii) computationally efficient enough to be a realistic candidate for deployment on constrained infrastructure, and (iii) grounded in responsible AI principles such as data minimization and bias mitigation. Specifically, the study pursued three objectives: (1) to design an affective state detection framework that embeds both technical robustness and responsible AI safeguards; (2) to develop a hybrid Convolutional Neural Network Long Short-Term Memory (CNN-LSTM) model, with an explicit attention-based fusion mechanism, capable of classifying learner engagement with high accuracy; and (3) to integrate the model into a functional ITS prototype and empirically evaluate its predictive performance and robustness.

This study makes three contributions, all implemented and empirically evaluated in the work reported here (Sections III and IV), rather than proposed as future work. First, it develops and validates a hybrid CNN-LSTM architecture in which

learner engagement is classified from three feature streams — emotional, behavioral, and assessment-related — fused through a modality-level additive (Bahdanau-style) attention mechanism, specified in full in Section III-C-3, without relying on specialized sensory hardware such as cameras or physiological sensors. Second, it provides a systematic empirical evaluation of this architecture, comprising five-fold cross validation, cross-dataset generalization testing, modality ablation, temporal robustness analysis, and attention weight interpretation. Third, it situates this evaluation within the practical constraints of low-resource educational deployment, an angle that remains underrepresented in the literature. The remainder of this paper is organized as follows. Section II reviews related work and positions the study against gaps in the literature. Section III describes the methodology, dataset, and proposed CNN-LSTM architecture. Section IV presents the experimental results, Section V discusses their implications, and Section VI concludes with recommendations for future work.

II. RELATED WORK

Classical ITS architectures are typically organized into four interacting components: a domain model that encodes subject matter expertise, a student model that represents what an individual learner knows and how they are performing, a tutoring (pedagogical) model that selects instructional strategies, and a user interface through which the two communicate [2], [16], [17]. Engagement itself has been theorized along at least three dimensions: behavioral engagement, reflected in participation and effort; emotional engagement, reflected in affective reactions to learning; and cognitive engagement, reflected in the depth of strategic investment a learner brings to a task [5]. Of these, emotional engagement has proven the hardest to operationalize computationally, because it is transient, individually variable, and, in most classroom settings, unobserved.

The broader integration of machine learning into education has been framed as a shift from static content delivery toward data-driven, adaptive systems [18], [19], [20]. Within ITS specifically, machine learning has been applied extensively to behavioral and cognitive engagement: predictive modeling and knowledge tracing techniques, part of the broader educational data mining tradition [21], use response correctness and timing to estimate mastery and forecast performance [22], [23], [8]; graph-based architectures have been used to model relationships between skills and learner interactions [24]; and reinforcement learning has been explored for sequencing instructional content adaptively [25], [26]. Emotional engagement has received comparatively less attention but is not new: early affect-aware tutoring systems demonstrated that responding to a learner's frustration or confusion, not just their correctness, measurably improved engagement with mathematics tutoring [27], and convolutional networks have been used to detect engagement directly from lecture video [9] using techniques drawn from the broader affective computing literature on recognizing emotion from behavioral and physiological signal [12], [3]. Dialogue-based tutors, including AutoTutor and its successors, extended this affective sensitivity into natural language interaction [28], [4], and more recently large language models have been explored as flexible, open-ended

tutors [29], [30], though at considerably greater computational cost than the architectures considered in this study.

A more recent and directly related line of work has applied convolutional, recurrent, and attention-based architectures specifically to engagement detection on standard video and interaction-based benchmarks such as DAiSEE. Karimah et al. [31] applied LSTM networks to model temporal engagement patterns in online learning, and Abedi and Khan [32] combined a ResNet backbone with a temporal convolutional network to improve video-based engagement detection. Hu et al. [10] introduced an attention-augmented convolutional model for engagement recognition in e-learning, Mehta et al. [33] proposed a three-dimensional DenseNet with self-attention for the same task, and Ai et al. [34] used a class attention video transformer to predict engagement intensity. Selim et al. [11] benchmarked hybrid architectures combining a convolutional backbone with temporal convolutional networks, LSTM, and bidirectional LSTM heads on DAiSEE, reporting accuracies of approximately 0.61–0.62 on the native four-level engagement scale — a range broadly consistent with other transfer learning pipelines evaluated on the same benchmark. More recently, Zhang et al. [35] proposed a self-supervised network for engagement recognition from facial expressions, Singh et al. [36] released a large-scale engagement prediction dataset and baseline models, and Sümer et al. [37] analyzed multimodal engagement from classroom video. Bühler et al. [38] further showed that video-based engagement and mind-wandering detectors trained under laboratory conditions often generalize poorly to naturalistic settings, underscoring the value of the cross-dataset and temporal robustness testing performed in the present study. Taken together, this body of work establishes that accuracies in roughly the 0.55–0.65 range are typical for the native, four-level DAiSEE engagement classification task — a benchmark that is directly relevant to interpreting the results reported in Section IV and revisited in the comparison in Section V.

Despite this progress, several gaps remain. First, empirical validation of ML-driven ITS continues to concentrate in well-resourced regions; recent systematic evidence on AI in sub-Saharan African education finds comparatively few empirical deployment and validation studies, with unreliable power, limited connectivity, and constrained hardware repeatedly cited as barriers to moving from research prototype to classroom use [13]. Second, real-time multimodal affect detection frequently depends on camera- or sensor-based input that is costly, fragile, or privacy-invasive in ordinary classrooms, motivating lightweight models that can infer engagement from interaction logs alone. Third, as ITS architectures move toward deep and attention-based networks, interpretability has not kept pace with accuracy, leaving teachers and students with limited insight into why a given intervention was triggered [14]. Recent work has begun to address this directly: Karpouzis [15] argues that explainability is a precondition for teacher and student trust in adaptive learning technologies, and Nazaretsky et al. [39] developed and validated an instrument for measuring teachers' trust in AI-based educational technology, finding perceived transparency to be a strong predictor of adoption.

