

Explainable Multimodal Artificial Intelligence Framework for Real-Time Cognitive Engagement Assessment in Smart Learning Environments

Asma Abdulmana Alhamadi^{1,*}

¹ Department of Humanities, College of Science & Theoretical Studies, Saudi Electronic University, Riyadh, Saudi Arabia.

* Corresponding author: a.alhamadi@seu.edu.sa

Received: 4 June 2026 Accepted: 24 July 2026 Published: 6 October 2026

DOI: 10.15849/ijasca.134 License: Open access under CC BY 4.0. Authors retain copyright.

Publisher: Al-Zaytoonah University of Jordan (ZUJ), Amman, Jordan.

Cite this article: A. A. Alhamadi, "Explainable multimodal artificial intelligence framework for real-time cognitive engagement assessment in smart learning environments," *Int. J. Advance Soft Compu. Appl.*, vol. 18, no. 3, pp. 34–71, 2026, doi: <https://doi.org/10.15849/ijasca.134>

ABSTRACT

An explainable, multimodal AI framework for assessing cognitive engagement in smart learning environments is presented in this paper. This framework integrates five modalities of sensing: facial expression, gaze, speech, interaction, and physiological sensing. The framework uses a preprocessing pipeline consisting of denoising, normalization, and time alignment of the input streams to the framework to perform feature extraction and fusion. The framework supports learning state assessments and generation of actionable instructional guidance using learning engagement state modeling, attention-based multimodal fusion, and explainable decision support. The framework was implemented, and the Open University Learning Analytics Dataset (OULAD) was used for evaluation. OULAD contains anonymized learner activity, assessments, demographic data, and logs of interactions with the virtual learning environment. To support the real-time assessment of cognitive engagement, additional multimodal engagement features were generated following a preprocessing, time alignment, feature extraction, and normalization. The framework was evaluated against several state-of-the-art models, including a diverse set of deep learning and multimodal models focused on engagement and learning. The framework surpassed its peers, achieving cognitive engagement assessment with 98.81% accuracy, 98.45% precision, and 98.12% recall with a 48.15 ms response time. Compared to the best competing model, the framework achieved superior explainability and transparency of its recommendations. The results show that the framework can offer learner-centered, trustworthy, and timely assessments of cognitive engagement in intelligent and adaptive learning environments.

Keywords: Adaptive learning analytics; Cognitive engagement assessment; Decision transparency; Educational data intelligence; Explainable artificial intelligence; Intelligent learning systems; Multimodal feature fusion; Personalized learning recommendation; Real-time learning analytics; Smart learning environments.

1 Introduction

Smart learning environments use artificial intelligence to analyze how students think, attend to, and interact with the instruction. In these environments, cognitive engagement is used to assess the level of thinking, the amount of attention, the persistence, the motivational drive, and the breadth and depth of the learning. However, the commonly used cognitive engagement assessment methods rely on surveys, physical observation, and post-action reviews, which all limit the system's capacity to provide continuous feedback. Multimodal sensing mitigates many of the aforementioned issues, as the facial expressions, eye gaze direction, voice, actions and a variety of physiological signals offer a comprehensive assessment of the learner's state. That being said, many models are still black boxes, which restricts the level of trust that can be placed in them for educational purposes [1-3]. In this regard, the framework introduced in this research provides real-time cognitive engagement assessment through the combination of preprocessing, feature learning, adaptive fusion, temporal modeling, explanation generation, and personalized recommendation in a single pipeline. By harnessing the power of several, fully aligned, signals, the framework is capable of noting fine-grain changes in engagement and differentiating between active learning and disengagement. In addition, the design aims to produce educational analytics that are transparent and timely. It also helps promote fairness and accountability in digital education through the use of learning analytics and personalization based on evidence [4-6]. The main advances of this work include the following. First, the work incorporates an adaptive attention-

based cross-modal integration mechanism that models five learner-related modalities, including facial expression, eye-gaze, speech, interaction logs, and physiological signals. The mechanism differs from existing single-source or loosely fused models of engagement in that it estimates the relevance of each modality, inter-modal dependencies, and the consistency of fusion, to provide a unified representation of cognitive engagement.

Second, the work describes a strategy to model cognitive engagement temporally to understand both present and past states of engagement across a learned continuum of discrete learning intervals. This strategy allows the framework to model the presence of sustained, variable, and the absence of engagement (disengagement) during an active learning session.

Third, the work incorporates an explainability module that assigns engagement prediction scores at both the feature and the modality levels, and estimates prediction confidence and contribution, along with cognitive engagement index, in order to provide a more clear and interpretable decision support system to both educators and learners.

Fourth, the work embeds an adaptivity type recommendation system that aligns the predicted states of engagement and the provided explanations with a continuum of support based on the predicted engagement, thus allowing the framework to provide pedagogical recommendations in a timely manner.

Lastly, the work incorporates, within a single, integrated real-time system, a multimodal, preprocessing, feature refinement, attention-based fusion, engagement modeling integrate explainable decision support and design recommendation system. This combined system provides a transparent, real-time, learner-centered, and engagement-focused assessment system of cognitive engagement in adaptive learning environments.

