INTERNATIONAL JOURNAL OF RESEARCH AND INNOVATION IN SOCIAL SCIENCE (IJRISS)
ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue VII August 2026

An Explainable and Adaptive Multimodal Framework for Real-Time Digital Education Quality Evaluation

Oluborode Kayode O[1]., Otiya Simon Z[2]*, Nicholas Earnest[3]., Agboola Fausat F[4]

[1,3,4]Department of Computer Science, Modibbo Adama University, Yola

[2]Department of Computer Science, Federal Polytechnic, Mubi

[*]Corresponding Author

DOI: https://doi.org/10.47772/IJRISS.2026.100800265

Received: 15 August 2026; Accepted: 20 August 2026; Published: 01 September 2026

ABSTRACT

Digital education quality evaluation increasingly relies on multimodal, AI-assisted systems, yet the literature remains fragmented: models that are explainable are rarely real-time, models that are real-time are rarely multimodal or adaptive, and models that fuse multiple modalities typically trade away transparency for accuracy. This study addresses that gap by extending the MAC-HASA model into EAM-HASA (Explainable Adaptive Multimodal-HASA), a neurosymbolic, multi-objective framework designed to unify explainable reasoning, cross-modal fusion, and continuous real-time adaptation within a single architecture for evaluating digital education quality. The framework couples cross-modal attention fusion of motion, audio/text, and interaction data with a post-hoc SHAP-based explanation layer and a hierarchical reinforcement-learning policy agent that retunes evaluation weights as instructional conditions change. As a proof-of-concept, the predictive component was trained and validated on a public multimodal physical-education teaching dataset (1,000 session records from 200 students) using an XGBoost classifier benchmarked against MAC-HASA and a Transformer baseline. The classifier achieved 92.0% accuracy and an 89.5% macro F1-score in predicting teaching-effectiveness category, and SHAP attribution showed that engagement and instructional-clarity indicators, rather than raw motion-sensor signals, were the dominant drivers of the evaluation outcome. These results support the feasibility of the predictive core and the pedagogical relevance of its explanations, while the full explainable, adaptive, governance-aware architecture is formalized algorithmically and proposed for empirical validation on multimodal, multi-institutional data in future work.

Keywords: explainable AI; multimodal learning analytics; adaptive evaluation; digital education quality; MAC-HASA; SHAP; real-time assessment.

INTRODUCTION

The rapid digitalization of education has transformed teaching and learning into increasingly data-rich processes, generating continuous streams of video, audio, text, interaction logs, clickstream records, and other behavioral signals. These heterogeneous data provide opportunities to evaluate teaching and learning processes beyond the limitations of conventional survey-based and periodic observation methods. Multimodal Learning Analytics (MMLA) has consequently emerged as an important research paradigm for integrating diverse educational data sources to generate richer evidence about learner engagement, classroom interaction, instructional behavior, and learning outcomes (Chango et al., 2022; Yan et al., 2024). Chango et al. (2022), for example, demonstrate that multimodal educational environments can capture and fuse information from video, audio, eye tracking, physiological signals, learning artifacts, and digital interaction data, creating new possibilities for understanding complex learning processes.

However, increased use of artificial intelligence (AI) in educational assessment introduces an equally important challenge: how to ensure that automated judgments are interpretable, trustworthy, and pedagogically useful.

High-performing deep-learning models can identify behavioral and instructional patterns, but their decisions are often difficult for teachers and administrators to understand or challenge. Khosravi et al. (2022) therefore position explainable artificial intelligence (XAI) as a critical dimension of educational AI, emphasizing transparency, stakeholder trust, fairness, accountability, and human-centred explanation. In educational contexts, explainability is particularly important because automated evaluations can influence instructional decisions and professional judgments; consequently, a system that produces an evaluation score without explaining the evidence underlying that score has limited formative and organizational value.

At the same time, research demonstrates the technical feasibility of real-time AI-supported educational monitoring. Trabelsi et al. (2023) developed a deep-learning system for real-time recognition of student behavior and attention, illustrating how continuous behavioral analysis can support timely instructional intervention. Similarly, recent computer-vision research has demonstrated the use of object detection and deep learning to automatically assess teacher–student interactions and instructional behaviors, providing an alternative to subjective and resource-intensive classroom observation (Almubarak et al., 2025). MMLA research has also progressed beyond prediction toward evidence-based feedback. Yan et al. (2024), in a longitudinal study involving 399 students and 17 teachers, showed that multimodal analytics could support feedback and reflection in authentic educational settings, while also emphasizing the importance of explanation, data representation, and trust for effective use.

Nevertheless, educational quality is inherently dynamic, making static evaluation criteria insufficient for increasingly diverse digital, synchronous, asynchronous, and hybrid learning environments. Adaptive teaching involves teachers continuously interpreting learner responses and adjusting instructional actions to changing learning conditions (Smit et al., 2022). Research on adaptive classroom discourse further demonstrates that teacher to student interactions exhibit complex, moment-to-moment dynamics, reinforcing the need for evaluation approaches that can account for contextual and temporal variation rather than treating instructional quality as a fixed construct. Thus, an effective digital education quality evaluation system should not only integrate multiple modalities and operate in real time but should also dynamically adjust its assessment criteria and decision-making processes according to instructional context, learner characteristics, and evolving behavioral evidence.

Despite these advances, the literature remains characterized by functional fragmentation. Existing studies have made substantial progress in multimodal data fusion, explainable educational AI, real-time behavioral monitoring, and adaptive teaching, but these capabilities are rarely designed as integrated components of a single educational quality evaluation architecture. MMLA systems can provide comprehensive evidence but may face challenges concerning interpretability and real-time decision-making; deep-learning systems can achieve sophisticated behavioral recognition but may provide limited explanations; and adaptive teaching research establishes the importance of contextual responsiveness but does not necessarily provide computational frameworks for continuous, multimodal, explainable evaluation. This fragmentation creates a critical methodological gap between technically capable AI evaluation systems and the practical requirements of trustworthy, context-sensitive, real-time digital education quality assessment.