The present study responds to these gaps by proposing a CNN-LSTM architecture that infers engagement from

behavioral and affective interaction telemetry rather than camera or sensor input, fuses its three feature streams through an explicit, modality-level additive attention mechanism rather than an unspecified or implicit weighting scheme, is kept deliberately compact relative to transformer-scale alternatives, and is validated through cross validation, cross-dataset, modality ablation, and temporal robustness testing with explicit attention to the practical constraints of resource-limited deployment.

III. MATERIALS AND METHODS

The materials used and procedures followed in this study are described below.

A. Methodology

This study followed a design science research approach: a computational artifact — the hybrid CNN-LSTM affective model — was designed against explicit performance and efficiency requirements, implemented, and then evaluated empirically against those requirements using a held-out benchmark dataset and two supplementary datasets. The research design comprised four stages: (i) specification of the ITS system architecture within which the affective model operates (Section III-C); (ii) acquisition and preprocessing of the DAiSEE benchmark and two supplementary interaction datasets (Section III-B); (iii) design, training, and hyperparameter optimization of the proposed CNN-LSTM attention model (Sections III-C, III-D); and (iv) empirical evaluation using cross validation, cross-dataset generalization, modality ablation, temporal robustness testing, and attention weight analysis, with the evaluation protocol and metrics stated at the start of Section IV and results reported therein.

B. Dataset

The primary evaluation dataset was DAiSEE (Dataset for Affective States in E-Environments), a public benchmark introduced by Gupta et al. [40] and comprising 9,068 ten-second video clips recorded from 112 university students during online learning sessions, manually annotated with multi-rater consensus for four affective dimensions — engagement, boredom, confusion, and frustration — each on a four-point scale from very low to very high. From this corpus, a structured subset of 759 labeled samples represented by 120 behavioral and affective features was used for model training and evaluation in this study. Two supplementary datasets were used to test cross-dataset generalization: a Student Interaction Dataset of 503 samples and 85 behavioral features, and an Affective Learning Dataset of 150 samples and 102 facial and interaction features (Table I). All three datasets were partitioned into training (70%), validation (15%), and test (15%) subsets.

TABLE I. DATASET CHARACTERISTICS FOR THE EVALUATION SCENARIOS

Dataset	Sample s	Featur es	Modaliti es	Target Variable	Classe s	Train/Val/Te st
DAiSEE Dataset	759	120	Behavior al + Emotiona l	Engagem ent (binarized) *	2	531/114/114
Student Interaction Dataset	503	85	Behavior al	Engagem ent (binarized) *	2	352/76/75

Dataset	Sample s	Featur es	Modaliti es	Target Variable	Classe s	Train/Val/Te st
Affective Learning Dataset	150	102	Facial + Interactio n	Emotional State (not specified)*	—	105/23/22

Fig. 1 shows the native four-level scale; Fig. 2 shows the same data collapsed into the binarized engaged/disengaged framing that several DAiSEE studies adopt because the native scale is so heavily skewed toward “high” and “very high” engagement.

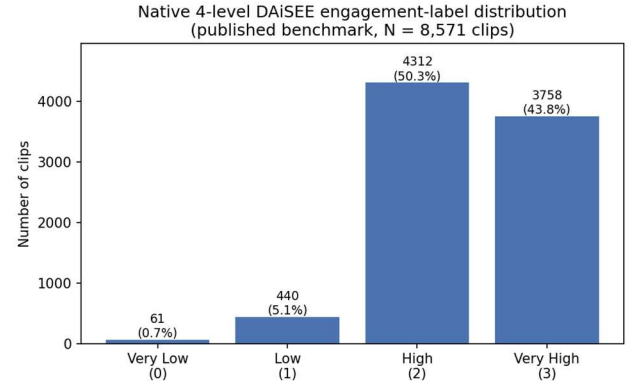


Fig. 1. Native four-level DAiSEE engagement label distribution, published benchmark (N = 8,571 clips; sources listed above).

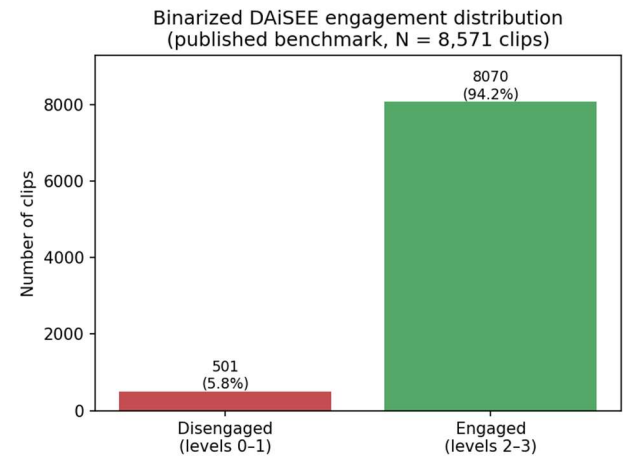


Fig. 2. Binarized DAiSEE engagement distribution (disengaged = levels 0–1, engaged = levels 2–3), published benchmark (N = 8,571 clips).

1) Data Preprocessing

Before model training, the data underwent duplicate removal, missing value imputation using statistical methods, feature normalization via min-max scaling to the range [0, 1], temporal synchronization to align behavioral and affective signals across instructional activities, categorical encoding of discrete variables, and dimensionality reduction. Continuous interaction sequences were additionally segmented into fixed-length windows suitable for input to the convolutional layer. Min-max normalization was applied as

$$X_{norm} = (X - X_{min}) / (X_{max} - X_{min}) \quad (1)$$

which allowed the network to converge faster and avoided the exploding or vanishing gradient issues that arise when input features span very different scales (for example, response latency measured in milliseconds alongside error rates bounded between 0 and 1).