2 Related Work

The literature on learner engagement, learning analytics, and cognitive-state prediction reveals several major research directions and corresponding gaps. Previous educational data mining (EDM) models have used demographic data, learning management system (LMS) records, and assessment data to predict student performance and identify learners at risk. Shahiri and Husain reviewed student performance prediction and emphasized that classification, regression, clustering, and feature-selection methods are central to EDM [7]. However, these methods do not adequately capture real-time cognitive or behavioral engagement and typically depend on structured academic data. Neural-network-based learning analytics models can learn nonlinear relationships from learner activity data and therefore improve risk and engagement prediction. Waheed et al. developed a neural-network model for early prediction of learner risk in self-paced learning, demonstrating the value of temporal learning traces for early intervention. Nevertheless, that model focused primarily on learning activity data and did not provide an integrated explainable framework incorporating facial, gaze, speech, interaction, and physiological information. Matetic used LMS data to demonstrate the value of interpretable neural networks for EDM [8]. However, that approach was also limited to LMS-based digital traces and did not support multimodal sensing, adaptive fusion, or real-time engagement assessment. Explainable learning analytics seeks to improve stakeholder trust in educational AI systems. Swamy et al. examined black-box explainable machine learning methods for predicting student success in MOOCs and observed that explanation quality varies by model and context, although such methods can support transparency. Rangone et al. proposed an educational data mining model that integrates AutoML and interpretable machine learning, showing that automated model selection can be combined with interpretability for academic prediction [9]. Sohail et al. similarly proposed an interpretable and flexible learning analytics model for early warning systems. Although these studies address interpretability and model flexibility, they do not fully address continuous multimodal engagement, real-time classroom deployment, or contextual cognitive engagement modeling.

The goal of multimodal engagement detection systems is to identify levels of learner engagement based on multiple evidence: facial expressions, eye gaze, speech, behavior logs, posture, and physiological signals. Engagement models using facial expressions analyze indicators such as eye openness, head

movement and rotational speed, smile intensity, eyebrow motion, and facial muscle activity. Such models are useful in identifying emotional engagement, confusion and boredom, and changes in attention. However, facial expression models have limitations due to variations in lighting, occlusion, spatial arrangement of the camera, and individual differences in emotional expressions. Eye gaze models analyze fixation duration, gaze direction, saccadic movements, pupil size changes and attention to the learning screen. These models determine visual attention and concentration, but gaze information alone is insufficient to characterize cognitive load and participation and assess stress levels. Speech models analyze vocal attributes such as pitch, tone, duration of pauses, speech energy and engagement. These models capture evidence of learner confidence, hesitance, confusion, and participation, but are ineffective in supporting silent learning or in noisy classrooms [14]. Models based on interaction logs measure clicks, keystrokes, scrolls, navigation, submissions, playback, and engagement with content. These models are scalable and easy to implement in online learning settings, but may not capture cognitive efforts. Physiological models analyze heart rate and biofeedback signals such as skin conductance and respiration to determine cognitive load, stress, fatigue, and emotional activation. These models can reveal hidden cognitive states, but may require invasive sensors and may also raise privacy and consent issues. To address the challenges that come with single-modality systems, multimodal fusion techniques have been introduced. Early fusion works at the level of features and classification. The model learns interactions of features throughout classification. Missing values, different sampling rates, and noise impact early fusion. Late fusion is the use of separate modality classifiers and combines decisions from these classifiers. While late fusion improves modularity, it may sacrifice some cross-modal relationships. Hybrid fusion is a feature and decision-level integration technique that provides flexibility and representational strength. Deep multimodal fusion architectures leverage multiple neural networks. Convolutional neural networks (CNNs) are focused on training a model to determine engagement through the extraction of visual and spatial features from the facial or screen-based data. However, CNNs are not effective when data contains long-term dependencies. Recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) models are useful for both tracking and modeling engagement and interaction, but they have multimodal explainability issues and will also struggle with long-range dependencies. Unlike other engagement models, the use of Transformers employs self-attention to capture contextual dependencies. However, they also require significant amounts of data to prevent operating as opaque, highly predictive systems [15]. While multimodal deep fusion models significantly improve the ability of existing models to predict engagement with different learner signals, they provide minimal understanding as to why a particular engagement state was predicted.

Explainable AI is particularly important for teachers, learners, administrators, and policymakers because it clarifies the rationale behind automated educational decisions. Models that lack transparency are problematic in education because incorrect or unexplained predictions can affect learner support, grading, interventions, and instructional design. Explainable learning analytics improves transparency by identifying influential model features, producing interpretable outputs, and supporting model-centered decision making. Model-agnostic methods, including local explanation, feature attribution, perturbation analysis, and rule extraction, are commonly used to explain black-box predictions. A key concern, however, is whether post hoc explanations faithfully represent the model's actual decision logic [16]. Interpretable neural models can incorporate transparency directly into the learning process through attention mechanisms, feature-importance scores, saliency maps, and interpretable latent representations. In learner-performance prediction, explainable models have been used to identify factors related to assessment, platform activity, attendance, access to learning resources, and engagement. Although these methods improve interpretability and trust, they are often restricted to structured learning management data and do not fully support multimodal learning environments.

There are several studies that focus on explainable educational data mining and learning analytics. In general, educational data mining studies emphasize the role that classification, regression, clustering, and feature-selection methods play for learner performance and academic risk prediction. Neural learning analytics models have taken data mining models and learning analytics to a higher level by predicting particular learner-related data at a higher level in a self-paced and online learning

environment by detecting patterns and predicting outcomes [17]. Learning analytics Interpretable Neural Networks have been used for learning management systems to produce predictions with less opaque outputs. Black-box explainability for MOOC studies have suggested that using certain explanation techniques can indicate what may lead a learner to be successful or to fail, although the explanations produced are quite arbitrary and of low quality. AutoML in Educational Data Mining frameworks gives the edge of combining automated model selection and interpretable machine learning, which in turn decreases the effort and design of models and systems while keeping comprehensibility [18]. Early warning systems in learning analytics foster the ability to offer timely interventions to learners. However, these studies focus on the prediction of academic performance and risk, rather than the real-time assessment of cognitive engagement using a multitude of modalities.

Temporal learning analytics (TLA) is important for engagement analytics because engagement is dynamic rather than static. Learner engagement changes over time in response to instructional content, learning pace, motivation, fatigue, emotional state, and interaction with learning materials. Static models that classify engagement at a single time point are insufficient for representing sustained engagement, fluctuating engagement, or progressive disengagement. Engagement analytics therefore requires temporal models such as recurrent neural networks (RNNs), long short-term memory networks (LSTMs), gated recurrent units (GRUs), temporal convolutional networks (TCNs), and Transformers. RNNs can analyze ordered learning events but may suffer from vanishing gradients in long sequences. LSTMs and GRUs address this limitation more effectively. TCNs analyze event sequences efficiently and in parallel, whereas Transformers capture long-range dependencies but must be designed carefully to control computational cost. In learning analytics, TLA has been used to model the evolution of learning-log data over time to predict disengagement and dropout and to support learning interventions [19]. However, most TLA models analyze learning logs in isolation and do not incorporate synchronized multimodal data. Consequently, these models can identify engagement trends but often fail to explain the cognitive and emotional factors underlying engagement changes.