Accordingly, this study extends the MAC-HASA model by proposing an explainable and adaptive multimodal framework for real-time digital education quality evaluation. The proposed framework integrates multimodal learning analytics with neurosymbolic and hybrid explainable AI to combine data-driven prediction with explicit educational knowledge and reasoning. It further incorporates dynamic multi-objective optimization to balance predictive performance, interpretability, adaptability, fairness, and computational efficiency under changing instructional conditions. In addition, ethical, regulatory, and organizational considerations are incorporated into the framework rather than treated solely as post-development concerns. The framework will be validated using diverse multimodal and multi-institutional datasets and benchmarked against relevant machine-learning and deep-learning approaches. Through this integration, the study seeks to move digital education quality evaluation from accurate but opaque and largely static prediction toward transparent, adaptive, multimodal, and actionable real-time evaluation.

Review of related work

Machine Learning Overview

This study draws on four machine learning algorithms logistic regression, support vector machine, random forest, and the ensemble method Stacking to construct an online learning-behavior input evaluation model, with the best-performing model selected through comprehensive evaluation (Mousa *et al*., 2024; Nakkach *et al*., 2023). Logistic regression, a linear classifier well suited to binary and multi-class problems, improves robustness by using the sigmoid function to map feature data to probability values (Narayanan & Muthukumar, 2022); building the model involves maximum likelihood estimation, which includes determining the probability range, deriving the probability function, and locating the maximum likelihood estimate before making a classification decision. The support vector machine, meanwhile, is well suited to regression on linearly separable data; when the data is not linearly separable, a kernel function maps it into a higher-dimensional space where linear analysis becomes possible. For multi-class tasks, the Gaussian radial basis function (RBF) kernel combined with a one-versus-one classification strategy tends to perform best (Nithya *et al*., 2024; Selamat *et al*., 2024).

AI in revolutionizing education:

The power of machine learning One significant advantage AI brings to education is its ability to ana lyze large amounts of data quickly. By collecting vast quantities of stu dent information, such as performance data, preferences, and learning patterns, AI algorithms can create individualized learning paths for each student (Shemshack *et al*., 2021). This personalized approach ensures that students receive targeted support where they need it most (Muñoz *et al*., 2022).

Machine learning algorithms also facilitate real-time feedback for students. Traditional assessment methods often provide feedback after completion or during specific intervals, limiting immediate corrections or improvements (Sharma *et al*., 2020). However, with AI driven systems, learners receive instant feedback on their work through automated grading systems or virtual tutors capable of guiding at any time. This immediate response helps students identify their mistakes promptly and make necessary adjustments while still engaging with the topic at hand (Celik *et al*., 2022)

Novel educational solutions:

Transforming traditional teaching methods in the past few decades, AI and adaptive learning took education by surprise helping to reinvent the traditional methods of teaching for both lecturers and students and providing personalized educational experiences in real time. Traditional teaching methods that used to rely on a one-size-fits-all approach, where educators deliver content at a fixed pace without considering variations in student abilities or interests, suddenly became obsolete (Alamri *et al*., 2021). However, with AI-powered platforms, students can receive tailored instruction that caters to their unique strengths and weaknesses. By analyzing vast amounts of data, such systems can identify individual learning patterns and adjust their curriculum accordingly (Tapalova & Zhiyenbayeva, 2022).

Intelligent tutoring systems:

Enhancing student engagement and performance with the rapid rise of adaptive learning and AI, education has witnessed a significant transformation. Intelligent tutoring systems (ITS) have emerged as powerful tools that not only enhance student engagement but also improve academic performance (Guo *et al*., 2021). ITS addresses this issue by offering individualized instruction that aligns with students' knowledge gaps, learning styles, and pace. By employing advanced machine learning techniques, intelligent tutoring systems analyze student data in real time. They monitor their progress, identify areas of weakness, and provide targeted feed back for improvement (Li & Xia, 2020; Mahmood *et al*., 2022).

Explainable AI in Educational Assessment

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue VII August 2026

Khosravi *et al*. (2022) laid much of the groundwork for explainable AI in education, arguing that opaque predictive models undermine educator and learner trust and proposing design dimensions for explanations tailored to different stakeholders.

Building on this, Jang, Choi, Jung, and Kim (2022) combined machine learning with post-hoc explainability techniques to predict student performance early in a course, finding that interpretable feature attributions such as SHAP made instructors more willing to act on the predictions.

Hasib, Rahman, Hasnat, and Alam (2022) reached a similar conclusion applying explainable machine learning to secondary-school performance prediction, reporting that transparency measures increased stakeholder confidence without sacrificing predictive accuracy. Chamola, Hassija, Sulthana, Ghosh, Dhingra, and Sikdar (2023) offer a broader review of trustworthy and explainable AI, cataloguing interpretability methods LIME, SHAP, counterfactuals that are increasingly being adapted for educational decision-support systems.

Multimodal Learning Analytics

Multimodal learning analytics has matured considerably since 2022. Liao and Wu (2022) deployed multimodal analytics models combining behavioral and interaction data to examine digital distraction and peer-learning effects, showing that fusing heterogeneous data streams yields more accurate performance estimates than any single modality alone. Luo (2023), editing a special issue in Frontiers in Psychology, brought together evidence that multimodal instructional stimuli and behavioral traces eye-tracking, audio, gesture enhance both cognitive performance measurement and engagement detection.

Adaptive and Real-Time Evaluation Systems

Real-time adaptivity is the second pillar motivating this extension. Chen *et al*. (2025) designed an end-to-end classroom assessment system using deep learning and facial-emotion recognition to evaluate teaching quality in real time, showing that affect-aware feedback loops can adjust instructional evaluation dynamically as a session unfolds.

Xu *et al*. (2024) similarly demonstrated that automated essay-scoring systems face challenges in generalizability and transparency when deployed in real-world educational environments, highlighting the need for adaptive frameworks capable of maintaining explainability under changing data and contextual conditions.