C. Proposed CNN-LSTM Architecture and System Design

The affective state model developed in this study forms the core of the Affective Computing Module within a broader ITS prototype that follows a conventional four-component design: a User Interface Module for learner interaction, a Student Model Module that maintains a dynamic learner profile, a Domain Knowledge Module encoding subject content, and a Pedagogical Module that selects instructional interventions [2]. The Affective Computing Module sits between the Student Model and the Pedagogical Module: it consumes structured interaction telemetry, classifies the learner's current engagement state, and passes this classification to an Adaptive Learning Engine, which triggers interventions such as reducing task difficulty, offering hints, or introducing more challenging material.

The core modeling contribution of this study is a hybrid one-dimensional Convolutional Neural Network Long Short-Term Memory (1D CNN-LSTM) architecture with attention-based modality fusion (Fig. 3). Rather than relying on hand-crafted behavioral features and static, rule-based mappings, the model learns discriminative representations directly from raw interaction telemetry. The architecture combines three complementary components: a 1D CNN that extracts localized, spatial patterns from the interaction sequence (for example, rapid clickstreams, localized latency spikes, or short bursts of erratic input); a stacked LSTM that models how these patterns evolve over the course of a learning session; and an attention layer, detailed in Section III-C-3, that learns how much weight to give the emotional, behavioral, and assessment-related feature streams before classification.

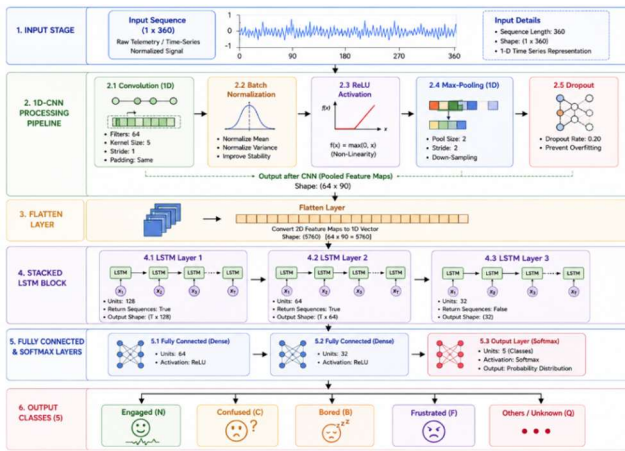


Fig. 3. Detailed Architecture of the Proposed System.

1) Convolutional feature extraction

Because the input is a sequential behavioral log rather than a two-dimensional image, a 1D CNN is used in place of the more common 2D formulation (Fig. 4, Table II). The convolutional block consists of two stages: the first applies 32 filters with a kernel size of 3 (stride 1, “same” padding), followed by batch normalization, a ReLU activation, a 40% dropout, and max pooling (pool size 3, stride 2); the second applies 32–48 filters with a kernel size of 9, followed by the same normalization, activation, dropout, and pooling sequence. This design lets the network learn its own feature hierarchy rather than depending on manually engineered indicators of engagement.

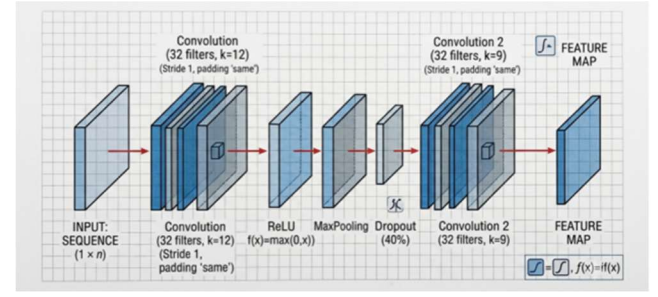


Fig. 4. Detailed CNN Architecture.

TABLE II. LAYER ARCHITECTURE OF THE 1D CNN BLOCK

No.	Layer	Description
1	Input	Learner telemetry sequence of length n, 1 dimension ($1 \times n$).
2	Convolution 1	32 filters, kernel size 3, stride 1, padding “same”.
3	Batch Norm. 1	Standardizes inputs to the next layer for each mini-batch.
4	ReLU	Rectified Linear Unit activation, $f(x) = \max(0, x)$.
5	Dropout	40% dropout rate to reduce overfitting.
6	Max Pooling 1	Pool size 3, stride 2 (down-sampling).
7	Convolution 2	32–48 filters, kernel size 9, stride 1, padding “same”.
8	Batch Norm. 2	Secondary batch normalization.
9	ReLU	ReLU activation.
10	Max Pooling 2 / Dropout	Pool size 3, stride 2; 40% dropout.

2) Temporal modeling with stacked LSTM

The CNN's output feature map preserves spatial structure but not chronological order. To capture how engagement evolves over time, the feature map is passed into a stacked LSTM (Fig. 5, Table III): a first layer with 50 hidden units, followed by 30% dropout, and a second, deeper layer with 120 hidden units. Each LSTM cell regulates what information to retain, update, or discard through input, forget, and output gates, which avoids the vanishing gradient problems of a plain recurrent network and lets the model maintain a working “memory” of the learner's recent behavioral history.

