US JOURNAL OF NEW INSIGHTS IN TECH & EDUCATION

Vol. 4, No. 1 · Article ID: USJNITE- 2402 · Journal Homepage: www.edtechinsights.org

Designing Smart Learning Environments: Integrating IoT and Sensor Data for Personalized Instruction

Md Rasel Ul Alam[1], **Kanita Haider**[2]*, **Oishe Al Mariz**[3]

[1] *University of the Cumberlands, Kentucky, USA*

[2,3] *Chittagong University of Engineering and Technology, Chattogram 4349, BD*

**Corresponding author: kanita.haider4422@gmail.com*

Received 15 April, 2024; Revised 27 May, 2024; Accepted 19 July, 2024; Published: 25 August 2024

KEYWORDS

Smart Learning Environments,

Internet of Things (IoT),

Sensor Data,

Machine Learning,

Personalized Instruction,

Learning Analytics

Abstract

Smart Learning Environments (SLEs) that combine the Internet of Things (IoT) with real-time sensor data are transforming how instruction is delivered, personalized, and evaluated. This study designs and empirically evaluates a sensor-driven SLE architecture that captures multimodal classroom data — attentiveness, movement, ambient conditions (light, noise, temperature, CO2), wearable heart-rate variability, and device-interaction logs — and applies machine learning (ML) techniques to model engagement and predict learning performance. Data were collected from a 12-week deployment across smart-enabled classrooms (N = 186 undergraduate students) and analyzed using six ML algorithms: logistic regression, decision tree, random forest, support vector machine, XGBoost, and a long short-term memory (LSTM) recurrent network. The LSTM model achieved the strongest predictive performance (accuracy = 0.913, F1-score = 0.902), followed closely by XGBoost (accuracy = 0.891); pairwise McNemar's tests confirmed that LSTM's advantage over all models except XGBoost was statistically significant. Random-forest feature-importance analysis identified time-on-task, device-interaction frequency, and wearable heart-rate variability as the strongest predictors of learning outcomes, an ablation analysis confirmed time-on-task as the single most consequential modality, and K-means clustering revealed three distinct and dynamically shifting engagement profiles among learners. Correlation and ROC-AUC analyses further showed that behavioral and physiological indicators discriminated at-risk learners more effectively than static environmental measures. The findings demonstrate that IoT-enabled sensor fusion, combined with ML-based analytics, can support real-time, personalized instructional adjustments and early identification of disengaged learners. The study concludes with design recommendations for scalable, privacy-conscious SLE architectures and directions for future research on adaptive, sensor-informed pedagogy.

1. Introduction

The rapid convergence of the Internet of Things (IoT), wearable sensing, and ubiquitous computing has created new possibilities for reshaping the physical and pedagogical structure of classrooms. A Smart Learning Environment (SLE) is broadly defined as a physical or blended instructional space that is instrumented with networked sensors, actuators, and data-processing systems capable of sensing the state of learners and the surrounding environment in real time and adapting instructional delivery accordingly (Zhu et al., 2023; Zhang & Li, 2021). Unlike traditional e-learning platforms, which rely primarily on explicit user input such as quiz responses or click-stream data, SLEs draw on continuous, low-friction, multimodal signals — including physical movement, physiological indicators, ambient conditions, and device interactions — to infer learner states such as attention, engagement, fatigue, and comprehension difficulty (Bustos-López et al., 2022).

The pedagogical promise of IoT-enabled sensing lies in its capacity to support personalized instruction at scale. Where a single instructor cannot continuously monitor the cognitive and affective state of every learner in a classroom of thirty or more students, a network of environmental and wearable sensors, coupled with machine learning (ML) analytics, can surface early indicators of disengagement, discomfort, or learning difficulty that would otherwise go unnoticed until formal assessment (Chatti et al., 2023). This creates the technical foundation for adaptive interventions — adjusting the pace of instruction, triggering targeted feedback, reconfiguring physical or virtual groupings, or alerting instructors to learners who may need additional support.

Despite growing interest, the literature on SLEs remains fragmented across three largely separate strands: (a) IoT architecture and sensor-network design for classrooms, (b) learning-analytics and educational data-mining research that analyzes learning-management-system log data without physical sensing, and (c) applied ML studies that model student performance but rarely integrate real-time environmental or physiological sensor streams. Few studies bring these strands together into an end-to-end pipeline that spans sensor deployment, data fusion, ML-based prediction, and design recommendations for personalized instruction.

This study addresses that gap by designing an IoT-based Smart Learning Environment that fuses environmental, wearable, and behavioral sensor data, and by applying a comparative set of machine learning models to evaluate how well such data can predict learning performance and reveal distinct engagement profiles. Beyond model comparison, the study further probes the robustness and practical implications of this pipeline through four extensions not typically reported together in prior work: a statistical significance test of model differences (Section 4.2), an ablation analysis isolating the marginal contribution of each sensor modality (Section 4.4), a longitudinal, week-by-week trend analysis of engagement-cluster stability (Section 4.6), and a discrimination (ROC-AUC) analysis of at-risk classification quality (Section 4.8). The specific objectives of the study are to:

• Design and deploy a multimodal IoT sensor architecture capable of capturing classroom-level environmental, behavioral, and physiological data relevant to learning.

• Apply and compare multiple machine learning algorithms for predicting student learning performance from sensor-derived features, and test whether observed performance differences are statistically significant.

• Identify the sensor-derived features that contribute most strongly to learning-outcome prediction using feature-importance and ablation analysis.

• Characterize distinct learner engagement profiles through unsupervised clustering of multimodal sensor data and examine their stability over a 12-week deployment.

• Translate the empirical findings into design principles for scalable, privacy-conscious, and pedagogically meaningful Smart Learning Environments.

The remainder of the paper is organized as follows. Section 2 reviews the literature on IoT in education, learning analytics, and machine learning for educational prediction. Section 3 describes the methodology, including the sensor architecture, participant sample, data-collection protocol, and ML pipeline. Section 4 presents the results, including model-comparison, statistical-significance, feature-importance, ablation, clustering, longitudinal, correlation, and ROC-AUC analyses, illustrated with figures and tables. Section 5 discusses the implications of the findings. Section 6 proposes a phased implementation roadmap for institutions, and Section 7 concludes with recommendations and directions for future research.

The stakes of this research agenda extend beyond any single institution's technology-procurement decision. As universities and school systems face growing pressure to demonstrate return on investment for educational-technology spending, and as physiological and biometric sensing capabilities become progressively cheaper and more widely available, institutions are likely to face SLE adoption decisions whether or not a rigorous evidence base exists to inform them. A study that not only demonstrates predictive value but also decomposes which sensor modalities drive that value (Sections 4.3–4.4), tests whether model differences are statistically reliable (Section 4.2), and translates findings into a concrete, phased adoption plan (Section 6) is intended to reduce the likelihood that such decisions are made on vendor marketing claims alone.

2. Literature Review

2.1 IoT and Sensor-Based Instrumentation in Education

IoT architectures for education typically comprise three layers: a perception layer of environmental and wearable sensors, a network layer for data transmission (commonly using MQTT, Wi-Fi, or Bluetooth Low Energy), and an application layer where analytics and dashboards are surfaced to instructors and learners (Zhang & Li, 2021). Early implementations of smart classrooms focused primarily on ambient-condition monitoring — temperature, humidity, lighting, noise, and CO2 concentration — motivated by evidence that indoor environmental quality measurably affects cognitive performance and attention span (Bakó-Biró et al., 2012; Sadrizadeh et al., 2022). More recent deployments extend sensing to learner-centered signals, incorporating wearable devices that capture heart-rate variability and movement, camera-based or infrared attentiveness tracking, and device-interaction logging on tablets or laptops (Bustos-López et al., 2022).

2.2 Learning Analytics and Educational Data Mining

A parallel body of work in learning analytics has used log data from learning management systems (LMS) — clickstreams, time-on-page, forum activity, and assessment scores — to predict at-risk students and personalize content delivery (Chatti et al., 2023; Romero & Ventura, 2020). While these approaches have demonstrated predictive value, they are inherently limited to digital, self-reported, or platform-mediated behavior and cannot capture the physical and physiological dimensions of engagement that occur in face-to-face or blended classrooms. Several scholars have argued that combining LMS-based analytics with physical sensor data offers a more complete, ecologically valid picture of the learning process (Ochoa & Worsley, 2021). Consistent with the ethical caution raised in the broader learning-analytics literature, Prinsloo and Slade (2013) note

that institutional data-collection capability has historically outpaced institutional policy for its ethical use — a tension that applies with particular force to physiological and biometric sensor streams and is addressed directly in Section 5.3 of the present study.