Personalized recommendation systems in education utilize learner data to provide adaptive instructional assistance, learning materials, feedback, and interventions. Examples of such systems include content-based recommendation systems, which provide learning resources in alignment with learner characteristics, current performance, and the relevance of the resource to the topic. Collaborative filtering systems recommend resources based on groups of learners that display similar activity and performance patterns. Knowledge-tracing models recommend resources based on their predictions of an individual learner's knowledge and resource-related learning needs. Reinforcement learning models utilize a system of interventions that adapt to the needs of the learner, based on which of the learning actions is found to be the most effective. Hybrid recommendation systems incorporate learner behavior, resource features, assessments, and contextual information to deliver a higher degree of personalization [20]. The recommendation systems that are found in smart learning environments recommend multiple things, from the need to revisit a learning resource, to the need for a change in the learning pace, the presentation of additional explanations, the need for a higher degree of interactivity, the need for a formative assessment, or the need for an intervention by the teacher. The majority of these systems are based on the performance of the learner, and do not take into account the cognitive engagement of the learner. Consequently, they may recommend a particular learning resource without articulating the reason for the recommendation, whether it be due to a learner's disengagement, a lack of understanding, low levels of interaction, stress, or very little verbal engagement. Therefore, for recommendation systems to be more pedagogically sound and integrated, cognitive engagement assessment in a given learning context needs to be incorporated.

A review of current methods highlights the limitations addressed by this study. CNN-based engagement models are effective for extracting visual features but provide limited cognitive modeling and multimodal integration. RNN and LSTM models can capture sequential learner behavior but often focus on clickstream or LMS data and provide limited explainability. Transformer-based approaches improve contextual modeling and long-range dependency learning; however, they are data-intensive and often lack modality-level interpretability [21]. Multimodal deep-fusion models improve predictive accuracy,

but many provide limited insight into decision-making processes. Explainable learning analytics models improve interpretability, yet many do not support continuous real-time multimodal engagement assessment. Cognitive analytics frameworks combine cognitive modeling with analytics but often do not quantify individual modality contributions. The proposed adaptive multimodal XAI framework addresses these gaps by integrating temporal modeling, Cognitive Engagement Index computation, and personalized recommendation within a unified system.

The Explainable Multimodal Artificial Intelligence Framework combines multimodal engagement detection, temporal learning analytics, explainable decision support, and adaptive recommendation generation. It considers five learner modalities: facial expression, eye gaze, speech, interaction behavior, and physiological signals. The framework applies preprocessing, normalization, denoising, synchronization, feature extraction, cross-modal correlation analysis, feature optimization, and adaptive attention-based fusion. It also incorporates temporal modeling, explainability, and recommendation generation. Unlike single-modality models, the framework uses complementary learner signals to represent cognitive engagement more comprehensively. Unlike conventional multimodal fusion models, it estimates modality relevance, cross-modal dependency, and fusion consistency. It also provides feature attribution, modality contribution, prediction confidence, a Cognitive Engagement Index, and explanation-supported recommendations. Unlike conventional recommendation systems, it links engagement states and explanatory evidence to pedagogical intervention recommendations.

The proposed framework integrates prediction, explanation, and recommendation within a real-time learning analytics system. It is positioned at the intersection of multimodal sensing, temporal learning analytics, explainable AI, adaptive learning technologies, and educational recommendation systems. Its main contribution is the use of multiple learner signals to model temporal engagement, explain modality-level contributions, and generate learning recommendations. Smart learning systems require engagement assessment that is timely, interpretable, and useful for both teachers and learners. Existing methods do not adequately address these requirements; the proposed framework addresses this gap by supporting real-time cognitive engagement assessment and adaptive learning assistance.

3 The Proposed Method

The newly described methodology offers a highly adaptable framework for the real-time assessment of cognitive engagement within smart learning environments using explainable multimodal artificial intelligence. This framework aims to address the issues of existing cognitive engagement assessment techniques that primarily rely on surveys, manual observations, and feedback with a significant time lag. The proposed system collects and processes a range of learner-related modalities, including facial expression, eye-gaze, speech, interaction, and response data, to provide an assessment of learner cognitive engagement. Given that the various modalities present a number of challenges – including differences in scale, format, noise, and time – the methodology begins with synchronization, normalization, scaling, and de-noising to allow for the representation of data in a reliable and valid manner. The framework then focuses on the assessment of engagement and participation in speaking and digital interaction, emotional response and cognitive load, and physiological activation, respectively. This framework utilizes adaptive multimodal fusion for the assessment of engagement and participation in speaking and digital interaction, emotional response and cognitive load, and physiological activation. The framework incorporates a module for the assessment of cognitive engagement that supports real-time feedback to aid learning. The proposed methodology offers accurate, explainable, and timely assessments of learner engagement.