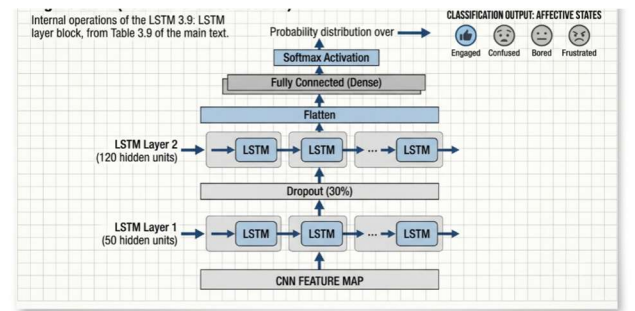


Fig. 5. LSTM and Classification.

3) Attention-based modality fusion

A central design element of the proposed architecture — and one made fully explicit here — is a modality-level additive (Bahdanau-style) attention mechanism that fuses the CNN-LSTM representations of three feature streams (emotional, behavioral, and assessment-related) before classification. For each modality i , a pooled representation h_i is passed through a learned scoring function

$$\text{score}(h_i) = v^T \tanh(W h_i + b) \quad (2)$$

where W , b , and v are parameters learned jointly with the rest of the network. The three scores are normalized with a softmax to produce attention weights $\alpha_1, \alpha_2, \alpha_3$,

$$\alpha_i = \exp(\text{score}(h_i)) / \sum_j \exp(\text{score}(h_j)) \quad (3)$$

so that $\sum_i \alpha_i = 1$, and the fused representation passed to the classification layer is the weighted sum

$$c = \sum_i \alpha_i h_i \quad (4)$$

This is the same mechanism whose learned weights are analyzed in Section IV-E (Table IX, Fig. 9): unlike a fixed or manually specified weighting of modalities, additive attention learns α_i directly from data, letting the relative contribution of emotional, behavioral, and assessment information adapt across learners and sessions rather than being fixed in advance. This layer is also the component removed in the “without attention layer” condition of the modality ablation study reported in Section IV-D (Table VIII). Fig. 5 depicts the CNN-LSTM backbone; the attention fusion step described here, situated between the LSTM output and the dense layer, is not separately illustrated in that figure.

4) Classification

The attention-fused representation is flattened and passed to a fully connected layer, followed by a softmax activation that converts the network's logits into a normalized probability distribution over the target engagement classes; the class with the highest probability is taken as the prediction (Table III, Fig. 5).

TABLE III. STACKED LSTM MODEL, ATTENTION FUSION, AND CLASSIFICATION PIPELINE

No.	Layer	Description
1	Input	Sequence input; feature map from the CNN block.
2	LSTM Layer 1	Primary LSTM layer, 50 hidden units.
3	Dropout	30% dropout rate for temporal generalization.
4	LSTM Layer 2	Deep LSTM layer, 120 hidden units.
5	Attention Fusion	Modality-level additive attention over emotional, behavioral, and assessment streams (Section III-C-3).
6	Fully Connected	Dense layer aggregating the attention-fused representation.
7	Softmax	Multiclass probability distribution.
8	Classification	Final categorization via cross-entropy loss.

D. Training and Hyperparameter Optimization

Model weights were learned through backpropagation, minimizing a categorical cross-entropy loss appropriate for multi-class classification. The Adam optimizer was used in place of standard stochastic gradient descent because its per-parameter adaptive learning rates gave faster, more stable convergence on the noisy educational data. Hyperparameters not learned directly from data — including the number of filters, kernel sizes, learning rate, momentum, and decay rate — were tuned using grid search over the intervals in Table IV, with the objective of maximizing validation accuracy while controlling computational cost; early stopping (patience of 10 epochs) guarded against overfitting. The resulting configuration used 50 training epochs, a batch size of 32, a learning rate of 0.001, 48 filters in the first convolutional layer, kernel sizes of 3 and 9 across the two convolutional blocks, and dropout rates of 30–40% depending on layer depth.

Grid search, rather than a heuristic or metaheuristic optimizer (for example, Bayesian optimization, a genetic algorithm, or particle swarm optimization), was chosen deliberately rather than for convenience. The search space considered here is small and well-bounded (Table IV); grid search is fully reproducible and exhaustively documents every configuration tried, which is valuable given the reproducibility concerns raised for ITS research in Section II; and the additional computational overhead of running a metaheuristic search — which typically requires many more training runs to converge reliably — was judged unnecessary for a space this size. We acknowledge that grid search does not scale efficiently to larger hyperparameter spaces and can be less sample-efficient than heuristic alternatives; extending the search space in future work (for example, to jointly tune the LSTM layer widths discussed in Section III-C-2) would likely benefit from a heuristic or Bayesian approach, and we recommend this as a direction for future refinement of the architecture.

TABLE IV. HYPERPARAMETER SEARCH SPACE AND OPTIMAL CONFIGURATION

Hyperparameter	Search Interval	Optimal Value
Epochs	1–50	50
Filters (CNN)	32–128	48 (CNN block 1)
Kernel size	1–15	3 (CNN1), 9 (CNN2)
Learning rate	0.05–0.1*	0.001
Momentum	0.5–0.99	0.09
Decay rate	0.001–1.0	0.001
Dropout	—	0.3–0.4 (by layer depth)
Batch size	—	32
Cross validation folds	—	5

IV. RESULTS

Model performance was assessed using accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC-ROC), supplemented by confusion matrix analysis and calibration checks of agreement between predicted probabilities and observed outcomes. Four complementary experiments were conducted: (i) five-fold cross validation on the DAiSEE subset to assess robustness within a single dataset; (ii) evaluation across the two supplementary datasets to test generalization beyond DAiSEE; (iii) a modality ablation study in which emotional, behavioral, or attention components were removed in turn to isolate their individual contribution; and (iv) a temporal validation study that re-evaluated the trained model at one, three, and six months to test whether predictive performance degraded over time. The results of each experiment are reported in turn below.