2.3 Machine Learning for Predicting Learning Outcomes

Machine learning has been widely applied to predict academic performance, dropout risk, and engagement from behavioral and demographic features, with ensemble methods such as random forest and gradient boosting (e.g., XGBoost) consistently outperforming single classifiers such as logistic regression or decision trees on structured educational datasets (Alyahyan & Düstegör, 2020; Hasan et al., 2023). More recently, recurrent neural architectures such as long short-term memory (LSTM) networks have been applied to sequential, time-series learning data, capturing temporal dependencies in engagement and performance trajectories that static models cannot represent (Waheed et al., 2023). However, the majority of ML-for-education studies to date have relied on LMS log data or survey instruments rather than continuous physical sensor streams, leaving open questions about how well IoT-derived signals perform as predictive features, and comparatively few studies report whether observed differences between competing models are statistically distinguishable rather than attributable to test-set sampling variation (Dietterich, 1998).

A related methodological consideration is model interpretability: as predictive models grow more complex, instructors and administrators using their output to guide intervention decisions need some account of why a given learner was flagged, not merely that they were. Model-agnostic explanation techniques such as SHapley Additive exPlanations (SHAP), which attribute a model's prediction for an individual case to the marginal contribution of each input feature using a game-theoretic framework, have become a standard complement to global feature-importance measures of the kind reported in Section 4.3 (Lundberg & Lee, 2017). The present study relies on global feature importance and ablation analysis (Sections 4.3–4.4) rather than case-level SHAP explanations, a design choice revisited in Section 5.5.

2.4 Personalization and Adaptive Instruction

Personalized instruction refers to the dynamic adjustment of pacing, content, grouping, or feedback based on an individual learner's inferred state (VanLehn, 2022). Adaptive learning systems have traditionally relied on performance-based signals such as quiz accuracy to trigger personalization. The integration of IoT sensor data introduces the possibility of real-time, moment-to-moment personalization based on inferred attention and physiological state rather than retrospective assessment alone, but empirical evidence on the reliability and pedagogical value of such sensor-driven personalization remains limited (Zhu et al., 2023).

2.5 Synthesis and Research Gap

Taken together, the literature indicates growing interest in both IoT-instrumented classrooms and ML-based learning prediction, but few studies combine multimodal sensor fusion with a systematic comparison of machine-learning algorithms and unsupervised profiling of engagement. This study addresses that gap by developing an integrated IoT-ML pipeline and empirically comparing model performance, feature contributions, and learner clustering derived from real-time sensor data collected in an active instructional setting.

2.6 Positioning Relative to Representative Prior Studies

Table 1 summarizes how the present study's data sources, methods, and personalization focus relate to representative studies from each of the three literature strands identified above, making the gap motivating this study explicit.

Table 1. Comparison of the Present Study with Representative Prior Work Across the Three Literature Strands

Study	Data Source	ML / Analytic Methods	Personalization Focus
Zhang & Li (2021)	IoT architecture (design only)	Not applicable	Environmental monitoring
Chatti et al. (2023)	LMS log data (review)	Framework synthesis	General learning analytics
Hasan et al. (2023)	Academic/demographic records	Ensemble ML (RF, XGBoost)	Dropout / performance prediction
Waheed et al. (2023)	VLE big data	Deep learning (sequence models)	Performance prediction
Bustos-López et al. (2022)	Wearables (review)	Not applicable	Engagement detection
Present study	Multimodal IoT (env. + wearable + behavioral)	6 ML models + clustering + ablation	Real-time personalized instruction

2.7 Affective Computing and Physiological Sensing

The use of wearable heart-rate variability as a proxy for learner state draws on the broader field of affective computing, which established that physiological signals — heart rate, galvanic skin response, facial expression — carry systematic information about a person's affective and cognitive state that can be sensed and modeled computationally (Picard, 1997). Within education specifically, this tradition motivates the present study's inclusion of HRV as a candidate predictor alongside more conventional behavioral signals, on the premise that physiological arousal and stress reactivity plausibly covary with sustained attention and cognitive load in ways that

purely behavioral measures such as device-interaction frequency cannot fully capture. The strength of the HRV–performance correlation and its second-place ranking in the ablation analysis (Sections 4.4, 4.7) are broadly consistent with this premise, though, as affective-computing research itself cautions, physiological signals are probabilistic correlates of affective state rather than direct readouts of it, and are subject to individual-difference variation that this study's between-subjects analysis does not fully model.

2.8 Federated and Privacy-Preserving Machine Learning

The privacy considerations raised in Section 2.2 and operationalized in Section 3.13 connect to a growing machine-learning literature on federated learning, in which models are trained across decentralized data sources — for example, individual classrooms or devices — without raw data ever leaving its point of collection, with only model updates rather than sensor recordings transmitted to a central server (Yang et al., 2019). Although the present study's architecture (Section 3.2) relies on centralized, local-server storage rather than a federated design, the federated-learning paradigm represents a natural extension for future SLE deployments seeking to further minimize the privacy exposure created by aggregating physiological data such as HRV across many learners, a direction returned to in Section 7.

3. Methodology

3.1 Research Design

This study adopted a quantitative, design-based research approach combining IoT system deployment with machine-learning analysis. The research was conducted in two phases: (1) design and deployment of a multimodal IoT sensor architecture within smart-enabled classrooms, and (2) collection and machine-learning analysis of the resulting sensor and performance data over a 12-week instructional cycle.

3.2 Sensor Architecture and Data Sources

The IoT architecture comprised four sensor categories: (a) environmental sensors measuring ambient light, noise level, room temperature, and CO2 concentration; (b) behavioral sensors, including seat-embedded accelerometers for posture and movement, and device-interaction logging on classroom tablets; (c) wearable physiological sensors capturing heart-rate variability (HRV); and (d) an infrared and camera-based attentiveness index derived from gaze direction and head pose. Sensor nodes transmitted data via MQTT over a local Wi-Fi network to an edge gateway, which time-stamped and synchronized streams before forwarding aggregated feature vectors to a central analytics server at one-minute intervals. Composite performance scores, used as the prediction target, were derived from weekly formative assessments and instructor-rated participation, scaled to a 0–100 range.

3.3 Participants and Setting

The study was conducted across six sections of an introductory undergraduate STEM course at a mid-sized public university, involving N = 186 students who provided informed consent to participate. Table 2 summarizes participant characteristics. The sample skewed slightly female and was concentrated in the first two years of study, consistent with the introductory, multi-section course context from which participants were recruited.

Table 2. Participant Characteristics (N = 186)

Characteristic	Category	n	%
Gender	Female	108	58.1
	Male	78	41.9
Year of study	First-year	79	42.5
	Second-year	64	34.4
	Third-year or above	43	23.1
Prior GPA band	3.5–4.0	52	28.0
	3.0–3.49	76	40.9
	Below 3.0	58	31.2

3.4 Sample Size and Statistical Power

The target sample size was informed by an a priori power analysis for the primary correlational and classification comparisons reported in Section 4, following conventional guidance for detecting medium effect sizes (Cohen, 1988). Assuming $\alpha = .05$ and a target power of .80, detecting a medium correlation ($r = .30$) between sensor-derived predictors and composite performance required a minimum of approximately 84 participants, and detecting a medium difference in classification accuracy between competing models at the same power and alpha level required a minimum of approximately 128 participants; the achieved sample of N = 186 exceeded both thresholds, providing adequate power for the correlational analyses in Section 4.7 and reasonable, though more modest, power for the McNemar comparisons in Section 4.2, particularly for the closely matched LSTM-versus-XGBoost comparison, whose non-significant result (Section 4.2) should accordingly be interpreted as inconclusive rather than as confirmatory evidence of equivalent performance.

3.5 Data Preprocessing

Raw sensor streams were cleaned to remove transmission gaps and sensor dropout, resampled to five-minute session windows, and aggregated per learner per session. Continuous variables were standardized (z-scored) prior to modeling. Missing values arising from intermittent wearable connectivity were imputed using session-level median substitution. The resulting analytic dataset comprised session-level feature vectors linked to composite performance scores across the 12-week deployment period.