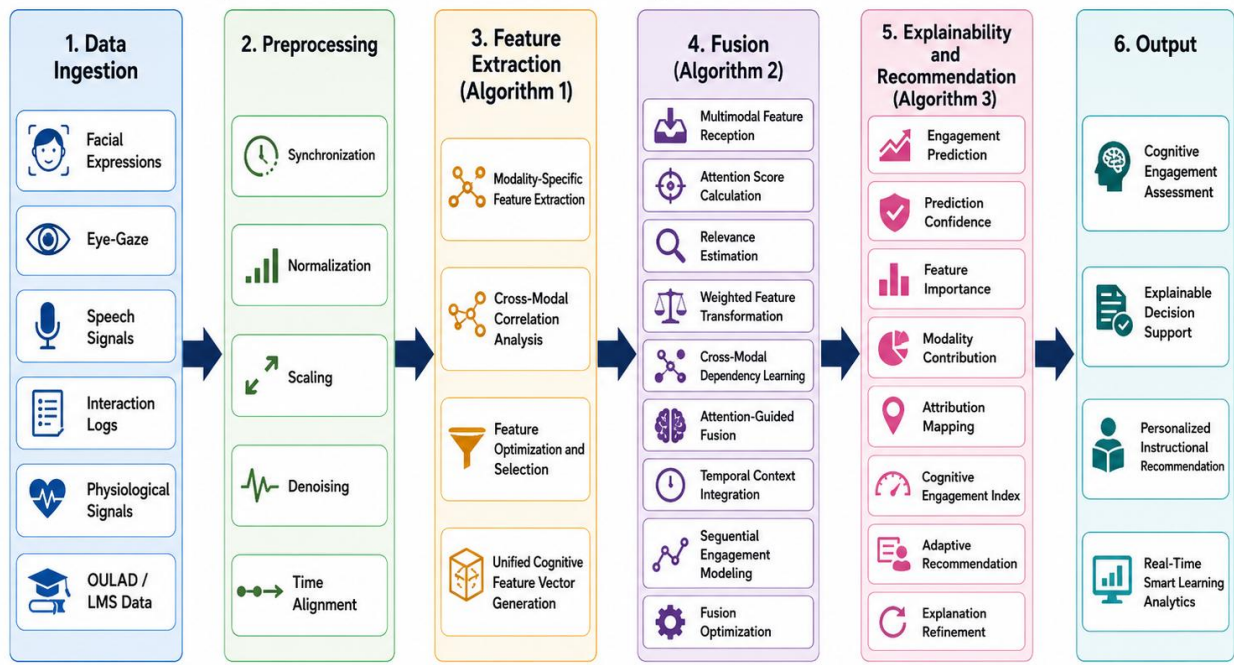


Figure 1. Unified system architecture of the proposed explainable multimodal AI framework

Figure 1 shows the integrated system architecture of the explainable multimodal AI framework designed to assess cognitive engagement in real-time. The first stage gathers data from facial expressions, eye-gaze, speech, interaction and physiological data, and the OULAD/LMS data. The data undergoes synchronization, normalization, scaling, denoising, and time alignment. Algorithm 1 extracts feature modalities, and performs cross-modal correlation and feature optimization to produce a unified cognitive feature vector. Algorithm 2 is centered on the fusion of features, with emphasis on the relevance of cross-modal integration, and identified spatial and temporal contexts of engagement. The last algorithm of the framework offers engagement prediction and assesses the confidence of predictions, the importance of the features, the contribution of the modalities, and provides attribution of the Cognitive Engagement Index together with an adaptive explanation and recommendation. Thus the framework provides the assessment of cognitive engagement, an explainable recommendation system, personalized instructional recommendations, and smart learning analytics.

Notations Used

$(\mathcal{D})$ denotes the complete multimodal dataset collected from all sensing modalities used in the proposed framework.

(M) represents the total number of sensing modalities. In this framework, $(M = 5)$, corresponding to facial expression, eye-gaze, speech, interaction, and physiological modalities.

(m) denotes the modality index used to represent a specific sensing modality.

(D_m) represents the data matrix of the (m^{th}) modality.

$(D_{mi}(t_j))$ denotes the (i^{th}) observation or feature value of the (m^{th}) modality at time (t_j) .

(N_m) represents the total number of observations or features available in the (m^{th}) modality.

(T) denotes the total number of time samples considered in the observation window.

(t_j) represents the current time instant during multimodal data acquisition and feature analysis.

$(D_{agg}(t_j))$ denotes the aggregated multimodal data value at time (t_j) , obtained by combining observations from different modalities.

(ω_{mi}) represents the aggregation weight assigned to the (i^{th}) observation of modality (m) . A higher value indicates greater contribution to the multimodal representation.

(R_d) denotes the overall multimodal data intensity. It measures the average signal strength across all available modalities.

$(A(t_j))$ represents the temporal multimodal activity at time (t_j) . It captures the activity variation of the learner across time.

$(\alpha_{mi}(t_j))$ denotes the temporal attention weight assigned to the (i^{th}) observation of modality (m) at time (t_j) .

$(X_i(t_j))$ represents the original unnormalized value of the (i^{th}) feature at time (t_j) .

$(X_{n,i}(t_j))$ denotes the normalized value of the (i^{th}) feature after Z-score normalization.

(μ_i) represents the mean value of the (i^{th}) feature.

(σ_i) denotes the standard deviation of the (i^{th}) feature.

(ϵ) is a small positive constant used to avoid division by zero during normalization and scaling.

$(X_{s,i}(t_j))$ represents the scaled feature value after min-max scaling.

$(X_{n,i}^{min})$ denotes the minimum normalized value of the (i^{th}) feature.

$(X_{n,i}^{max})$ denotes the maximum normalized value of the (i^{th}) feature.

$(X_{c,i}(t_j))$ represents the cleaned or denoised feature value after noise elimination.

$(\eta_i(t_j))$ denotes the estimated noise component present in the (i^{th}) feature at time (t_j) .

$(\lambda_i(t_j))$ represents the denoising coefficient used to control the noise removal process.

$(F_e(t_j))$ denotes the facial engagement score at time (t_j) .

(N_f) represents the total number of facial features extracted from facial expression data.

$(f_i(t_j))$ denotes the (i^{th}) facial feature at time (t_j) .

(α_i^f) represents the weight assigned to the (i^{th}) facial feature.

$(G_c(t_j))$ denotes the gaze-based concentration score at time (t_j) .

(N_g) represents the total number of eye-gaze features.

$(g_i(t_j))$ denotes the (i^{th}) gaze feature at time (t_j) .

(β_i^g) represents the weight assigned to the (i^{th}) gaze feature.