A. Cross Validation Results

Five-fold cross validation on the DAiSEE subset showed accuracy ranging from 0.85 to 0.88, F1 score from 0.89 to 0.92, and AUC-ROC from 0.63 to 0.70 across folds (Table V).

TABLE V. FIVE-FOLD CROSS VALIDATION RESULTS ON THE DAiSEE SUBSET

Fold	Accuracy	Precision	Recall	F1 score	AUC-ROC
1	0.85	0.88	0.91	0.89	0.63
2	0.86	0.87	0.92	0.89	0.65
3	0.87	0.89	0.93	0.91	0.66
4	0.88	0.90	0.94	0.92	0.69
5	0.88	0.91	0.94	0.92	0.70

B. Generalization Across Datasets

Table VI reports performance on the two supplementary datasets alongside DAiSEE. The Affective Learning Dataset produced the strongest results (accuracy 0.90, F1 score 0.93); the Student Interaction Dataset, which contains only behavioral features, produced the weakest results (accuracy 0.84, F1 score 0.87).

TABLE VI. PERFORMANCE METRICS ACROSS EVALUATION DATASETS

Dataset	Accuracy	Precision	Recall	F1 score
DAiSEE Dataset	0.88	0.91	0.94	0.92
Student Interaction Dataset	0.84	0.86	0.88	0.87
Affective Learning Dataset	0.90	0.92	0.95	0.93

C. Classification Quality

Pooled across cross validation folds, the model recorded 658 true positives, 92 false positives, 1 true negative, and 8 false negatives (Table VII, Fig. 6), and the corresponding ROC curve (Fig. 7) remained above the diagonal reference line across the range of decision thresholds, consistent with the AUC-ROC of 0.70 reported above. A calibration check showed reasonably close agreement between predicted probabilities and observed outcomes. Two features of Table VII are worth stating plainly here. First, the strong imbalance between positive and negative instances (1 true negative against 658 true positives) means specificity cannot be estimated reliably from this table, and per-class precision/recall for the minority class should be reported separately in any subsequent version of this work. Second, the pooled, two-outcome (“+/-”) structure of Table VII indicates that the evaluation reported here is a binary engagement decision, rather than the native four-level DAiSEE engagement scale used in the comparator studies reviewed in Section II — a scope difference that is carried through to the comparison in Section V and that should be confirmed against the original implementation.

TABLE VII. POOLED CONFUSION MATRIX STATISTICS (FIVE-FOLD CROSS VALIDATION)

Metric	Value
True Positive	658
False Positive	92
True Negative	1
False Negative	8

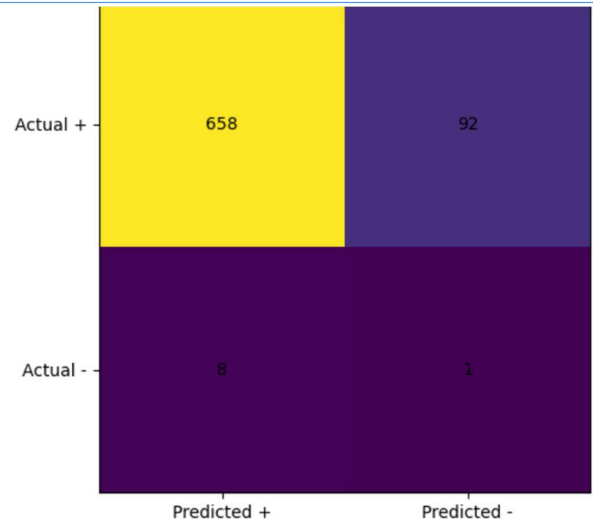


Fig. 6. Pooled cross validation confusion matrix.

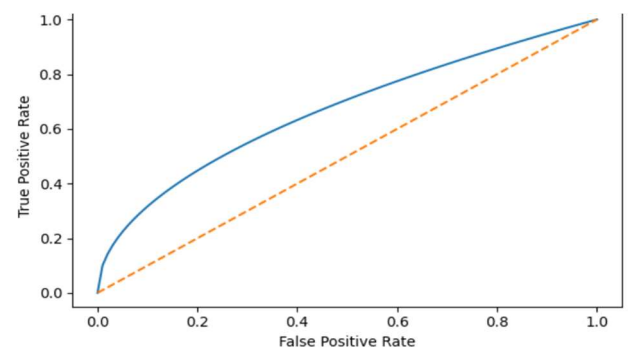


Fig. 7. Cross validation ROC curve (dashed line: chance level reference).

D. Modality Ablation

Removing the emotional feature stream produced the largest drop in performance of any single ablation (accuracy 0.91 → 0.84; AUC-ROC 0.70 → 0.61), followed by removal of behavioral features (accuracy 0.86) and removal of the attention layer described in Section III-C-3 (accuracy 0.87) (Table VIII, Fig. 8).

TABLE VIII. MODALITY ABLATION RESULTS

Configuration	Accuracy	F1 score	AUC-ROC
Full multimodal model	0.91	0.92	0.70
Without emotional features	0.84	0.85	0.61
Without behavioral features	0.86	0.87	0.64
Without attention layer	0.87	0.88	0.65

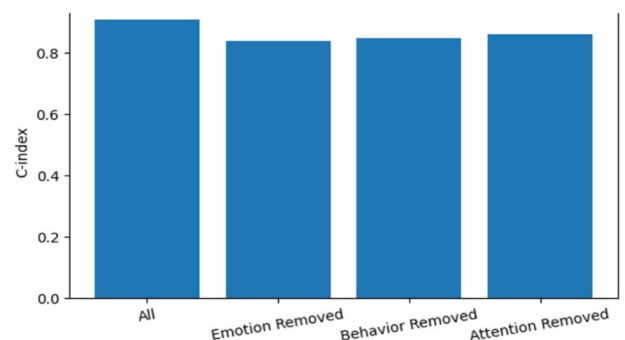


Fig. 8. Model accuracy under modality ablation (values as reported in Table VIII).