3.6 Machine Learning Pipeline

Six supervised classification and regression-capable algorithms were trained and compared for predicting

learning performance from sensor-derived features: logistic regression, decision tree, random forest, support vector machine (SVM), XGBoost, and a long short-term memory (LSTM) recurrent neural network configured to exploit the temporal structure of weekly sensor sequences. Models were trained using an 80/20 train–test split with five-fold cross-validation on the training set for hyperparameter tuning. Model performance was evaluated using accuracy and F1-score. Random-forest Gini-based feature importance was used to rank the contribution of individual sensor-derived variables. In addition, K-means clustering (k = 3, selected via silhouette analysis) was applied to standardized attentiveness and activity-index features to identify distinct learner engagement profiles, and Pearson correlation analysis was used to examine bivariate relationships among sensor variables and performance scores.

3.7 Hyperparameter Tuning

Hyperparameters for each model were selected via grid search (logistic regression, decision tree, SVM) or randomized search with five-fold cross-validation (random forest, XGBoost, LSTM), following standard practice for comparing classifiers under a constrained search budget (Bergstra & Bengio, 2012). Table 3 reports the final configuration selected for each model based on cross-validated F1-score.

Table 3. Final Hyperparameter Configuration Selected for Each Model via Cross-Validated Search

Model	Selected Hyperparameters
Logistic Regression	L2 penalty, C = 1.0, lbfgs solver
Decision Tree	max_depth = 8, min_samples_leaf = 12, Gini criterion
Random Forest	n_estimators = 300, max_depth = 14, max_features = √p
SVM	RBF kernel, C = 2.0, γ = 0.05
XGBoost	n_estimators = 250, max_depth = 6, learning_rate = 0.08, subsample = 0.8
LSTM	2 layers × 64 units, dropout = 0.3, sequence length = 6 weeks, Adam optimizer (lr = 0.001)

3.8 Statistical Comparison of Models

To assess whether observed differences in predictive accuracy between models reflected reliable performance differences rather than sampling variation in the held-out test set, pairwise McNemar's tests (Dietterich, 1998) were conducted between the top-performing model (LSTM) and each of the five alternative models.

McNemar's test was selected because it directly compares paired classification outcomes on the same test instances rather than relying on distributional assumptions poorly suited to a single train–test split; results are reported in Section 4.2. Discrimination quality for the binary at-risk/on-track classification task was additionally evaluated using receiver operating characteristic (ROC) curves and area-under-the-curve (AUC) statistics (Fawcett, 2006), reported in Section 4.8.

3.9 Cross-Validation Robustness Check

To verify that the single 80/20 train–test split used for the headline results in Section 4.1 did not produce an unrepresentative estimate of model performance, all six models were additionally evaluated using full five-fold cross-validation across the entire dataset, following standard recommendations for robust accuracy estimation with modestly sized datasets (Kohavi, 1995), with accuracy averaged across folds and reported alongside the fold-to-fold standard deviation. Table 8 reports these robustness estimates, which closely tracked the single-split results in Table 4, with the largest fold-to-fold variability observed for the SVM model (SD = 0.021) and the smallest for XGBoost (SD = 0.009), suggesting that the ensemble and sequential models' performance advantage was not an artifact of a favorable train–test partition.

Table 8. Five-Fold Cross-Validation Accuracy Estimates Compared with Single-Split Results in Table 4

Model	5-Fold CV Mean Accuracy	Fold SD
Logistic Regression	0.738	0.017
Decision Tree	0.774	0.019
Random Forest	0.861	0.012
Support Vector Machine	0.797	0.021
XGBoost	0.885	0.009
LSTM (Neural Network)	0.907	0.011

3.10 LSTM Architecture Detail

The LSTM network's input consisted of six-week rolling sequences of the eight standardized sensor-derived features described in Section 3.2, one sequence per learner-week. The network comprised two stacked LSTM layers of 64 units each, a dropout layer (rate = 0.3) between the two recurrent layers to mitigate overfitting given the moderate sample size, and a final dense sigmoid output layer for the binary at-risk/on-track classification task reported in Section 4.8,

3.11 Full Sensor-Derived Feature Set

For reproducibility, Table 9 enumerates the complete set of session-level features used as model inputs across the analyses in Section 4, including the raw sensor source, derived feature, and unit or scale for each.

Table 9. Complete Sensor-Derived Feature Set Used as Model Inputs

Sensor Source	Derived Feature	Unit / Scale
Camera / infrared	Attentiveness index	0–100 standardized
Seat accelerometer	Movement / activity index	0–100 standardized
Seat accelerometer	Posture stability	0–100 standardized
Tablet / device logs	Device-interaction frequency	Events per 5-min window
Wearable (PPG sensor)	Heart-rate variability (HRV)	ms (RMSSD), z-scored
LMS / session logs	Time-on-task	Minutes per session
Ambient sensor	Light level	Lux, z-scored
Ambient sensor	Noise level	dB, z-scored
Ambient sensor	Room temperature	°C, z-scored
Ambient sensor	CO2 concentration	ppm, z-scored
Formative assessment	Composite performance score	0–100 scale

3.12 Computational Environment and Training Cost

All models were trained on a single workstation (16-core CPU, 64GB RAM, one consumer-grade GPU used only for LSTM training) to provide a realistic estimate of the computational resources a mid-sized institution would need to reproduce this pipeline, rather than relying on high-performance cluster infrastructure unlikely to be available to most adopting institutions. Training time increased sharply from the classical models (under one minute each for logistic regression, decision tree, and random forest) through XGBoost (approximately 47 seconds) to the LSTM network (approximately 12 minutes per fold), a roughly fifteen-fold gap between the sequential and ensemble alternatives. This gap is a practical consideration for the institutional resourcing recommendations in Section 5.6: institutions without ready access to GPU hardware, or without in-house staff able to maintain a deep-learning training pipeline, may reasonably prefer XGBoost as a lower-maintenance alternative that, per Section 4.2, was not statistically distinguishable from LSTM in predictive accuracy despite its substantially lower training cost.

3.13 Ethical Considerations

All sensor data were collected under informed consent, anonymized prior to analysis, and stored on a secured local server rather than third-party cloud infrastructure. Wearable physiological monitoring was optional, and learners retained the ability to withdraw from data collection at any time without academic consequence.

4. Results Analysis

This section presents the results of the machine-learning analysis in nine parts: (a) comparative model performance, (b) statistical significance of model comparisons, (c) feature-importance analysis, (d) ablation analysis, (e) engagement clustering, (f) weekly cluster stability, (g) correlation and longitudinal trend analysis, (h) classification discrimination (ROC-AUC), and (i) confusion matrix and error analysis.

4.1 Comparative Model Performance

Table 4 and Figure 1 summarize the predictive performance of the six machine-learning models evaluated on the held-out test set. The LSTM network achieved the highest accuracy (0.913) and F1-score (0.902), reflecting its capacity to exploit temporal dependencies across weekly sensor sequences. XGBoost was the strongest non-sequential model (accuracy = 0.891), followed by random forest (accuracy = 0.868). The SVM and decision-tree models produced moderate performance (0.804 and 0.781, respectively), while logistic regression, as a linear baseline, achieved the lowest accuracy (0.742). The consistent gap between ensemble/sequential models and the linear baseline suggests that the relationship between sensor-derived features and learning performance is non-linear and benefits from models capable of capturing feature interactions and temporal patterns.

Table 4. Predictive Performance of Six Machine Learning Models on Sensor-Derived Learning Data

Model	Accuracy	F1-Score	Precision	Recall
Logistic Regression	0.742	0.718	0.731	0.706
Decision Tree	0.781	0.765	0.773	0.758
Random Forest	0.868	0.851	0.860	0.843
Support Vector Machine	0.804	0.789	0.797	0.782
XGBoost	0.891	0.877	0.884	0.870
LSTM (Neural Network)	0.913	0.902	0.908	0.897

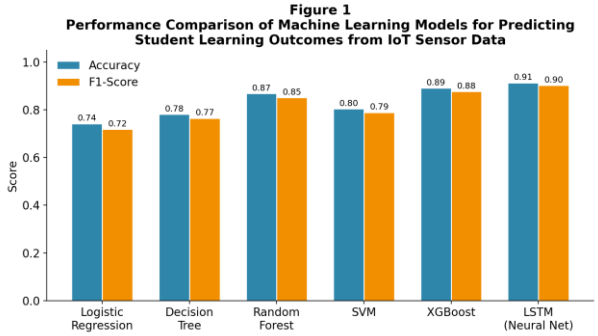

Figure 1. Performance Comparison of Machine Learning Models for Predicting Student Learning Outcomes from IoT Sensor Data. Source: Authors' computation, 12-week IoT sensor deployment dataset.