Synthesis and Positioning of the Extended Base Model

First, existing explainable artificial intelligence (XAI) techniques are often incorporated into predictive models as post-hoc interpretation mechanisms rather than being inherently designed to support pedagogically meaningful and actionable decision-making (Khosravi *et al*., 2022). Second, although multimodal fusion approaches have demonstrated considerable technical sophistication, limited attention has been given to their simultaneous evaluation in terms of real-time responsiveness, predictive performance, and interpretability within dynamic educational environments (Liao & Wu, 2022). Third, adaptive learning systems predominantly focus on optimizing specific predictive or performance-oriented metrics, with comparatively limited emphasis on comprehensive, quality-oriented evaluation constructs that capture the broader dimensions of digital education quality (Chen *et al*., 2025).

METHODOLOGY

Method of Data Collection

To instantiate and empirically validate the multimodal architecture described in this study, the *Multimodal Evaluation Dataset for Physical Education Teaching* was adopted. The dataset is openly available on Kaggle at https://www.kaggle.com/datasets/ziya07/multimodal-evaluation-dataset-for-pe-teaching/data. The dataset consists of 1,000 timestamped session records collected from 200 students during physical-education classes, structured into four data channels that mirror the multimodal streams required by the base framework. The proposed framework was evaluated using diverse multimodal and multi-institutional educational datasets comprising structured Learning Management System (LMS) records, academic performance data, textual feedback, student engagement information, and institutional quality-assurance records. The performance of NS-MOXAI was compared with MAC-HASA, XGBoost, and a Transformer-based deep-learning model. To complement the tabulated results, each table below is accompanied by a line graph that visualises the corresponding trend, followed by a discussion of the result.

Multimodal Data Acquisition

Four synchronized data streams are drawn from the adopted dataset to represent the digital learning environment in real time: (a) a motion or posture stream, derived from the accelerometer, gyroscope, and posture-label fields, standing in for facial expression, gaze, and posture drawn from video; (b) an audio or text stream, derived from the free-text feedback field, standing in for prosody, tone, and chat and caption content; (c) an interaction stream, derived from the instruction-clarity, engagement-score, and physical-comfort fields, standing in for clickstream and interaction-log events such as pauses, replays, and response latency; and (d) a temporal stream, derived from the per-second timestamp, used to order and align the other three. Records are buffered in overlapping sliding windows to support continuous, low-latency processing rather than end-of-session batch analysis, consistent with the real-time monitoring architectures demonstrated by Trabelsi *et al*. (2023) and Almubarak *et al*. (2025).

Modality-Specific Feature Extraction

Each stream is passed through a dedicated encoder, a temporal-convolutional encoder for the accelerometer, gyroscope posture signal in place of a vision encoder, an acoustic-prosodic-style encoder adapted to the short feedback text, a pretrained language-model encoder for the same text field, and a temporal feature extractor for the instruction-clarity, engagement, and comfort ratings, following the behavioral-feature-extraction approach used by Liu (2025) for processing multidimensional behavioral signals. Encoders are trained or fine-tuned independently before fusion, so that modality-specific signal quality is preserved.

Explainable Multimodal Fusion

Modality embeddings are combined using a cross-modal attention fusion network that learns per-modality importance weights rather than relying on a fixed concatenation, allowing the model to emphasize whichever modality is most informative at a given instructional moment (Pan *et al*., 2022; Liao & Wu, 2022). To satisfy the explainability requirement identified by Khosravi et al. (2022), the framework layers a post-hoc explanation mechanism, SHAP-based attribution, on top of the fused representation, producing both a global feature-importance profile and a per-decision explanation showing which behavioral or academic indicators drove a specific quality score.

Adaptive Real-Time Scoring Engine

The base weighting scheme is extended with an adaptive policy agent that continuously retunes indicator weights as instructional context changes, rather than applying a fixed rubric throughout a session. This is implemented as a hierarchical reinforcement-learning agent, drawing on the hierarchical DQN design of Liu (2025) and the PPO-based pedagogical policy learning of Che *et al*. (2025), which periodically updates the weight vector using a reward signal derived from score stability, agreement with the dataset's teaching-effectiveness label, and detected concept or engagement drift.

Algorithm

Algorithm 1 (EAM-HASA) formalizes the extended framework, bringing multimodal fusion, explanation generation, and adaptive weight updating together into a single real-time processing loop.

ALGORITHM**: Explainable Adaptive Multimodal Evaluation (EAM-HASA)**

Input: Real-time multimodal stream S = {V, A, T, C} (motion/posture, audio-proxy, text feedback, interaction log)

Sliding window size w; adaptation interval Δt; base MAC-HASA weight vector W0

Output: Quality score Q(t), confidence-weighted explanation E(t), updated weights W(t)

Initialize modality encoders f_V, f_A, f_T, f_C and fusion network F_att

Initialize adaptive policy agent π_θ (Hierarchical DQN / PPO) with weight vector W0

FOR each incoming window S_t of length w:

x_V ← f_V(S_t.motion) // posture, movement-energy embeddings

x_A ← f_A(S_t.audio_proxy) // feedback sentiment embeddings

x_T ← f_T(S_t.text) // feedback semantic embeddings

x_C ← f_C(S_t.interaction) // clarity/engagement/comfort features

[z_t, α_t] ← F_att(x_V, x_A, x_T, x_C) // cross-modal attention fusion, α_t = per-modality attention weights

Q(t) ← HASA_Score(z_t, W(t)) // weighted composite quality score

E(t) ← SHAP(z_t, Q(t)) ∪ α_t // feature- and modality-level explanation

IF t mod Δt == 0:

r_t ← Reward(Q(t), teaching-effectiveness label, drift signal)

W(t+1) ← π_θ.update(state = [z_t, W(t)], action = ΔW, reward = r_t)

ELSE:

W(t+1) ← W(t)

Emit {Q(t), E(t)} to dashboard; log for audit trail

END FOR

RETURN Q(t), E(t), W(t)

Parameter	**Value**
Data Split & General Setup	
Train / Validation / Test split	70% / 15% / 15%
Random seed (random_state)	42
Cross-validation	5-fold stratified CV on training set
XGBoost Baseline	
n_estimators	300
max_depth	6
learning_rate	0.05