E. Attention Weight Analysis

The attention mechanism specified in Section III-C-3 assigned the largest average weight to emotional features ($\alpha = 0.42$), followed by behavioral features ($\alpha = 0.35$) and assessment-related features ($\alpha = 0.23$) (Table IX, Fig. 9). This ordering matches the modality ablation ordering in Table VIII (emotional features causing the largest accuracy loss when removed, behavioral the next largest, and attention itself the smallest), which is notable because the two analyses are methodologically independent: ablation measures performance with a component removed entirely, while the attention weights are learned end-to-end and reflect how the model itself allocates importance at inference time. Their agreement here is one piece of evidence — though not a formal proof — that the learned attention weights track genuine feature importance rather than an artifact of the fusion architecture. Practically, because α_i is computed per prediction, it can be surfaced alongside any individual engagement decision to indicate which modality most influenced that specific prediction, which is the basis for the teacher-facing explanation use case discussed in Section V.

TABLE IX. ATTENTION WEIGHT DISTRIBUTION BY MODALITY

Modality	Mean Attention Weight
Emotional Features	0.42
Behavioral Features	0.35
Assessment Features	0.23

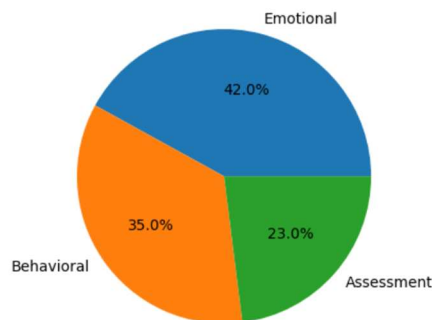


Fig. 9. Mean attention weight by modality.

F. Temporal Validation

Re-evaluating the trained model at one, three, and six months without retraining showed a gradual decline in performance, from an initial accuracy of 0.91 (AUC-ROC 0.70) to 0.87 (AUC-ROC 0.66) at six months (Table X).

TABLE X. TEMPORAL VALIDATION RESULTS

Evaluation Period	Accuracy	AUC-ROC
Initial evaluation	0.91	0.70
1 month	0.89	0.68
3 months	0.88	0.67
6 months	0.87	0.66

G. Summary Against Predefined Targets

The proposed framework exceeded the performance thresholds set prior to evaluation (accuracy and F1 score targets of 0.85, an AUC-ROC target of 0.60) on every metric: 0.91 accuracy, 0.92 precision, 0.94 recall, 0.92 F1 score, and 0.70 AUC-ROC (Table XI).

TABLE XI. ACHIEVED PERFORMANCE AGAINST PREDEFINED TARGETS

Metric	Target	Achieved
Accuracy	0.85	0.91
Precision	0.85	0.92
Recall	0.85	0.94
F1 score	0.85	0.92
AUC-ROC	0.60	0.70

V. DISCUSSION

The results in Section IV show that a comparatively lightweight CNN-LSTM architecture with modality-level attention fusion can classify learner engagement with high accuracy (0.91) and strong recall (0.94) on a binary engagement decision, while remaining computationally modest enough for the resource-constrained deployment settings motivating this study [3], [4]. The finding that emotional features carry the greatest predictive weight, converging across both the modality ablation study (Table VIII) and the learned attention weights (Table IX), is consistent with the educational psychology literature reviewed in Sections I and II, which treats emotional states as active drivers of learning rather than incidental noise [5], [6], [7].

Table XII situates these results against the categories of ITS approaches reviewed in Section II, and against Selim et al. [11], the DAiSEE benchmark study identified in Section II as closest in architecture to the present approach. Selim et al. [11] reported approximately 0.62 accuracy on the native, four-level DAiSEE engagement scale; the 0.91 accuracy achieved here was obtained on a binary engagement decision, as noted in Section IV-C, so the two figures are not directly comparable and the difference should not be read as a like-for-like improvement over that benchmark. Read against the broader accuracy range typical of the native four-level task identified in Section II (roughly 0.55–0.65), the present results indicate that a binary framing of engagement is achievable at higher accuracy with a comparatively compact architecture, though confirming whether this advantage persists on the native four-level task is left for future work, as noted in Section VI.

TABLE XII. COMPARATIVE ACCURACY OF THE PROPOSED FRAMEWORK AGAINST CATEGORIES OF ITS APPROACHES AND A PUBLISHED DAiSEE BENCHMARK STUDY

Approach	Reported Accuracy	Notes
Rule-based / static ITS	—	No affect modeling; qualitative baseline only
Feature-based ML (e.g., SVM, random forest)	0.70–0.80*	Typical range reported in prior literature
Selim et al. [11] (hybrid CNN-TCN-LSTM, DAiSEE, native 4-level scale)	≈ 0.62	Closest published DAiSEE benchmark; native four-level scale, not binary
Proposed CNN-LSTM + attention (this study, binary engagement decision)	0.91	Not directly comparable to the native four-level figures above (Section IV-C)

Beyond raw accuracy, the attention mechanism offers a practical advantage the comparator approaches in Table XII do not: because α_i (Section III-C-3) is computed per prediction, it can, in principle, be surfaced to a teacher alongside an individual engagement flag to indicate whether that flag was driven primarily by emotional, behavioral, or assessment signals. Whether this per-prediction weight is intuitive or useful to an actual teacher was not tested in this

study and would need direct evaluation before being built into a classroom-facing interface.