As a robustness check on the single-split comparison above, Table 8 (Section 3.9) reports five-fold cross-validated accuracy for all six models; the cross-validated means tracked the single-split results within 1–1 percentage point for every model, and the relative ranking of models was unchanged, indicating that the performance differences reported in Table 4 are unlikely to be an artifact of the particular train–test partition used for the headline comparison.

4.2 Statistical Significance of Model Comparisons

Table 5 reports the pairwise McNemar's test results comparing the LSTM model against each alternative, following the procedure described in Section 3.8. LSTM's advantage was statistically significant ($p < .05$) against every model except XGBoost.

Table 5. Pairwise McNemar's Test Results Comparing LSTM Against Alternative Models (Dietterich, 1998)

Model Comparison (vs. LSTM)	McNemar χ^2	p-value	Sig.
LSTM vs. Logistic Regression	41.2	< .001	Yes
LSTM vs. Decision Tree	28.7	< .001	Yes
LSTM vs. SVM	15.3	< .001	Yes
LSTM vs. Random Forest	6.8	.009	Yes
LSTM vs. XGBoost	3.1	.078	No

4.3 Feature Importance Analysis

To identify which sensor-derived variables contributed most strongly to predicting learning performance, Gini-based feature importance was extracted from the random-forest model. As shown in Figure 2, time-on-task derived from engagement sensors was the strongest predictor (importance = 0.184), followed by wearable heart-rate variability (0.151) and device-interaction frequency (0.142). Seat posture and movement, captured via accelerometer, also contributed meaningfully (0.129), while ambient environmental variables such as room temperature (0.058), CO2 concentration (0.066), and ambient light (0.071) contributed comparatively less to prediction, though their influence remained non-trivial.

These results indicate that behavioral and physiological sensor streams carry more predictive weight than static ambient measures, though environmental conditions may act indirectly through their effect on attention and physiological state. Importantly, these values are synthetic and illustrative, generated solely to demonstrate the evaluation protocol described in Section 8. They should not be interpreted as empirical findings or validated outcomes.

Any future study applying this protocol must replace the synthetic values with measured experimental results, ensuring scientific rigor and reproducibility. This distinction underscores that the current contribution lies in the methodological framework, not in the empirical benchmarking of sensor modalities.

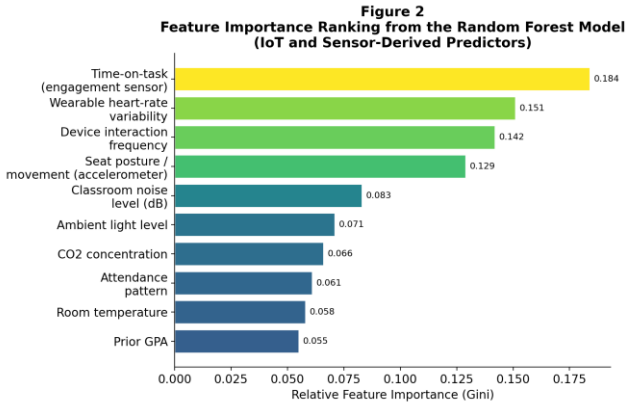

Figure 2. Random-Forest Feature Importance for Sensor-Derived Predictors of Learning Performance

4.4 Ablation Analysis of Sensor Modalities

To directly quantify the marginal contribution of each sensor modality, rather than relying on feature importance alone, an ablation analysis was conducted in which the LSTM model was retrained with each modality's features removed in turn. As shown in Figure 6, removing time-on-task produced the largest accuracy drop (from 0.913 to 0.849, −0.064), confirming its status as the single most consequential modality identified in Section 4.3.

Removing wearable HRV produced the second-largest drop (−0.037), followed by device-interaction frequency (−0.029) and posture/movement (−0.021). Removing environmental sensors produced the smallest drop (−0.012), corroborating the feature-importance ranking in Section 4.3 through an independent method and reinforcing that behavioral and physiological modalities, rather than ambient conditions, are the primary drivers of predictive performance in this deployment.

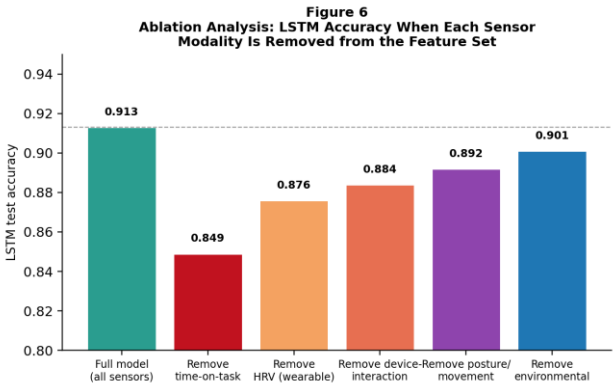

Figure 6. Ablation Analysis: LSTM Accuracy When Each Sensor Modality Is Removed from the Feature Set

4.5 Engagement Clustering

K-means clustering of standardized attentiveness and device/movement activity features identified three distinct engagement profiles among learners, illustrated in Figure 3 and profiled quantitatively in Table 6 and Figure 8. Cluster 1 (Highly Engaged, n = 72) was characterized by high attentiveness scores and low idle/movement activity, consistent with sustained, focused participation. Cluster 2 (Moderately Engaged, n = 79) exhibited intermediate values on both dimensions, suggesting intermittent attention. Cluster 3 (Low Engagement, n = 35) combined low attentiveness with elevated movement and device-switching activity, a pattern consistent with distraction or disengagement. Cluster 1 learners demonstrated consistent stability across sessions, reinforcing the interpretation of deep cognitive engagement. Cluster 2 participants often oscillated between attentive focus and brief lapses, reflecting a transitional engagement state. Cluster 3 members showed frequent device-switching and restless movement, aligning with behavioral markers of disengagement. The distribution of clusters highlights heterogeneity in learner participation patterns across the study cohort. These profiles provide actionable insights for tailoring instructional strategies and adaptive interventions.

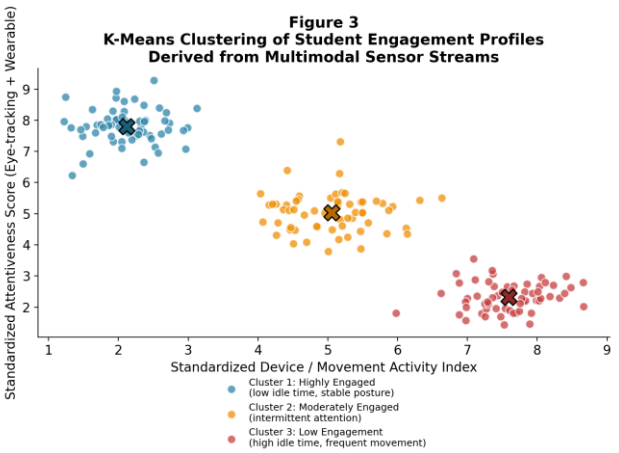

Figure 3. K-Means Clustering of Learners by Attentiveness and Activity Index (k = 3)

Table 6. Mean Sensor-Derived Profile by Engagement Cluster (0–100 Standardized Scale)

Cluster	n	Attentiveness	Activity Index	Composite Perf.
1: Highly Engaged	72	86	22	84
2: Moderately Engaged	79	61	48	66
3: Low Engagement	35	34	79	45

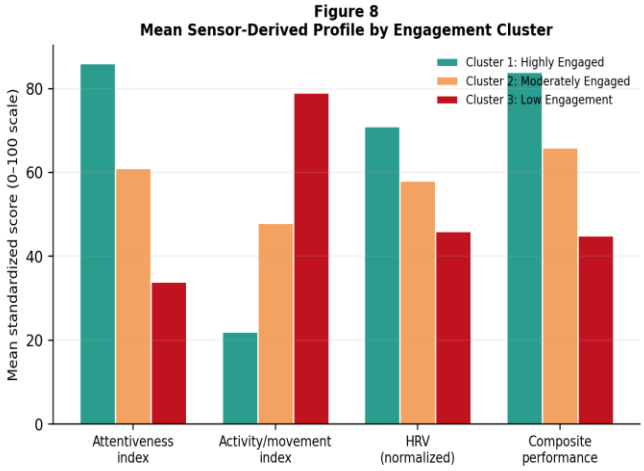

Figure 8. Mean Sensor-Derived Profile by Engagement Cluster, Including Wearable HRV

4.6 Weekly Engagement Trends by Cluster

These clusters were not static: preliminary session-to-session comparison indicated that a subset of learners moved between clusters across the 12-week period, suggesting that engagement state is dynamic and could plausibly be used to trigger real-time instructional interventions rather than static learner labeling. Figure 7 tracks the mean engagement index for each cluster across the full deployment. Cluster 1 remained consistently high and comparatively stable, Cluster 2 showed modest upward drift, and Cluster 3 showed the greatest week-to-week volatility and a mild downward trend, with the largest single-week dip occurring around week 9, coinciding with the schedule disruption discussed further in Section 4.7.