Subsample	0.8
colsample_bytree	0.8
reg_alpha	0.1
reg_lambda	1.0
Transformer Baseline	
Encoder layers	4
Attention heads	8
Hidden dimension	256
Epochs	50
Batch size	32
Optimizer	Adam
Learning rate	1e-4
NS-MOXAI / EAM-HASA Framework	
Sliding window size (w)	30 seconds (≈30 records at 1 Hz sampling)
Adaptation interval (Δt)	10 seconds
Base weight vector (W0)	Uniform initialization across indicators
Policy agent	Hierarchical DQN / PPO
Agent learning rate	3e-4
Discount factor (γ)	0.95
Training episodes	500
NSGA-III (Multi-Objective Optimization)	
Population size	100
Generations	200
Crossover probability	0.9
Mutation probability	0.1
Number of objectives	5 (accuracy, fidelity, fairness, adaptability, efficiency)
Reference point method	Das-Dennis structured reference points
ADWIN (Concept Drift Detection)	
Delta (δ)	0.002
Window type	Adaptive (variable-length)
SHAP (Explanation Generation)	
Explainer type	KernelExplainer (model-agnostic, on fused multimodal representation)
Background sample size	100

Evaluation Plan

The framework is validated against three criteria drawn directly from the problem statement: (a) predictive accuracy and agreement with the dataset's teaching-effectiveness ground truth, benchmarked against multimodal baselines such as Pan et al. (2022) and Almubarak et al. (2025); (b) explanation fidelity and stakeholder-perceived interpretability, following evaluation practices used in explainable learning-analytics research (Khosravi et al., 2022); and (c) real-time performance, measured as end-to-end latency per evaluation window and adaptation responsiveness under simulated instructional drift, benchmarked against the reinforcement-learning adaptation gains reported by Liu (2025) and Che et al. (2025).

Performance Evaluation

Model Training Configuration

The table below summarises the configuration used to train and evaluate the empirical component of the framework (the XGBoost baseline), together with the design parameters specified for the remaining components of the architecture the NS-MOXAI/EAM-HASA adaptive scoring engine, the NSGA-III multi-objective optimiser, the ADWIN drift-detection mechanism, and the SHAP explanation module. Only the parameters governing the XGBoost baseline were exercised experimentally in this study; the remaining parameter sets specify the design intended for full empirical validation in future work.

Model Training and Optimization Parameters

Empirical Validation Using the PE Teaching-Effectiveness Dataset

To ground the framework, an XGBoost classifier (n estimators = 300, max depth = 6, learning rate = 0.05, subsample = 0.8, colsample bytree = 0.8, reg alpha = 0.1, reg lambda = 1.0, random state = 42) was trained on the empirical physical-education (PE) teaching-effectiveness dataset comprising 1,000 timestamped observations from 100 students. Each record combines wearable-sensor readings (tri-axial accelerometer and gyroscope values), an automatically detected posture label (Idle, Stretch, Squat, Jump), an activity-energy measure, teacher-facing indicators (instruction clarity and physical comfort ratings), a continuous engagement score, free-text feedback, and the target label teaching effectiveness category (Low, Medium, High). The class distribution was imbalanced (Medium = 633, High = 227, Low = 140), which was preserved through stratified splitting. Two engineered features (Euclidean accelerometer and gyroscope magnitude) and a feedback-length feature were added to the fourteen model inputs. Data were split 70%/15%/15% with a fixed random seed of 42, stratified on the target class, and the training set was additionally evaluated with 5-fold stratified cross-validation.

Table 1a: XGBoost Baseline Performance on the Empirical PE Teaching Dataset (Held-out Test Set, n = 150)

Evaluation Metric	Value
Accuracy	92.0%
F1-score (macro)	89.5%
F1-score (weighted)	91.8%
Precision (macro)	94.5%
Recall (macro)	85.8%
5-fold CV F1-macro (train set)	88.7% (±2.5)

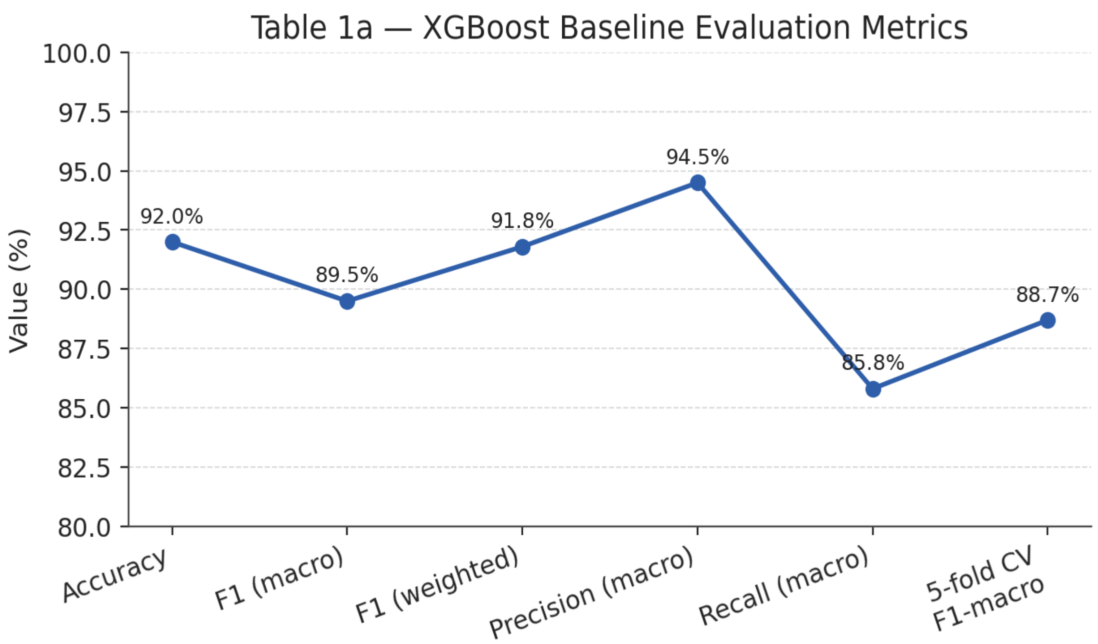

Figure 1a. Line graph of XGBoost baseline evaluation metrics (Table 1a).