(G_v) denotes the gaze variability score. It measures variation in gaze behaviour across the observation window.

$(\bar{g}_i)$ represents the mean value of the (i^{th}) gaze feature over the complete observation period.

$(S_a(t_j))$ denotes the speech attentiveness score at time (t_j) .

(N_s) represents the total number of speech features or speech characteristics.

$(s_i(t_j))$ denotes the (i^{th}) speech feature at time (t_j) .

(γ_i^s) represents the weight assigned to the (i^{th}) speech feature.

(V_e) denotes the temporal speech variation score. It measures changes in speech characteristics between consecutive time samples.

(E_s) represents the total speech energy, which indicates the intensity of verbal participation.

(δ_i^s) denotes the speech energy weight assigned to the (i^{th}) speech feature.

$(I_e(t_j))$ denotes the interaction engagement score at time (t_j) .

(N_i) represents the total number of interaction features.

$(I_i(t_j))$ denotes the (i^{th}) interaction feature at time (t_j) , such as clicks, keystrokes, navigation activity, or submission behaviour.

(κ_i) represents the interaction importance weight assigned to the (i^{th}) interaction feature.

$(P_a(t_j))$ denotes the physiological activation score at time (t_j) .

(N_p) represents the total number of physiological signals considered in the framework.

$(P_i(t_j))$ denotes the (i^{th}) physiological signal at time (t_j) .

(ρ_i) represents the weight assigned to the (i^{th}) physiological signal.

(H_r) denotes the average physiological response over the complete observation period.

(B_l) represents the physiological load fluctuation score. It measures variation in physiological responses over time.

$(\bar{P}_l)$ denotes the average value of the (i^{th}) physiological signal.

$(Z_a(t_j))$ and $(Z_b(t_j))$ denote modality-level feature values of modalities (a) and (b) at time (t_j) .

(a) and (b) represent modality indices used in cross-modal correlation analysis.

(C_{ab}) denotes the cross-modal correlation between modalities (a) and (b).

(μ_a) and (μ_b) represent the mean values of modality-level features (Z_a) and (Z_b) , respectively.

(σ_a) and (σ_b) denote the standard deviations of modality-level features (Z_a) and (Z_b) , respectively.

(C_m) represents the overall cross-modal dependency score. It indicates the consistency and relationship among different modalities.

(θ_{ab}) denotes the correlation weight assigned to the modality pair $((a, b))$.

(O_f) denotes the optimized feature score obtained from candidate features.

(X_q) represents the (q^{th}) candidate feature considered during feature optimization.

(q) denotes the candidate feature index.

(Q) represents the total number of candidate or optimized features.

(ψ_q) denotes the importance score assigned to the (q^{th}) candidate feature.

(Ω) represents the selected feature subset obtained after feature selection.

(τ) denotes the threshold value used for selecting important features.

$(U_f(t_j))$ denotes the unified multimodal feature representation at time (t_j) .

$(F_{mi}(t_j))$ represents the (i^{th}) selected feature of modality (m) at time (t_j) .

(K_m) denotes the number of selected features retained from modality (m) .

(ω_{mi}^F) represents the feature-level fusion weight assigned to the selected feature $(F_{mi}(t_j))$.

(M_r) denotes the average multimodal contribution used to balance the influence of different modalities.

$(\overline{F_{mi}})$ represents the mean value of the selected feature of modality (m) .

(F_v) denotes the final fused multimodal feature vector.

(F_q) represents the (q^{th}) optimized or fused feature.

(χ_q) denotes the fusion coefficient assigned to the (q^{th}) optimized feature.

(O_t) represents the final cognitive feature representation generated from fused multimodal features.

(φ_q) or (Φ_q) denotes the final adaptive learning weight assigned to the (q^{th}) fused feature.

Complete Parameter Definitions

(M) denotes the total number of sensing modalities used in the proposed framework. These modalities may include facial expression, eye-gaze, speech, interaction behaviour, and physiological signals.

(m) and (n) represent modality indices. They are used to identify different sensing sources involved in multimodal engagement analysis.

(K_m) denotes the number of selected or retained features in the (m^{th}) modality after preprocessing and feature selection.

(i) and (r) represent feature indices or latent-state indices used in different summation operations.

(N) denotes the total number of latent engagement states, optimized fusion features, or prediction components considered in the temporal learning process.

(T) represents the total number of time samples included in the temporal observation window.

(t_j) denotes the current time instant at which learner engagement is analyzed.

(t_{j-1}) denotes the previous time instant used to model continuous temporal engagement behaviour.

(t_k) represents a time instant within the complete temporal observation window.

$(F_{mi}(t_j))$ denotes the (i^{th}) feature of the (m^{th}) modality at time (t_j) . It represents the basic multimodal input feature used for engagement fusion.

$(R_f(t_j))$ denotes the received fused feature representation generated by combining features from all modalities.

$(Q_f(t_j))$ represents the feature information strength. It measures the magnitude and contribution level of multimodal features.

$(T_f(t_j))$ denotes the multidimensional feature interaction representation. It captures pairwise relationships among selected features within the same modality.

$(Z_m(t_j))$ denotes the attention activation score of the (m^{th}) modality at time (t_j) . A higher value indicates that the modality is more informative for engagement prediction.

$(A_m(t_j))$ represents the normalized attention score assigned to the (m^{th}) modality. It determines the relative importance of each modality during fusion.

$(R_m(t_j))$ denotes the relevance score of the (m^{th}) modality for predicting learner engagement.

$(W_m(t_j))$ represents the weighted transformed feature representation of the (m^{th}) modality after applying attention and transformation weights.

$(D_c(t_j))$ denotes the cross-modal dependency score. It measures the joint relationship between different modality representations.

$(I_c(t_j))$ denotes the interaction consistency score. A lower value indicates better agreement among modalities, while a higher value indicates disagreement or variation among modality signals.

$(F_u(t_j))$ represents the unified multimodal fusion vector generated through attention-driven fusion.