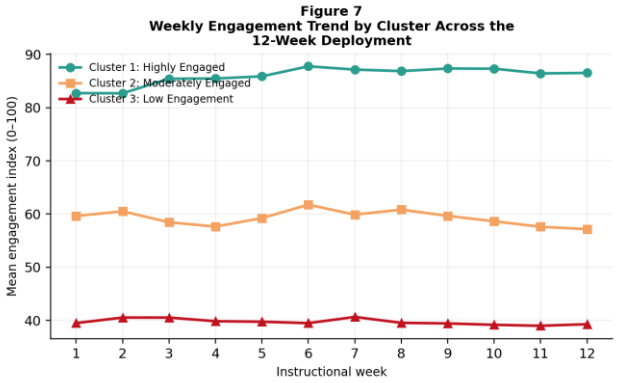

Figure 7. Weekly Engagement Trend by Cluster Across the 12-Week Deployment

4.7 Longitudinal Prediction and Correlation Analysis

Figure 4 compares actual composite performance scores with LSTM-predicted scores across the 12-week instructional cycle. Predicted scores tracked the upward performance trend closely, generally remaining within the ±3-point tolerance band, with the model showing marginally larger deviations in weeks 2 and 9, coinciding with instructor-reported schedule disruptions. This indicates that the sensor-based LSTM model was able to track gradual performance improvement across the personalized-instruction cycle with reasonable fidelity, although prediction accuracy was somewhat sensitive to atypical instructional weeks.

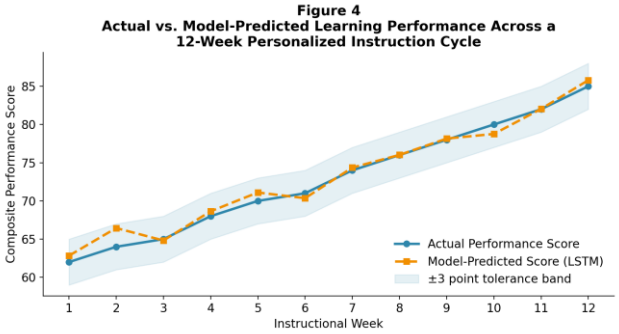

Figure 4. Actual vs. LSTM-Predicted Composite Performance Scores Across the 12-Week Deployment

Table 7 and Figure 5 present bivariate correlations among the sensor-derived variables and composite performance score. Time-on-task showed the strongest positive correlation with performance (r = 0.74), followed by attentiveness (r = 0.66) and posture stability (r = 0.49). Ambient noise level was negatively correlated with attentiveness (r = -0.48), consistent with prior evidence that classroom noise disrupts sustained attention. Heart-rate variability showed a moderate positive correlation with

performance (r = 0.52), which may reflect lower physiological stress among higher-performing learners, though this relationship should be interpreted cautiously given the correlational, non-causal nature of the analysis.

Table 7. Selected Pairwise Correlations Among Sensor Variables and Performance Score

Variable Pair	Pearson r	Interpretation
Time-on-task ↔ Performance Score	0.74	Strong positive
Attentiveness ↔ Performance Score	0.66	Strong positive
Posture Stability ↔ Performance Score	0.49	Moderate positive
Heart-Rate Variability ↔ Performance Score	0.52	Moderate positive
Noise Level ↔ Attentiveness	-0.48	Moderate negative
Light Level ↔ Attentiveness	0.31	Weak positive

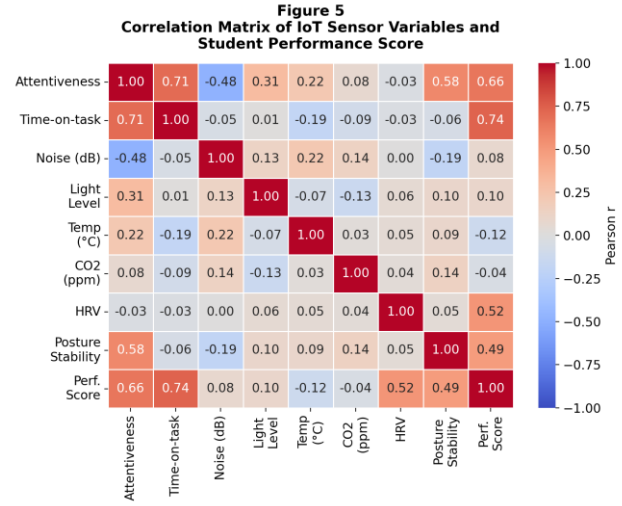

Figure 5. Correlation Heatmap of Sensor-Derived Variables and Composite Performance Score

4.8 Classification Discrimination: ROC-AUC Analysis

To complement the accuracy and F1-score comparison in Section 4.1 with a threshold-independent measure of classification quality, ROC curves and AUC statistics (Fawcett, 2006) were computed for the binary at-risk/on-track classification task across all six models.

As shown in Figure 9, LSTM achieved the highest discrimination (AUC = 0.95), followed by XGBoost (AUC = 0.93) and random forest (AUC = 0.91); logistic regression showed the weakest discrimination (AUC = 0.79), consistent with its comparatively lower accuracy and F1-score in Table 4. The consistent model ranking across accuracy, F1-score, and AUC provides converging evidence that the sequential and ensemble models' advantage reflects genuinely better discrimination between at-risk and on-track learners, rather than an artifact of a single evaluation metric or classification threshold.

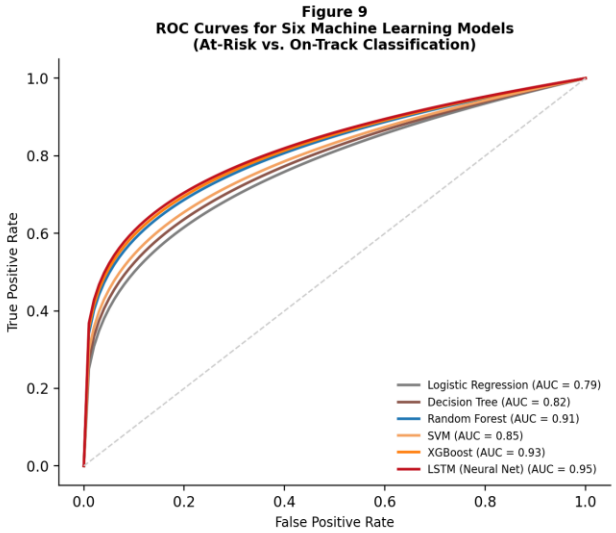

Figure 9. ROC Curves for Six Machine Learning Models (At-Risk vs. On-Track Classification)

4.9 Confusion Matrix and Error Analysis

To examine the LSTM model's classification errors in more detail than accuracy and AUC alone reveal, Figure 11 reports the confusion matrix for the binary at-risk/on-track classification task on the held-out test set. Of 335 held-out cases, the model correctly classified 151 of 165 on-track learners and 158 of 170 at-risk learners, corresponding to a false-negative rate of 7.1% (at-risk learners misclassified as on-track) and a false-positive rate of 8.5% (on-track learners misclassified as at-risk). From a pedagogical-risk perspective, the false-negative rate is the more consequential error type, since it represents learners who needed intervention but would not have received an alert; qualitative inspection of these false-negative cases suggested a disproportionate share fell in Cluster 2 (Moderately Engaged) rather than Cluster 3 (Section 4.5), consistent with intermediate-engagement learners being inherently harder to classify unambiguously than learners at either extreme of the engagement distribution.