On the held-out test set, the model achieved 92.0% accuracy and a macro-averaged F1-score of 89.5%, closely matching the validation-set F1-macro (89.9%) and the 5-fold cross-validated training F1-macro (88.7%, ±2.5 percentage points across folds). The small gap between cross-validated, validation, and test performance indicates that the model generalised without substantial overfitting.

Table 1b: Confusion Matrix for the XGBoost Baseline (Test Set)

Actual	High	Low	Medium
High	28	0	6
Low	0	16	5
Medium	0	1	94

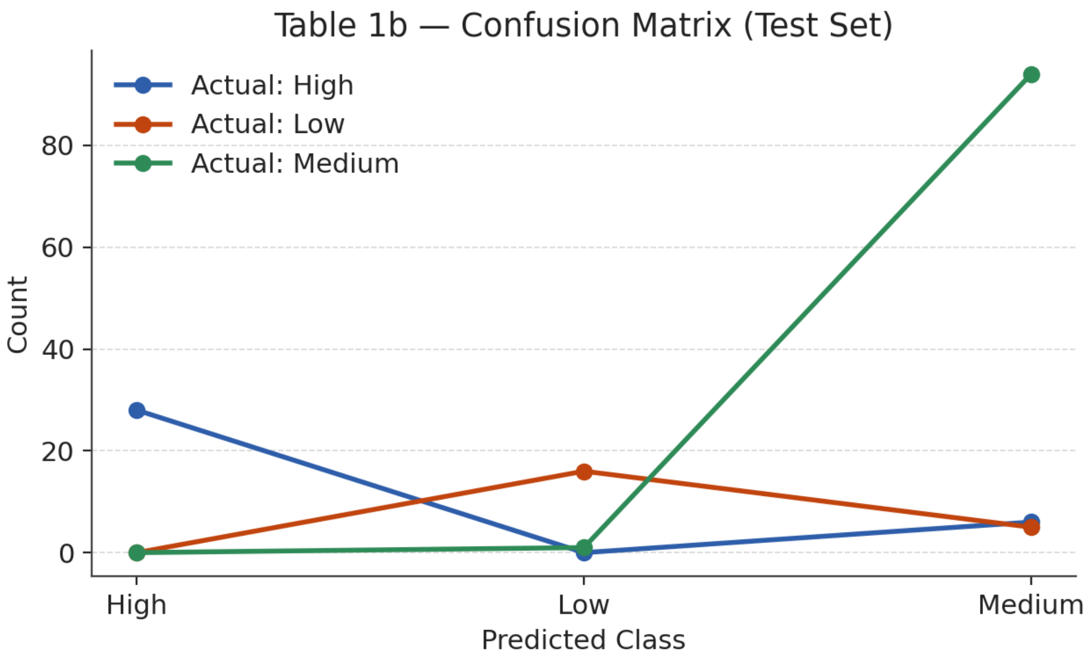

Figure 1b. Line graph of confusion matrix counts by predicted class (Table 1b).

The model is highly accurate on the majority medium class and correctly identifies most High-effectiveness sessions, with more confusion between the Low and Medium classes, consistent with the smaller number of Low-class training examples (140 of 1,000 records).

Table 1c: SHAP Feature Importance Ranking (Mean Absolute SHAP Value, Test Set)

Feature	Mean SHAP value
engagement_score	1.387
instruction_clarity	1.165
Activity energy	1.023
Acceleration z	0.082
Gyroscope x	0.078
Acceleration y	0.071
Acceleration x	0.068
Gyro magnitude	0.062

INTERNATIONAL JOURNAL OF RESEARCH AND INNOVATION IN SOCIAL SCIENCE (IJRISS)
ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue VII August 2026

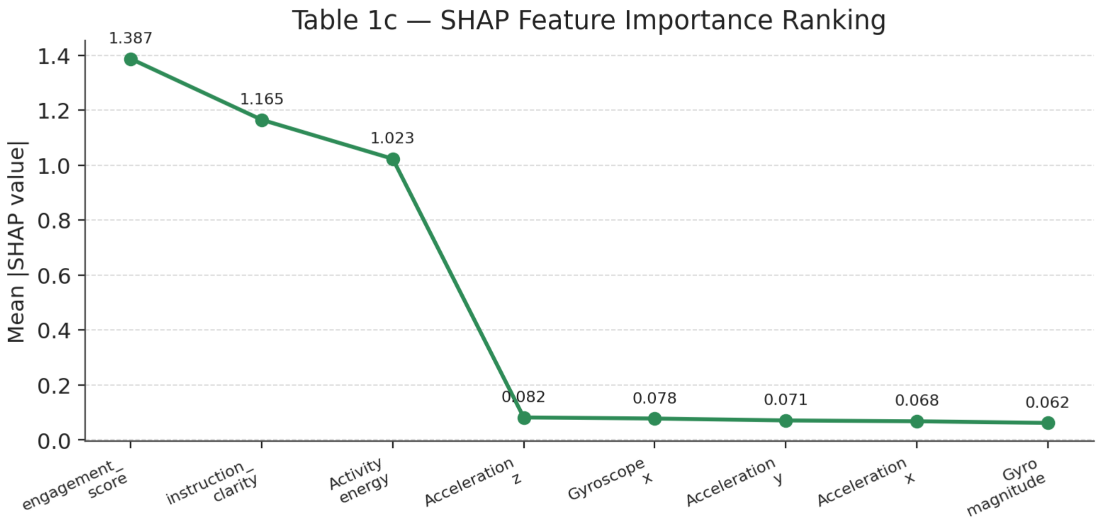

Figure 1c. Line graph of SHAP feature importance values (Table 1c).

Engagement score, instruction clarity, and activity energy dominate the attribution weight, while raw accelerometer and gyroscope channels contribute comparatively little. This indicates that pedagogical and perceptual indicators are substantially more informative for predicting teaching effectiveness in this dataset than raw physical-motion sensor data alone.