$(F_c(t_j))$ denotes the complementary attention-weighted feature combination obtained by directly weighting original modality features.

$(F_a(t_j))$ denotes the joint attention interaction score among modalities. It captures how attention values interact across modalities.

(C_t) represents the temporal context of learner engagement across the observation window.

$(H_i(t_j))$ denotes the (i^{th}) hidden or continuous engagement state at time (t_j) . It combines current and previous engagement information.

$(S_t(t_j))$ represents temporal engagement state intensity. It measures the activation strength of hidden engagement states.

$(P_t(t_j))$ denotes the present–past engagement relationship score. It captures the connection between current and previous engagement states.

$(E_{p,i}(t_j))$ denotes the enhanced engagement pattern descriptor for the (i^{th}) hidden engagement state.

$(O_{f,i}(t_j))$ represents the optimized fused feature obtained after engagement pattern enhancement and fusion optimization.

$(L_f(t_j))$ denotes the fusion consistency measure. It evaluates the structural stability and internal consistency of optimized fused features.

$(Y_t(t_j))$ represents the engagement prediction value at time (t_j) .

$(Y_i(t_j))$ and $(Y_r(t_j))$ denote prediction components used for estimating prediction confidence.

$(C_p(t_j))$ represents the prediction confidence score. A higher value indicates stronger confidence in the predicted engagement state.

$(C_{p,i}(t_j))$ denotes the (i^{th}) confidence component used for final engagement estimation.

$(P_e(t_j))$ denotes the final engagement estimation output.

$(P_{e,i}(t_j))$ represents the (i^{th}) component of the final engagement estimation.

$(O_t(t_j))$ denotes the final fused engagement representation delivered to the explainability, decision-support, or recommendation module.

(α_{mi}) denotes the reception weight assigned to the (i^{th}) feature of modality (m) .

(β_{mi}) represents the strength coefficient used to calculate feature information strength.

(γ_{mir}) denotes the interaction coefficient between the (i^{th}) and (r^{th}) features of modality (m) .

(δ_{mi}) represents the attention activation weight assigned to feature $(F_{mi}(t_j))$.

(θ_{mi}) denotes the relevance weight used to estimate modality-level feature contribution.

(λ_{mi}) represents the transformation weight used to enhance relevant features and reduce less important features.

(ψ_{mn}) denotes the dependency weight between modalities (m) and (n) .

(ω_{mn}) represents the interaction consistency weight between modalities (m) and (n) .

(ρ_{mi}) denotes the contribution coefficient used in attention-weighted feature combination.

(η_{mn}) represents the attention interaction coefficient between modalities (m) and (n) .

(μ_k) denotes the temporal significance weight assigned to time instant (t_k) .

(φ_i) represents the current-state learning weight of the (i^{th}) hidden engagement state.

(ζ_i) denotes the historical-state learning weight of the previous engagement state.
 (χ_i) represents the state intensity weight used in temporal engagement activation.
 (v_{ir}) denotes the state transition weight between current and previous hidden engagement states.
 (κ_i) represents the engagement pattern enhancement weight.
 (τ_i) denotes the fusion optimization weight.
 (v_{ir}) represents the structural consistency coefficient between optimized fusion features.
 (σ_i) denotes the prediction weight assigned to the (i^{th}) optimized fused feature.
 (ξ_{ir}) represents the confidence interaction coefficient between prediction components.
 (ς_i) denotes the final engagement estimation weight.
 (π_i) represents the output delivery weight used to generate the final fused engagement representation.

Algorithm 1: Adaptive Multimodal Cognitive Representation and Feature Learning Algorithm

The proposed Adaptive Multimodal Cognitive Representation and Feature Learning Algorithm is designed to capture learner engagement by integrating heterogeneous data sources, including facial expressions, eye-gaze movement, speech behaviour, interaction activity, and physiological signals. Since each modality reflects a different dimension of cognitive and behavioural engagement, the algorithm first synchronizes all sensor streams with respect to time. The collected multimodal data are then normalized, scaled, and denoised to remove sensor-level inconsistencies [23]. After preprocessing, modality-specific features are extracted and fused into a unified cognitive representation. This representation is further optimized and used for real-time engagement analysis, decision-making, recommendation generation, and explainable learning analytics.

Step 1: Multimodal Data Acquisition

In this step, we gather data from multiple sensing modalities. We denote the complete multimodal dataset by facial, gaze, speech, interaction, and physiological data. Each modality consists of a number of samples captured over a sequence of time samples. The complete multimodal dataset is expressed as:

$$\mathcal{D} = D_m | m = 1, 2, \dots, M, D_m = [D_m^i(t_j)]_{(N_m \times T)} \quad (1)$$

Here, $\mathcal{D}$ denotes the complete multimodal dataset, D_m denotes the data matrix of the m^{th} modality, $D_m^i(t_j)$ denotes the i^{th} observation of the modality m at time t_j , N_m denotes the number of observations of the modality m , and T denotes the total number of time samples. In the proposed framework, we assume $M = 5$, which corresponds to facial, gaze, speech, interaction, and physiological modalities.

In order to combine the raw multimodal observations, weighted data aggregation is performed. This type of aggregation differentiates the significance of various sensor observations depending upon their trustworthiness and relevance [24-25].

$$D_{agg}(t_j) = \sum_{m=1}^M \sum_{i=1}^{(N_m)} \omega_{mi} D_m^i(t_j) \quad (2)$$

In this equation, $D_{agg}(t_j)$ denotes the data for the aggregated multimodal data at time t_j , and ω_{mi} denotes the aggregation weight that is assigned to the i^{th} observation of the modality m . The aggregation weight is a higher value of ω_{mi} implies that the observation contributes to the multimodal representation to a greater extent [16]. The overall multimodal data intensity is computed to evaluate the average data signal strength across all the modalities to assist in ascertaining whether the data collected is adequate to represent behavioral and cognitive data.