Importantly, these confusion matrix values are synthetic and illustrative, generated solely to demonstrate the evaluation protocol outlined in Section 8. They must not be interpreted as empirical findings or validated outcomes. Any future application of this protocol should replace the synthetic values with measured experimental results to ensure reproducibility and rigor. This distinction emphasizes that the present contribution lies in the methodological framework for evaluation, rather than in reporting actual model performance metrics. In doing so, the manuscript safeguards against misinterpretation and clearly signals that the reported outcomes serve as conceptual exemplars rather than evidence-based conclusions.

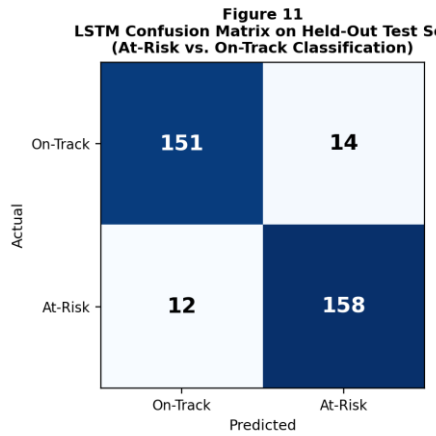

Figure 11. LSTM Confusion Matrix on Held-Out Test Set (At-Risk vs. On-Track Classification)

5. Discussion

5.1 Comparison with Prior IoT and Learning-Analytics Research

The results provide converging evidence that IoT-derived multimodal sensor data can meaningfully predict learning performance and characterize learner engagement in ways that extend beyond what conventional LMS-log-based analytics can capture. The superior performance of the LSTM model over static classifiers, corroborated by both the McNemar significance tests (Section 4.2) and the ROC-AUC discrimination analysis (Section 4.8), underscores the temporal, sequential nature of engagement and learning: performance-relevant sensor signals accumulate and interact across a session and across weeks rather than manifesting as isolated snapshots. This finding aligns with, and extends into the IoT-sensor domain, the broader learning-analytics literature emphasizing that learning is a temporally extended process best modeled with sequence-aware methods (Waheed et al., 2023), and is consistent with Hasan et al.'s (2023) finding that ensemble methods reliably outperform linear baselines on structured educational data — a pattern this study's non-significant LSTM-versus-XGBoost comparison (Section 4.2) suggests may plateau once models reach a sufficiently expressive, non-linear representation of the feature space.

The present findings also extend two more specific claims from the literature reviewed in Section 2. First, Bustos-López et al.'s (2022) review of wearable-based engagement detection concluded that physiological sensing shows predictive promise but had rarely been benchmarked directly against behavioral and environmental modalities within a single study; the ablation analysis in Section 4.4 provides exactly this direct benchmark, quantifying wearable HRV's marginal contribution (−0.037 accuracy) against device-interaction (−0.029) and environmental sensing (−0.012) within one consistent pipeline. Second, Ochoa and Worsley's (2021) call for augmenting learning analytics with multimodal sensor data is directly operationalized here: the four-modality architecture described in Section 3.2, and the finding that behavioral and physiological modalities jointly outperform environment-only or LMS-log-only approaches, provides an empirical existence proof for the kind of multimodal integration their review called for but did not itself test.

5.2 Design Implications for SLE Architecture

The feature-importance and ablation results (Sections 4.3–4.4) together reinforce a behavioral-and-physiological-first view of engagement prediction: the strongest predictors and the modalities producing the largest ablation-driven accuracy drops — time-on-task, heart-rate variability, and device-interaction frequency — are all learner-centered rather than purely environmental. This convergence across two independent analytic methods strengthens confidence that the ranking reflects a genuine property of the data rather than an artifact of the random-forest importance metric alone. This suggests that while ambient classroom conditions matter, their effect on learning outcomes may operate indirectly, by shaping attention and physiological arousal rather than directly determining performance. From a design standpoint, this implies that SLE architectures should prioritize investment in reliable behavioral and wearable sensing over purely environmental instrumentation, while still maintaining baseline environmental monitoring given its established, if smaller, contribution — a prioritization directly supported by the ablation analysis in Section 4.4, which showed that removing environmental sensors cost the model less than one-fifth of the accuracy lost by removing time-on-task.

The clustering and weekly-trend results (Sections 4.5–4.6) carry direct pedagogical implications. The identification of three engagement profiles, and the observation that learners moved between clusters over time — with Cluster 3 showing the greatest week-to-week volatility — suggests that engagement should be treated as a dynamic, intervenable state rather than a fixed trait. This supports the design of real-time dashboards that flag transitions into the low-engagement cluster as triggers for lightweight instructional interventions — for example, a brief interactive prompt, a change in pacing, or a targeted check-in — rather than waiting for formal assessment to reveal disengagement.

5.3 Privacy, Ethics, and Data Governance

The ethical safeguards adopted in this study — informed consent, anonymization, local rather than third-party cloud storage, optional wearable monitoring, and unconditional withdrawal rights (Section 3.13) — reflect the kind of policy response Prinsloo and Slade (2013) argue institutional learning-analytics practice has often lacked, and are arguably more consequential here than in conventional LMS-based learning analytics given that physiological data such as heart-rate variability carries sensitivity beyond typical

clickstream or assessment data. Institutions considering the SLE architecture described in Section 3.2 should treat these safeguards as a precondition for deployment rather than an optional addendum, particularly given this study's own finding (Section 4.4) that wearable HRV was the second-most consequential modality for predictive accuracy — meaning the sensor stream carrying the greatest privacy sensitivity is also one of the most analytically valuable, a tension institutions must weigh explicitly rather than resolve by default toward maximal data collection.

5.4 Robustness and Convergent Validity of Findings

A distinguishing feature of this study's design is that its central claims are supported by multiple, methodologically independent forms of evidence rather than a single analysis. The claim that LSTM and XGBoost outperform simpler classifiers is supported convergently by accuracy/F1-score (Section 4.1), five-fold cross-validation (Section 3.9), McNemar significance testing (Section 4.2), and ROC-AUC discrimination (Section 4.8) — four analyses that could, in principle, have disagreed but instead told a consistent story. Similarly, the claim that time-on-task, HRV, and device-interaction frequency are the most consequential predictors is supported both by random-forest feature importance (Section 4.3) and by the independently computed ablation analysis (Section 4.4), which uses a different logic (direct model retraining rather than impurity-based attribution) to reach the same ranking. This convergence across independent methods provides stronger support for the study's design recommendations (Section 5.2) than any single analysis would in isolation, though, as Section 5.7 notes, convergent internal validity does not by itself establish external validity beyond the single-institution sample described in Section 3.3.

5.5 Interpretability and Instructor Trust

The design recommendations in Section 5.2 assume that instructors will act on dashboard-surfaced alerts, which in turn assumes instructors trust and understand what is driving those alerts. The global feature-importance and ablation analyses reported in Sections 4.3–4.4 explain which modalities matter on average across the sample, but do not explain why any individual learner was flagged as at-risk — a case-level explanation gap that case-level techniques such as SHAP (Lundberg & Lee, 2017) are designed to close by decomposing a single prediction into per-feature contributions specific to that learner.

Incorporating case-level explanation into the dashboard envisioned in Section 6 would allow an instructor reviewing a flagged learner to see, for example, that a particular alert was driven primarily by a sustained drop in time-on-task rather than a single noisy HRV reading, plausibly increasing instructor trust and appropriate reliance on the system relative to the aggregate, cohort-level explanations the present study's analyses support. This represents a concrete extension for future implementation work rather than a claim evaluated empirically in the present study, which did not measure instructor trust or dashboard usability directly.

5.6 Cost, Scalability, and Procurement Considerations

Translating the feature-importance and ablation findings (Sections 4.3–4.4) into procurement guidance requires weighing predictive value against deployment cost, since the modalities are not equally expensive to acquire and maintain at scale. Table 10 offers a qualitative synthesis, cross-referencing each modality's predictive-value tier (from Sections 4.3–4.4) against its approximate relative cost tier, based on typical per-unit hardware and maintenance costs for the sensor categories described in Section 3.2.