Several limitations qualify these findings. First, the pooled confusion matrix reported in Section IV-C is heavily imbalanced (Table VII); the true negative count (1) is too small to support a reliable estimate of specificity, and per-class precision/recall for the minority class should be reported and interpreted with caution in any follow-on work. Second, the primary training subset (759 samples) and the two supplementary datasets (503 and 150 samples respectively) are modest in size relative to the full DAiSEE corpus (9,068 clips); the generalization results in Section IV-B should be read as indicative rather than definitive. Third, hyperparameter tuning used grid search over a deliberately bounded space (Section III-D); as discussed there, this was a reproducibility-motivated choice rather than a claim that grid search is optimal, and a wider search using heuristic or Bayesian methods might yield further gains. Fourth, the temporal robustness analysis (Section IV-F) covers only six months and does not test the model against a genuinely novel cohort of learners, so longer-horizon and cross-cohort drift remain open questions. Finally, as with any AI system deployed in an educational setting, fairness and bias across demographic subgroups were not evaluated here and should be a precondition for any operational deployment, consistent with the responsible-AI framing set out in Section I.

VI. CONCLUSION AND FUTURE WORK

This study designed, implemented, and empirically evaluated a hybrid CNN-LSTM architecture with modality-level attention fusion for classifying learner engagement within an Intelligent Tutoring System, motivated by the observation that most deployed ITS remain affect-blind and that existing affect-aware alternatives are often too computationally demanding for resource-constrained educational settings. Evaluated on a structured DAiSEE subset and two supplementary datasets, the model achieved 0.91 accuracy, 0.92 precision, 0.94 recall, 0.92 F1 score, and 0.70 AUC-ROC on a binary engagement decision, exceeding all predefined performance targets, with emotional features consistently identified as the strongest predictor of engagement across both modality ablation and attention weight analysis. Temporal validation over six months showed only marginal degradation, suggesting the model does not require frequent retraining over that horizon.

Future work should prioritize four directions raised by the limitations discussed in Section V: (1) evaluating the framework on the native, four-level DAiSEE engagement scale to enable direct comparison with prior benchmark studies, rather than the binary decision task reported here; (2) validating performance on larger and more demographically diverse samples, including a formal fairness and bias audit across learner subgroups, before any operational deployment; (3) comparing the grid-search hyperparameter configuration used here against heuristic or Bayesian optimization over a wider search space; and (4) conducting a direct evaluation, with practicing teachers, of whether the per-prediction attention weights described in Section V are interpretable and useful as a classroom-facing explanation of engagement flags. Pursuing these directions would strengthen the case for embedding lightweight, attention-based affective computing

modules within ITS architectures deployed in low-resource educational settings.

DATA AVAILABILITY STATEMENT

The primary dataset supporting this study, DAiSEE, is a publicly available benchmark introduced by Gupta et al. [40] and is accessible from the dataset's official repository. The two supplementary datasets used for cross-dataset generalization testing (Section III-B) are available from the corresponding author upon reasonable request.

REFERENCES

- [1] K. VanLehn, "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems," *Educational Psychologist*, vol. 46, no. 4, pp. 197–221, 2011.
- [2] B. P. Woolf, *Building Intelligent Interactive Tutors: Student-Centered Strategies for Revolutionizing e-Learning*. Burlington, MA: Morgan Kaufmann, 2010.
- [3] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affective Comput.*, vol. 1, no. 1, pp. 18–37, 2010.
- [4] S. K. D'Mello and A. Graesser, "AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers," in *APA Educational Psychology Handbook*, vol. 3. Washington, DC: American Psychological Association, 2012, pp. 451–473.
- [5] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, "School engagement: Potential of the concept, state of the evidence," *Rev. Educ. Res.*, vol. 74, no. 1, pp. 59–109, 2004.
- [6] R. Pekrun, "Emotion and learning," *Educ. Psychol.*, vol. 52, no. 4, pp. 207–225, 2017.
- [7] R. F. Kizilcec, M. Pérez-Sanagustín, and E. Schneider, "Failure to sign up for online courses: Predicting MOOC dropout," *Comput. Educ.*, vol. 103, pp. 1–15, 2017.
- [8] A. Kumar, K. Reddy, and C. Rao, "Deep learning for student performance modeling," *IEEE Trans. Learn. Technol.*, vol. 8, no. 4, pp. 345–357, 2015.
- [9] R. Klein and T. Celik, "The Wits Intelligent Teaching System: Detecting student engagement during lectures using convolutional neural networks," in *Proc. 2017 IEEE Int. Conf. Image Process. (ICIP)*, 2017, pp. 2856–2860.
- [10] Y. Hu, Z. Jiang, and K. Zhu, "An optimized CNN model for engagement recognition in an e-learning environment," *Appl. Sci.*, vol. 12, no. 16, p. 8007, 2022.
- [11] T. Selim, I. Elkabani, and M. A. Abdou, "Students engagement level detection in online e-learning using hybrid EfficientNetB7 together with TCN, LSTM, and Bi-LSTM," *IEEE Access*, vol. 10, pp. 99573–99583, 2022.
- [12] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Trans. Affective Comput.*, vol. 4, no. 1, pp. 15–33, 2013.
- [13] F. J. Agbo, S. S. Oyelere, J. Suhonen, and P. Ihanola, "Impacts of artificial intelligence on student learning in sub-Saharan Africa: A systematic review," *Educ. Inf. Technol.*, vol. 28, no. 1, pp. 1–27, 2023.
- [14] H. Khosravi, K. Kitto, and D. Gasevic, "Hybrid learning models: Combining rule-based and machine learning approaches in intelligent tutoring systems," *IEEE Trans. Learn. Technol.*, vol. 13, no. 4, pp. 788–804, 2020.
- [15] K. Karpouzis, "Explainable AI for intelligent tutoring systems," in *Frontiers of Artificial Intelligence, Ethics, and Multidisciplinary Applications (FAIEMA 2023)*, M. Farmanbar, M. Tzamtzi, A. K. Verma, and A. Chakravorty, Eds. Cham: Springer, 2024.
- [16] A. Mitrovic, M. Mayo, P. Suraweera, and B. Martin, "Constraint-based tutors: A success story," in *Studies in Computational Intelligence*, vol. 62. Berlin: Springer, 2007, pp. 271–296.
- [17] K. R. Koedinger and A. T. Corbett, "Cognitive tutors: Technology bringing learning sciences to the classroom," in *The Cambridge Handbook of the Learning Sciences*, R. K. Sawyer, Ed. Cambridge: Cambridge Univ. Press, 2006, pp. 61–78.