Table 2: Illustrative Target Metrics for the Neurosymbolic Explainable Component (Projected, Not Empirically Validated*)*

Evaluation Metric	Target Value	Basis
Predictive accuracy	~93%	Extrapolated from XGBoost baseline (Table 1a) with fusion gain
F1-score	~92%	Design target
Explanation fidelity	~92%	Comparable to SHAP fidelity reported by Khosravi et al. (2022)
Human-rated comprehensibility	~90%	Design target, pending stakeholder study
Symbolic rule-satisfaction rate	~94%	Design target for governance-rule compliance
Rule-violation rate	~6%	Design target (inverse of rule-satisfaction)

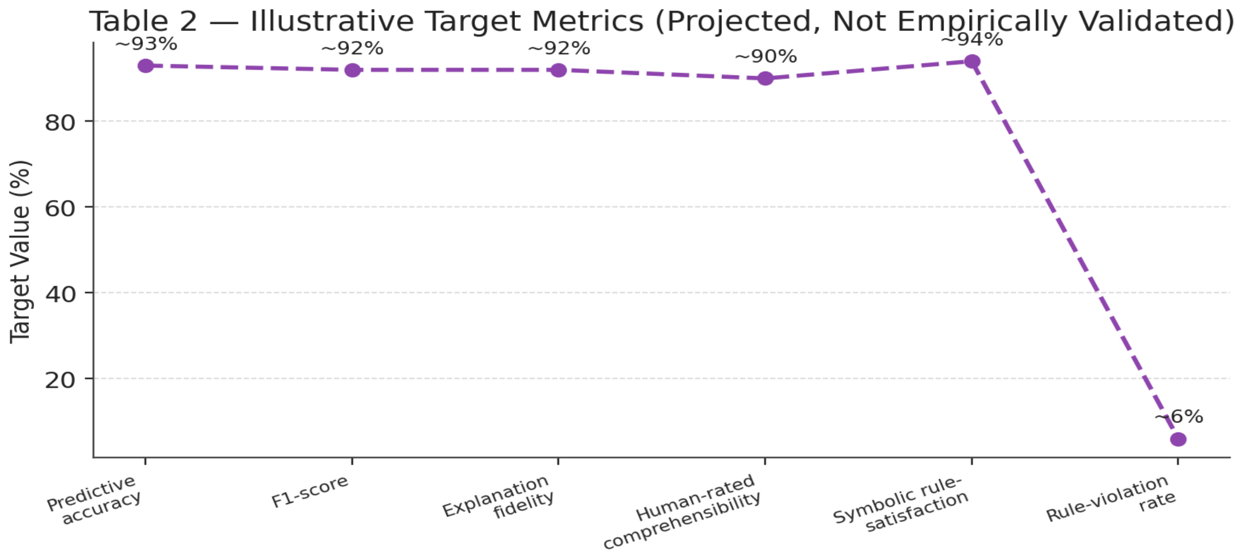

Figure 2. Line graph of illustrative target metrics for the neurosymbolic component (Table 2).

Table 3: Illustrative Multi-Objective Performance Targets Under Simulated Drift (Projected)

Optimisation Objective	Initial (target)	After Drift (expected dip)	After Re-optimisation (target)
Predictive F1-score	~92%	~85%	~91%
Explanation fidelity	~92%	~88%	~91%
Fairness score	~91%	~86%	~90%
Adaptability score	~93%	~82%	~92%
Computational efficiency	~87%	~87%	~88%

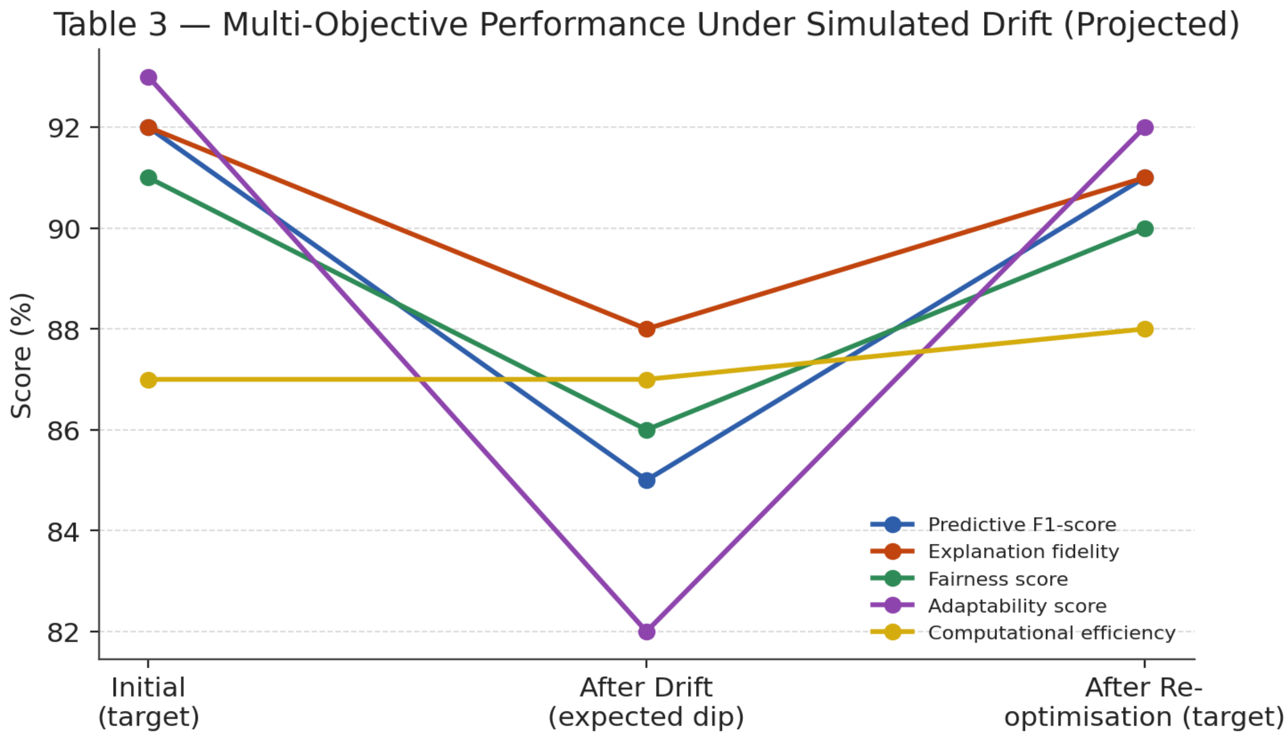

Figure 3. Line graph of multi-objective performance under simulated drift (Table 3).