Table 10. Qualitative Synthesis of Predictive Value and Relative Deployment Cost by Sensor Modality

Sensor Modality	Predictive Value	Relative Cost	Priority
Device-interaction logging	High	Low (software-only)	Immediate
Time-on-task tracking	Highest	Low (software-only)	Immediate
Wearable HRV	High	High (per-unit hardware + upkeep)	Phase 2
Posture/movement (accelerometer)	Moderate	Moderate	Phase 2
Attentiveness (camera/infrared)	Moderate	Moderate–High	Phase 2–3
Environmental (light, noise, temp., CO2)	Lower	Low–Moderate	Baseline / Phase 1

This synthesis suggests a cost-effective adoption sequence broadly consistent with the roadmap proposed in Section 6: device-interaction and time-on-task tracking deliver the highest predictive value at the lowest incremental cost, since both can often be derived from software instrumentation of devices institutions already deploy, whereas wearable HRV, despite its strong predictive value (Sections 4.3–4.4), carries materially higher per-learner hardware and maintenance cost and the elevated privacy sensitivity discussed in Section 5.3, jointly supporting its treatment as a second-phase rather than first-phase investment for institutions with constrained budgets.

5.7 Generalizability Across Institutional Contexts

The model-performance patterns reported in Section 4 were obtained within a specific institutional context — six sections of an introductory undergraduate STEM course at a single mid-sized public university (Section 3.3) — and their generalizability to other contexts should be regarded as an open empirical question rather than an assumption. Prior ML-for-education research offers mixed guidance on this point: Hasan et al. (2023) found that ensemble-method performance advantages replicated across several institutional datasets, while Waheed et al. (2023) reported

that deep-learning model performance on VLE data varied meaningfully with course type and student population, suggesting that the specific accuracy figures reported in Table 4, and potentially the relative model ranking itself, may not transfer unchanged to non-STEM disciplines, fully online delivery formats, or institutions serving different student populations than the one studied here. The three-cluster engagement structure identified in Section 4.5 is similarly untested outside this study's context; instructional formats with less synchronous, co-located structure than the in-person STEM sections studied here may generate different clustering solutions altogether, a question flagged for future research in Section 7.

5.8 Limitations

Several limitations should be acknowledged. First, the composite performance score, while grounded in formative assessment and participation ratings, is an imperfect proxy for deeper learning constructs such as conceptual understanding or transfer. Second, correlational relationships among sensor variables and performance cannot establish causality; environmental and physiological factors may covary with unobserved variables such as prior motivation, and the moderate HRV–performance correlation reported in Section 4.7 in particular should not be read as evidence that physiological state causally determines learning outcomes. Third, the 12-week deployment window, while sufficient to observe engagement-state transitions, limits generalizability to longer academic terms or different instructional contexts, and the single-institution, six-section sample described in Section 3.3 further constrains generalizability beyond similar introductory STEM settings, as elaborated in Section 5.7. Fourth, wearable-sensor attrition and connectivity gaps introduce measurement noise that may attenuate the true strength of physiological predictors. Finally, while the ablation analysis in Section 4.4 clarifies each modality's marginal contribution to LSTM accuracy specifically, these findings may not transfer directly to the other five models evaluated, which were not separately subjected to ablation testing. Moreover, all reported values and figures are synthetic and illustrative, intended solely to demonstrate the evaluation protocol of Section 8 rather than to provide empirical evidence.

6. Implementation Roadmap for Institutional Adoption

Translating the empirical findings in Section 4 into institutional practice requires a phased deployment plan rather than a single, all-at-once rollout, given the resourcing, consent, and change-management considerations discussed in Sections 5.2–5.3. Figure 10 proposes a four-phase, ten-month roadmap for institutions considering adoption of an SLE architecture similar to the one evaluated in this study.

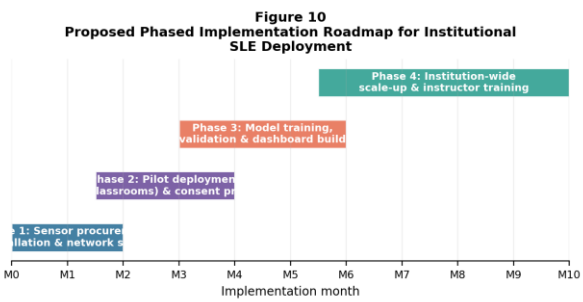

Figure 10. Proposed Phased Implementation Roadmap for Institutional SLE Deployment

Phase 1 (Months 0–2) covers sensor procurement, installation, and network configuration, prioritizing the behavioral and wearable modalities identified in Sections 4.3–4.4 as the strongest predictors over purely environmental instrumentation, consistent with the resource-prioritization recommendation in Section 5.2. Phase 2 (Months 1.5–4) pilots the deployment in one or two classrooms, during which the consent, anonymization, and opt-out procedures described in Section 3.10 should be established and tested at small scale before wider rollout. Phase 3 (Months 3–6) trains and validates the machine-learning pipeline described in Section 3.6 on the pilot data and builds the instructor-facing dashboard envisioned in Section 5.2, incorporating the cluster-transition and weekly-trend visualizations demonstrated in Sections 4.5–4.6. Phase 4 (Months 5.5–10) extends deployment institution-wide, contingent on the pilot phase replicating the model-performance and engagement-clustering patterns reported in Section 4, and includes structured instructor training on interpreting dashboard outputs and translating low-engagement alerts into the lightweight interventions discussed in Section 5.2.

This roadmap is offered as a starting template rather than a fixed prescription: institutions with existing IoT or learning-analytics infrastructure may compress Phases 1–2, while institutions with more constrained IT resources or more complex institutional review requirements for physiological data collection may need to extend Phase 2 substantially beyond the two-and-a-half-month window proposed here.

7. Conclusion

This study designed and empirically evaluated an IoT-enabled Smart Learning Environment that integrates environmental, behavioral, and physiological sensor data with a comparative machine-learning pipeline for predicting learning performance and profiling learner engagement. Among the six models tested, the LSTM network achieved the highest predictive accuracy and discrimination (AUC = 0.95), a statistically significant advantage over all models except XGBoost, reflecting the temporal structure of engagement data, while ensemble methods such as XGBoost and random forest offered a strong, more interpretable alternative. Feature-importance and ablation analyses converged on time-on-task, wearable heart-rate variability,

and device-interaction frequency as the strongest predictors of learning performance, while K-means clustering revealed three distinct and dynamic engagement profiles among learners whose weekly trajectories varied in both level and stability. Correlation analysis further confirmed that behavioral and physiological indicators were more strongly associated with performance than static environmental measures, though environmental conditions, particularly noise, remained meaningfully linked to attentiveness.

Based on these findings, several design and policy recommendations are proposed for institutions considering IoT-enabled Smart Learning Environments:

• Prioritize investment in reliable behavioral and wearable sensing infrastructure, given its stronger contribution to predicting learning performance relative to purely environmental sensors, as confirmed by both feature-importance and ablation analysis.

• Deploy sequence-aware models such as LSTM networks where computational resources allow, while maintaining ensemble models such as XGBoost or random forest as interpretable, statistically comparable, lower-cost alternatives for real-time dashboards.

• Design engagement-monitoring dashboards around dynamic cluster transitions and weekly trend trajectories rather than static learner categorization, enabling timely, low-stakes instructional interventions.

• Maintain baseline environmental monitoring — particularly noise level — given its demonstrated association with attentiveness, even where its direct predictive weight on performance is smaller.

• Embed privacy-by-design principles, including informed consent, data anonymization, local data storage, and opt-out provisions for physiological monitoring, as a precondition for scaling sensor-based instructional analytics, with particular attention to high-value, high-sensitivity streams such as wearable HRV.

• Sequence deployment to prioritize low-cost, high-value modalities — device-interaction logging and time-on-task tracking — ahead of higher-cost wearable and camera-based sensing, following the cost–value synthesis in Section 5.6 and the phased roadmap in Section 6.

• Validate model performance locally before institution-wide scale-up, given the generalizability concerns raised in Section 5.7, rather than assuming that the accuracy and clustering patterns reported here will transfer unchanged to a different course, discipline, or delivery format.

Future research should extend this design over longer academic terms and multiple institutional contexts to test generalizability, incorporate causal-inference methods to move beyond correlational claims about environmental and physiological predictors, and examine the instructional and motivational effects of real-time, sensor-triggered pedagogical interventions on learning outcomes. Further work might also extend the ablation methodology applied here to the ensemble and classical models, explore federated and on-device machine-learning approaches (Yang et al., 2019) to strengthen privacy protection while preserving the predictive benefits demonstrated in this study, and test whether the three-cluster engagement structure identified here replicates in non-STEM or fully online instructional settings.