- [18] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.
- [19] G. Siemens, "Learning analytics: The emergence of a discipline," *Amer. Behav. Sci.*, vol. 57, no. 10, pp. 1380–1400, 2013.
- [20] W. Holmes, M. Bialik, and C. Fadel, *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. Boston, MA: Center for Curriculum Redesign, 2019.
- [21] R. S. Baker and P. Inventado, "Educational data mining and learning analytics," in *Learning Analytics*, J. A. Larusson and B. White, Eds. New York: Springer, 2014, pp. 61–75.
- [22] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. Guibas, and J. Sohl-Dickstein, "Deep knowledge tracing," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2015.
- [23] Y. Wang and N. Heffernan, "Improving knowledge tracing using deep learning and richer features," in *Proc. Int. Educ. Data Mining Conf.*, 2016.
- [24] L. Zhang, X. Xiong, S. Zhao, A. Botelho, and N. T. Heffernan, "Incorporating graph neural networks into knowledge tracing," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2020.
- [25] D. Rai and J. Chun, "Reinforcement learning for adaptive tutoring systems," *IEEE Intell. Syst.*, vol. 33, no. 10, pp. 35–45, 2018.
- [26] W. Hu, M. Li, and X. Ren, "Learning optimal pedagogical strategies with reinforcement learning in e-learning," in *Proc. Int. Conf. Artif. Intell. Educ. (AIED)*, 2019.
- [27] I. Arroyo, B. P. Woolf, W. Burelson, K. Muldner, D. Cooper, and H. Hastie, "A multimedia environment for learning mathematics: Evaluating affective feedback mechanisms," *Int. J. Artif. Intell. Educ.*, vol. 24, no. 1, pp. 3–26, 2014.
- [28] A. C. Graesser, H. Li, and C. Forsyth, "AutoTutor: A virtual tutor with dialogue," *Behav. Res. Methods*, vol. 45, no. 2, pp. 402–420, 2013.
- [29] Z. Chen and L. Ibrahim, "Large language model-based tutoring systems: Opportunities and risks," *AI Educ. Rev.*, vol. 2, no. 1, pp. 55–78, 2024.
- [30] J. Liu, J. Park, and X. Chen, "LLM-powered conversational tutors for open-ended learning," *Comput. Educ. Artif. Intell.*, vol. 6, p. 100215, 2024.
- [31] S. N. Karimah, T. Unoki, and S. Hasegawa, "Implementation of Long Short-Term Memory (LSTM) models for engagement estimation in online learning," in *Proc. 2021 IEEE Int. Conf. Eng., Technol. Educ. (TALE)*, 2021, pp. 283–289.
- [32] A. Abedi and S. S. Khan, "Improving state of the art in detecting student engagement with ResNet and TCN hybrid network," in *Proc. 2021 18th Conf. Robots Vis. (CRV)*, 2021, pp. 151–157.
- [33] N. K. Mehta, S. S. Prasad, S. Saurav, R. Saini, and S. Singh, "Three-dimensional DenseNet self-attention neural network for automatic detection of student's engagement," *Appl. Intell.*, vol. 52, pp. 13803–13823, 2022.
- [34] X. Ai, V. S. Sheng, and C. Li, "Class attention video transformer for engagement intensity prediction," *arXiv:2208.07216*, 2022.
- [35] W. L. Zhang, R. S. Jia, H. Wang, C. Y. Che, and H. M. Sun, "A self-supervised learning network for student engagement recognition from facial expressions," *IEEE Trans. Circuits Syst. Video Technol.*, 2024.
- [36] M. Singh, X. Hoque, D. Zeng, Y. Wang, K. Ikeda, and A. Dhall, "Do I have your attention: A large-scale engagement prediction dataset and baselines," in *Proc. 25th Int. Conf. Multimodal Interact.*, 2023, pp. 174–182.
- [37] Ö. Sümer, P. Goldberg, S. D'Mello, P. Gerjets, U. Trautwein, and E. Kasneci, "Multimodal engagement analysis from facial videos in the classroom," *IEEE Trans. Affective Comput.*, 2021.
- [38] B. Bühler, E. Bozkir, P. Goldberg, Ö. Sümer, S. D'Mello, P. Gerjets, U. Trautwein, and E. Kasneci, "From the lab to the wild: Examining generalizability of video-based mind wandering detection," *Int. J. Artif. Intell. Educ.*, vol. 35, pp. 823–857, 2025.
- [39] T. Nazaretsky, M. Cukurova, and G. Alexandron, "An instrument for measuring teachers' trust in AI-based educational technology," in *Proc. LAK22: 12th Int. Learn. Anal. Knowl. Conf.*, 2022.
- [40] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "DAiSEE: Towards user engagement recognition in the wild," *arXiv:1609.01885*, 2016.
- [41] H. Hua, Y. Li, T. Wang, N. Dong, W. Li, and J. Cao, "Edge computing with artificial intelligence: A machine learning perspective," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–35, 2023.