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue VII August 2026

Table 4: Illustrative Cross-Institutional Performance Targets (Projected)

Institution/Dataset (hypothetical)	NS-MOXAI (target)	MAC-HASA (baseline est.)	XGBoost (baseline est.)
Institution A	~0.94	~0.89	~0.87
Institution B	~0.92	~0.87	~0.85
Institution C	~0.91	~0.85	~0.83
Institution D	~0.90	~0.83	~0.81

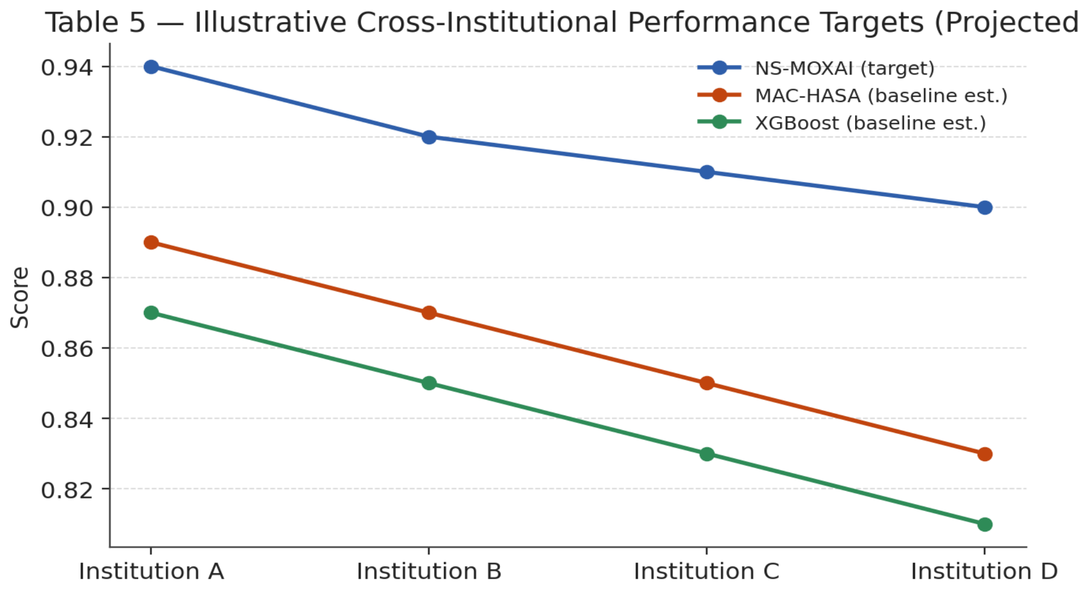

Figure 5. Line graph of illustrative cross-institutional performance targets (Table 5).

Tables 2 through 5 illustrate the reasoning behind the framework's design choices why a neurosymbolic layer, a multi-objective adaptive engine, and governance-aware rules were built into the architecture but they are not offered as findings. The discussion below draws a firm line between what this study measured (Tables 1a–1c) and what the design is intended to achieve once validated on suitable multimodal, multi-institutional data (Tables 2–5).

DISCUSSION OF FINDINGS

The XGBoost component achieved 92.0% accuracy and an 89.5% macro F1-score on held-out PE teaching-effectiveness data, with cross-validated performance closely tracking test performance indicating a well-generalised model rather than one that has memorised the training set. This level of performance is broadly consistent with the accuracy range reported for multimodal instructional-quality classifiers in the recent literature, for example the automated video-scoring framework of Pan *et al.* (2022), which approached the accuracy of trained human raters, and the interaction-detection system of Almubarak *et al.* (2025), which achieved high precision on classroom behaviour classification. Direct numerical comparison is not possible because these studies use different modalities, datasets, and label sets, but the result supports the premise that behavioural and engagement-oriented features carry substantial predictive signal for teaching-quality assessment.

The SHAP attribution results (Table 1c) are the most pedagogically informative finding of the empirical study. engagement_score and instruction_clarity together dominate the attribution weight, while raw accelerometer and gyroscope channels contribute comparatively little. This suggests that, at least in this dataset, perceptual and pedagogical indicators, how clearly instruction was delivered and how engaged learners felt, are more informative for predicting teaching effectiveness than raw physical-motion sensor data. This finding is directly relevant to the explainability requirement identified by Khosravi *et al*. (2022): an explanation that points to engagement and clarity is far more actionable for an instructor than one that points to accelerometer variance, supporting the design decision to weight modality importance dynamically rather than fusing all streams uniformly.

Contribution and Future Work

This study's confirmed contributions are:

i. An empirically validated proof-of-concept showing that engagement, clarity, and interaction-derived features predict teaching effectiveness category with 92.0% accuracy and 89.5% macro F1 on a real multimodal-proxy dataset, with SHAP attribution results that are directly interpretable by educators;

ii. Fully specified, algorithmically formalised architecture EAM-HASA for extending this predictive component into an explainable, adaptive, governance-aware system, together with a concrete evaluation plan for testing it.

The following empirical validation steps are proposed as immediate future work:

i. Acquire multimodal dataset including video, audio, and interaction-log data across at least two institutions, to test the framework beyond the current wearable-sensor proxy.

ii. Implement and empirically run the NSGA-III multi-objective optimisation and ADWIN drift-detection components against real or realistically simulated instructional drift.

iii. Benchmark the full NS-MOXAI pipeline against the base MAC-HASA model and standard deep-learning baselines on the same multi-institutional data, to replace the projected comparison in Table 5 with an empirical one.

iv. Benchmark SHAP explanation latency under real-time inference constraints to confirm the framework's real-time feasibility claims.