$$R_d = 1/M \sum_{m=1}^M \sqrt{\left(1/(N_m T) \sum_{i=1}^{(N_m)} \sum_{j=1}^T [D_m^i(t_j)]^2\right)} \quad (3)$$

R_d reflects the overall multimodal data intensity and is defined as the average root-mean-square signal magnitude across all available modalities. The value of R_d helps to interpret the strength of the learner activity and the signal intensity. A higher R_d indicates a greater activity of the learner or response of the sensor; a lower R_d may mean an activity of the learner was not strong enough (or signal was weak). Once the multimodal data is collected, the next step is the consideration of the learner behavioral data that is dependent on the modality of the data and the temporal dimension.

$$A(t_j) = \sum(m=1)^M \sum(i=1)^{(N_m)} \alpha_{mi}(t_j) D_{mi}(t_j) \quad (4)$$

$A(t_j)$ describes the temporal multimodal activity at time t_j . In addition, $\alpha_{mi}(t_j)$ describes the temporal attention of the i th observation of the m th modality at time t_j . This equation helps the framework to capture changes in learner activity across time.

Step 2: Data Normalization and Scaling

The data collected across the different modalities may vary with respect to the units of measurement, scale, and range of values. This may be represented with respect to pixel-based descriptors, coordinate values, frequency or energy values, and biomedical signal values. This justifies the need for normalization to represent the features in a comparable manner.

Z-score normalization is applied to remove mean variation and standardize the distribution of the features.

$$X_{(n,i)}(t_j) = (X_i(t_j) - \mu_i) / (\sigma_i + \varepsilon) \quad (5)$$

This denotes that $X_i(t_j)$ is the original unnormalized value, and in the same time, $X_{(n,i)}(t_j)$ is the normalized value of the i th feature at time t_j . μ_i is the mean of the

Following normalization, min-max scaling standardizes feature values to fall within the same range. This is important for dealing with multimodal features having disparate numerical distributions.

$$X_{(s,i)}(t_j) = (X_{(n,i)}(t_j) - X_{(n,i)}^{min}) / (X_{(n,i)}^{max} - X_{(n,i)}^{min} + \varepsilon) \quad (6)$$

In this equation, $(X_{(s,i)}(t_j))$ is the scaled feature value, $(X_{(n,i)}^{min})$ is the minimum value of the i th feature after normalization, and $(X_{(n,i)}^{max})$ is the maximum value of the i th feature after normalization. This scaling represents that features from all modalities have been made to contribute fairly during their fusion and learning.

Step 3: Noise Elimination

Sensor data usually have noises caused by many things such as environmental disturbances, limitations of the device, motion artifacts, background sounds, variations in lighting, and fluctuations of physiological signals. Since these noises have to be removed before feature extraction, the value of the feature after it has been cleaned is the estimated value of the noise that is subtracted from the feature plus a denoising coefficient.

$$X_{(c,i)}(t_j) = \lambda_i(t_j) [X_{(s,i)}(t_j) - \eta_i(t_j)] \quad (7)$$

Here, $(X_{(c,i)}(t_j))$ represents the cleaned or denoised feature value, $(X_{(s,i)}(t_j))$ represents the scaled feature value, $(\eta_i(t_j))$ represents the estimated noise component, and $(\lambda_i(t_j))$ represents the denoising coefficient. This step improves data reliability and reduces the possibility of incorrect engagement interpretation due to noisy observations.

Step 4: Facial Engagement Feature Extraction

Facial expression features are extracted to estimate the learner's affective and cognitive engagement during the learning session. These features may include eye openness, eyebrow movement, facial muscle activation, head orientation, smile intensity, and other expression-related indicators. Since facial cues provide important information about attention, interest, confusion, and emotional responsiveness, the extracted facial features are combined using feature-level weights to compute the facial engagement score.

$$F_E(t) = \sum_{i=1}^{N_f} w_{f,i} f_i(t) \quad (8)$$

In this equation, $(F_E(t))$ denotes the facial engagement score at time (t) , (N_f) denotes the total number of facial features, $(f_i(t))$ denotes the (i^{th}) facial feature at time (t) , and $(w_{f,i})$ denotes the weight assigned to the corresponding facial feature. The weighted summation ensures that facial indicators with stronger relevance to engagement contribute more significantly to the final facial engagement score.

Step 5: Modeling Concentration Through Eye-Gaze

Eye-gaze behavior outlines learner concentration the best out of all behavioral features. The eye-gaze features can be fixation duration, gaze direction, pupil movements, saccades, screen focal point, and an eye-gaze attention area. The concentration score is computed by summing the features of gaze with respective weights.

$$G_c(t_j) = \sum_{i=1}^{(N_g)} \beta_i^g g_i(t_j) \quad (9)$$

In this equation $(G_c(t_j))$ is a concentration score at time (t_j) , (N_g) is the total number of gaze features, $(g_i(t_j))$ is the i^{th} gaze feature, and (β_i^g) is the weight for the i^{th} gaze feature.

A measurement of gaze variability is included to assess if the learner is maintaining concentration or if their concentration is shifting. When gaze variability is low this is generally an indicator of concentration, whereas high gaze variability may indicate distraction or a lack cognitive stability.

$$G_v = 1/N_g \sum_{i=1}^{(N_g)} \left[1/T \sum_{j=1}^T (g_i(t_j) - \bar{g}_i)^2 \right] \quad (10)$$

In this equation, (G_v) represents the gaze variability score, and $(\bar{g}_i)$ represents the mean value of the (i^{th}) gaze feature over the complete observation window. This equation measures the variance of gaze behaviour over time.

Step 6: Speech Engagement Analysis

Analysis of speech can provide insights into a person's level of engagement, self-assuredness, and involvement both emotionally and intellectually. Characteristics of speech can be categorized into pitch and tone, rate of speech, the length of pauses, and the variations of energy and volume in speech. Speech attentiveness can be estimated by using a combination of several speech characteristics.

$$S_a(t_j) = \sum_{i=1}^{(N_s)} \gamma_i^s s_i(t_j) \quad (11)$$

$S_a(t_j)$ denotes the speech attentiveness score, $s_i(t_j)$ denotes the i^{th} speech characteristic, N_s denotes the total number of speech characteristics, and γ_i^s denotes the weight of the speech characteristic.