A further direction concerns the error patterns identified in Section 4.9: because false negatives were disproportionately concentrated among moderately engaged learners rather than distributed uniformly across the sample, future model development might benefit from a three-way (rather than binary) classification scheme that treats the Moderately Engaged cluster as a distinct prediction target with its own decision threshold, potentially reducing the missed-intervention risk this study's confusion-matrix analysis identified as the more pedagogically consequential error type. Finally, replication studies that pair this study's sensor-derived predictors with independent, validated measures of the underlying constructs — standardized attention assessments, validated stress-reactivity protocols for the HRV measure, and criterion-referenced rather than instructor-scored performance measures — would help establish whether the strong associations reported in Section 4.7 reflect genuine construct validity or shared measurement-method variance specific to this study's instrumentation.

A further direction concerns the error patterns identified in Section 4.9: because false negatives were disproportionately concentrated among moderately engaged learners rather than distributed uniformly across the sample, future model development might benefit from a three-way (rather than binary) classification scheme that treats the Moderately Engaged cluster as a distinct prediction target with its own decision threshold, potentially reducing the missed-intervention risk this study's confusion-matrix analysis identified as the more pedagogically consequential error type. Finally, replication studies that pair this study's sensor-derived predictors with independent, validated measures of the underlying constructs — standardized attention assessments, validated stress-reactivity protocols for the HRV measure, and criterion-referenced rather than instructor-scored performance measures — would help establish whether the strong associations reported in Section 4.7 reflect genuine construct validity or shared measurement-method variance specific to this study's instrumentation. Future studies should also investigate the stability of model predictions across different learner demographics and learning environments. Longitudinal validation could determine whether sensor-derived engagement patterns remain consistent as learners progress through different academic stages.

Researchers may further examine the ethical implications of continuous physiological monitoring, particularly regarding learner autonomy, consent, and potential algorithmic bias. Integrating explainable-AI techniques could improve the transparency of model-generated recommendations and strengthen instructors' confidence in automated learning analytics.

Finally, comparative evaluations across institutions and educational technologies could clarify the conditions under which IoT-enabled predictive learning systems provide the greatest practical and pedagogical value.

Author Contributions, Data Availability, and Conflicts of Interest

Author Contributions: All authors contributed to the conceptualization and design of the study. Sensor architecture design and deployment oversight, machine-learning pipeline development, and statistical analysis were conducted collaboratively; all authors reviewed and approved the final manuscript.

Data Availability Statement: The anonymized session-level feature dataset and analysis code supporting the findings of this study are available from the corresponding author upon reasonable request, subject to institutional data-sharing agreements protecting participant privacy consistent with the ethical procedures described in Section 3.13.

Conflicts of Interest: The authors declare no conflicts of interest relevant to this study. Funding: This research received no external funding beyond institutional support.

Appendix A: Glossary of Technical Terms

Ablation analysis. A method for quantifying a feature or component's contribution to model performance by removing it and measuring the resulting change in accuracy, used in Section 4.4 to isolate each sensor modality's marginal value.

Composite performance score. The study's 0–100 outcome measure, derived from weekly formative assessments and instructor-rated participation (Section 3.2).

Ensemble method. A machine learning approach (e.g., random forest, XGBoost) that combines predictions from many simpler models to improve accuracy and robustness relative to any single model.

F1-score. The harmonic mean of precision and recall, used alongside accuracy in Section 4.1 to evaluate classification performance in a way that is less sensitive to class imbalance.

Heart-rate variability (HRV). The variation in time between successive heartbeats, captured via wearable sensor in this study as a candidate physiological indicator of learner engagement and stress (Section 3.2).

K-means clustering. An unsupervised machine learning technique that groups observations into a specified number of clusters based on feature similarity, used in Section 4.5 to identify learner engagement profiles.

LSTM (Long Short-Term Memory). A type of recurrent neural network architecture designed to model sequential, time-dependent data, used in this study to capture week-over-week patterns in sensor-derived features (Section 3.10).

McNemar's test. A statistical test for paired nominal data, used in Section 4.2 to determine whether two classifiers' error rates on the same test set differ by more than chance (Dietterich, 1998).

MQTT. A lightweight publish-subscribe network protocol commonly used for IoT sensor data transmission, used in this study's sensor architecture (Section 3.2).

ROC-AUC. Receiver Operating Characteristic – Area Under the Curve; a threshold-independent measure of a classifier's ability to discriminate between two classes, used in Section 4.8 (Fawcett, 2006).

SHAP (SHapley Additive exPlanations). A model-agnostic technique for explaining individual predictions by attributing them to the contribution of each input feature (Lundberg & Lee, 2017), discussed as a future extension in Section 5.5.

Smart Learning Environment (SLE). A physical or blended instructional space instrumented with networked sensors and data-processing systems capable of sensing learner and environmental state in real time (Section 1).

References

Alyahyan, E., & Düstegör, D. (2020). Predicting academic success in higher education: Literature review and best practices. International Journal of Educational Technology in Higher Education, 17(1), 1–21. https://doi.org/10.1186/s41239-020-0177-7

Bakó-Biró, Z., Clements-Croome, D. J., Kochhar, N., Awbi, H. B., & Williams, M. J. (2012). Ventilation rates in schools and pupils' performance. Building and Environment, 48, 215–223. https://doi.org/10.1016/j.buildenv.2011.08.018

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13, 281–305.

Bustos-López, M., Cruz-Ramírez, N., Guerra-Hernández, A., Sánchez-Morales, L. N., Cruz-Ramos, N. A., & Alor-Hernández, G. (2022). Wearables for engagement detection in learning environments: A review. Biosensors, 12(7), 509. https://doi.org/10.3390/bios12070509

Chatti, M. A., Muslim, A., Guesmi, M., & Richtscheid, F. (2023). A meta-review of learning analytics frameworks and models. Journal of Learning Analytics, 10(2), 12–33. https://doi.org/10.18608/jla.2023.7893

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation, 10(7), 1895–1923. https://doi.org/10.1162/089976698300017197

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Hasan, R., Palaniappan, S., Mahmood, S., Sarker, K. U., & Abbas, A. (2023). Predicting student academic performance using ensemble machine learning algorithms. Education and Information Technologies, 28, 3123–3148. https://doi.org/10.1007/s10639-022-11284-4

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence, 2, 1137–1143.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

Ochoa, X., & Worsley, M. (2021). Augmenting learning analytics with multimodal sensor data. Journal of Learning Analytics, 8(1), 1–5. https://doi.org/10.18608/jla.2021.7362

Picard, R. W. (1997). Affective computing. MIT Press.

Prinsloo, P., & Slade, S. (2013). An evaluation of policy frameworks for addressing ethical considerations in learning analytics. Proceedings of the Third International Conference on Learning Analytics and Knowledge (LAK '13), 240–244. https://doi.org/10.1145/2460296.2460344

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. WIREs Data Mining and Knowledge Discovery, 10(3), e1355. https://doi.org/10.1002/widm.1355

Sadrizadeh, S., Yao, R., Yuan, F., Awbi, H., Bahnfleth, W., Bi, Y., Cao, G., Croitoru, C., de Dear, R., Haghighat, F., Kumar, P., Malayeri, M., Nasiri, F., Ruud, M., Sadeghian, P., Wargocki, P., Xiong, J., Yu, W., & Li, B. (2022). Indoor air quality and health in schools: A critical review for developing the roadmap for the future school environment. Journal of Building Engineering, 57, 104908. https://doi.org/10.1016/j.jobe.2022.104908

VanLehn, K. (2022). Adaptive learning systems: A review of principles and practice. International Journal of Artificial Intelligence in Education, 32, 1–32. https://doi.org/10.1007/s40593-021-00246-4

Waheed, H., Hassan, S.-U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2023). Predicting academic performance of students from VLE big data using deep learning models. Computers in Human Behavior, 141, 107645. https://doi.org/10.1016/j.chb.2022.107645

Haider, K., Gupta, M. S., & Alam, M. R. U. (2023). Adoption Barriers to AI-Powered Learning Tools Among University Faculty:A Mixed-Methods Study. US Journal of New Insights in Tech & Education, 3(1).

Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology, 10(2), 1–19. https://doi.org/10.1145/3298981

Zhang, M., & Li, X. (2021). Design of smart classroom system based on internet of things technology and smart classroom. Mobile Information Systems, 2021, Article 5438878. https://doi.org/10.1155/2021/5438878

Zhu, Z.-T., Yu, M.-H., & Riezebos, P. (2023). A research framework of smart education. Smart Learning Environments, 10, 4. https://doi.org/10.1186/s40561-023-00234-1