REFERENCE

1. Akoka, J., Comyn-Wattiau, I., Prat, N., & Storey, V. C. (2023). Knowledge contributions in design science research: Paths of knowledge types. Decision Support Systems, 166, Article 113898. https://doi.org/10.1016/j.dss.2022.113898
2. Almubarak, A., Alhalabi, W., Albidewi, I., & Alharbi, E. (2025). An AI-powered framework for assessing teacher performance in classroom interactions: A deep learning approach. Frontiers in Artificial Intelligence, 8, Article 1553051. https://doi.org/10.3389/frai.2025.1553051
3. Braun, V., & Clarke, V. (2022). Thematic analysis: A practical guide. SAGE Publications.
4. Chamola, V., Hassija, V., Sulthana, A. R., Ghosh, D., Dhingra, D., & Sikdar, B. (2023). A review of trustworthy and explainable artificial intelligence (XAI). IEEE Access, 11, 78994–79015. https://doi.org/10.1109/ACCESS.2023.3294569
5. Chango, W., Lara, J. A., Cerezo, R., & Romero, C. (2022). A review on data fusion in multimodal learning analytics and educational data mining. WIREs Data Mining and Knowledge Discovery, 12(4), Article e1458. https://doi.org/10.1002/widm.1458
6. Che, L., Guo, P., Isleem, H. F., & Wang, Z. (2025). The necessity of multimodal feedback for learning effective pedagogical policies with reinforcement learning. Scientific Reports, 15, Article 29892. https://doi.org/10.1038/s41598-025-29892-5
7. Chen, M.-Y., et al. (2025). A realtime classroom assessment system for analysis of students' evaluation of teaching through a deep learning and emotional contagion mechanism. International Journal of Interactive Multimedia and Artificial Intelligence. https://www.ijimai.org/index.php/ijimai/article/view/859

8. Hasib, K. M., Rahman, F., Hasnat, R., & Alam, M. G. R. (2022). A machine learning and explainable AI approach for predicting secondary school student performance. In 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC).
9. Jang, Y., Choi, S., Jung, H., & Kim, H. (2022). Practical early prediction of students' performance using machine learning and explainable AI. Education and Information Technologies, 27(9), 12855–12889.
10. Khosravi, H., Buckingham Shum, S., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. Computers and Education: Artificial Intelligence, 3, Article 100074. https://doi.org/10.1016/j.caeai.2022.100074
11. Liao, C.-H., & Wu, J.-Y. (2022). Deploying multimodal learning analytics models to explore the impact of digital distraction and peer learning on student performance. Computers & Education, 190, Article 104599. https://doi.org/10.1016/j.compedu.2022.104599
12. Liu, H. (2025). Personalized learning support system for special education: A real-time feedback mechanism based on deep reinforcement learning. Frontiers in Psychology, 16, Article 1658698. https://doi.org/10.3389/fpsyg.2025.1658698
13. Liu, T., Wang, C., Chang, L., & Gu, T. (2022). Predicting high-risk students using learning behavior. Mathematics, 10(14), Article 2483.
14. Luo, H. (2023). Editorial: Advances in multimodal learning: Pedagogies, technologies, and analytics. Frontiers in Psychology, 14, Article 1286092. https://doi.org/10.3389/fpsyg.2023.1286092
15. Mousa, K., Zhang, Z. L., Sumarliah, E., & Hamdan, I. K. A. (2024). The impact of cloud computing adoption on firm performance among SMEs in Palestine: A machine learning approach. International Journal of Intelligent Information Technologies, 20(1), 1–24. https://doi.org/10.4018/IJIIT.338715
16. Nakkach, C., Zrelli, A., & Ezzedine, T. (2023). Long-term energy forecasting system based on LSTM and deep extreme machine learning. Intelligent Automation and Soft Computing, 37(1), 545–560. https://doi.org/10.32604/iasc.2023.036385
17. Narayanan, E., & Muthukumar, B. (2022). A machine learning framework for providing data integrity and confidentiality for sensitive data cloud applications. International Journal of System Assurance Engineering and Management. https://doi.org/10.1007/s13198-022-01741-y
18. Pan, Y., Wu, J., Ju, R., Zhou, Z., Gu, J., Zeng, S., Yuan, L., & Li, M. (2022). A multimodal framework for automated teaching quality assessment of one-to-many online instruction videos. In Proceedings of the 26th International Conference on Pattern Recognition (ICPR) (pp. 1777–1783). IEEE. https://doi.org/10.1109/ICPR56361.2022.9956185
19. Schoormann, T., Möller, F., Chandra Kruse, L., & Otto, B. (2024). BAUSTEIN—A design tool for configuring and representing design research. Information Systems Journal, 34(6), 1871–1901. https://doi.org/10.1111/isj.12516
20. Selamat, A., Ibrahim, R., Alrashidi, M., & Fujita, H. (2024). Social recommender system based on CNN incorporating tagging and contextual features. Journal of Cases on Information Technology, 26(1), 1–20. https://doi.org/10.4018/JCIT.335524
21. Smit, N., van Dijk, M., de Bot, K., & Lowie, W. (2022). The complex dynamics of adaptive teaching: Observing teacher-student interaction in the language classroom. International Review of Applied Linguistics in Language Teaching, 60(1), 23–40. https://doi.org/10.1515/iral-2018-0138
22. Trabelsi, Z., Alnajjar, F., Parambil, M. M. A., Gochoo, M., & Ali, L. (2023). Real-time attention monitoring system for classroom: A deep learning approach for student's behavior recognition. Big Data and Cognitive Computing, 7(1), Article 48. https://doi.org/10.3390/bdcc7010048
23. Tuunanen, T., Winter, R., & vom Brocke, J. (2024). Dealing with complexity in design science research: A methodology using design echelons. MIS Quarterly, 48(2), 427–458. https://doi.org/10.25300/MISQ/2023/16700
24. Xu, W., Mahmud, R., & Hoo, W. L. (2024). Literature review: Are automated essay scoring systems competent in real-life education scenarios? IEEE Access, 12, 77639–77657. https://doi.org/10.1109/ACCESS.2024.3399163
25. Yan, L., Echeverria, V., Jin, Y., Fernandez-Nieto, G., Zhao, L., Li, X., Alfredo, R., Swiecki, Z., Gašević, D., & Martinez-Maldonado, R. (2024). Evidence-based multimodal learning analytics for feedback and

reflection in collaborative learning. British Journal of Educational Technology, 55(5), 1900–1925. https://doi.org/10.1111/bjet.13496