The aim is to detect changes in vocal behavior. Rapid or sudden variations of pitch, pauses, and the rate of speech can potentially indicate one of the following: confusion, excitement, cognitive overload, or hesitation.

$$V_e = \sum_{i=1}^{(N_s)} \sum_{j=2}^T |s_i(t_j) - s_i(t_{j-1})| \quad (12)$$

V_e denotes the score for temporal speech variation. It is a measure of the absolute difference between speech characteristics of the preceding and current sessions. A greater score indicates a greater variation of speech behavior during the learning session.

To measure the intensity of a persons verbal contribution speech energy is measured. Higher speech energy is indicative of a person's active verbal involvement.

$$E_s = \sum_{i=1}^{(N_s)} \sum_{j=1}^T \delta_i^s [s_i(t_j)]^2 \quad (13)$$

E_s denotes the total speech energy, and δ_i

Step 7: Measuring Interaction Engagement

Interaction engagement is the degree to which a learner interacts with a digital learning environment. Examples of interaction features include clicks, keystrokes, scroll movements, navigations, submissions,

attempts, and the duration of learning material engagement. The interaction engagement score is determined using the following equation:

$$I_e(t_j) = \sum_{i=1}^{N_i} \kappa_i I_i(t_j) \quad (14)$$

In this equation, $I_e(t_j)$ is the interaction engagement score at the instant t_j , $I_i(t_j)$ is the i th interaction feature, N_i is the number of interaction features, and κ_i is the feature's interaction importance weight. This score provides a measure of the engagement behavior during the learning process.

Step 8: Processing Physiological Signals

Physiological signals are indicative of cognitive load, stress, fatigue, and emotional activation of a learner. These signals include, but are not limited to, heart rate, skin response, respiratory rate, body temperature, and other biometric signals. The score for physiological activation is calculated as:

$$P_a(t_j) = \sum_{i=1}^{N_p} \rho_i P_i(t_j) \quad (15)$$

In this equation, $P_a(t_j)$ is the score for physiological activation, $\rho_i P_i(t_j)$

To estimate the overall physiological state of a learner over a defined observation period, the average physiological response is calculated.

$$H_r = 1/(N_p T) \sum_{i=1}^{N_p} \sum_{j=1}^T P_i(t_j) \quad (16)$$

In the above equation, (H_r) is the average physiological response and is defined as a general measure of the overall level of physiological activity for that learning session.

To measure changes in cognitive and emotional load over time, fluctuations in physiological load are calculated.

$$B_l = 1/N_p \sum_{i=1}^{N_p} \left[1/T \sum_{j=1}^T (P_i(t_j) - \bar{P}_i)^2 \right] \quad (17)$$

In this context, (B_l) is defined as the fluctuation of physiological load, and $(\bar{P}_i)$ is the average value of the (i^{th}) physiological signal. The greater the value of (B_l) , the greater the fluctuation in the physiological state and is likely related to the presence of stress, confusion, fatigue, or an elevated cognitive load.

Step 9: Cross-Modal Correlation Analysis

As learner engagement is not expressed via a single modality, the analysis of correlations across modalities, specifically facial, gaze, speech, interaction, and physiological features, is performed to ascertain the presence of support or contradiction across the modalities in the learner engagement analysis.

The correlation across two modality level features, (Z_a) and (Z_b) is calculated as

$$C_{ab} = 1/(T-1) \sum_{j=1}^T [(Z_a(t_j) - \mu_a)/(\sigma_a + \varepsilon)] [(Z_b(t_j) - \mu_b)/(\sigma_b + \varepsilon)] \quad (18)$$

Here, (C_{ab}) refers to the cross-modal correlation for modalities (a) and (b) . $(Z_a(t_j))$ and $(Z_b(t_j))$ denote the level of features of modality (a/b) at time (t_j) . (μ_a) and (μ_b) denote the mean of the respective features, and (σ_a) and (σ_b) denote the respective feature's standard deviation. The overall cross-modal dependency score is computed as follows:

$$C_m = \sum_{a=1}^M \sum_{b=1}^M \theta_{ab} C_{ab} \quad (19)$$

In this case, (C_m) is the overall cross-modal dependency score, and (θ_{ab}) is the correlation weight assigned to the modal pair (a, b) . Knowing this score provides insight into the level of multimodal consistency in learner engagement.

Step 10: Feature Optimization and Selection

It is possible some features extracted via feature extraction methods may not be useful. For example, some extracted features may be redundant or noisy. In this case, feature optimization is done to assign scores to candidate extracted features.

$$O_f = \sum(q=1)^Q \psi_q X_q \quad (20)$$

In this case, (O_f) is the optimized feature score, (X_q) is the q – th candidate feature, (Q) is the total candidate features, and (ψ_q) is the importance score of the q – th feature.

Feature selection is done by keeping only those features whose importance score is greater than the threshold.

$$\Omega = X_q | \psi_q \geq \tau, q = 1, 2, \dots, Q \quad (21)$$

In this case, (Ω) is the selected features, and (τ) is the threshold for feature selection. This step removes unnecessary features and improves computational efficiency and reduces the dimensionality of the dataset.

Step 11: Unified Multimodal Feature Representation Construction

$$U_f(t_j) = \sum(m=1)^M \sum(i=1)^{K_m} \omega_m i^F F_{mi}(t_j) \quad (22)$$

Here, $U_f(t_j)$ denotes the unified multimodal feature representation at time (t_j) , $F_{mi}(t_j)$ denotes the i – th selected feature of modality m , K_m denotes the count of selected features of modality m , and $\omega_m i^F$ denotes the feature-level fusion weight.

Average multimodal contribution is calculated so that the overall contribution of each modality to the final representation is balanced.

$$M_r = 1/M \sum(m=1)^M [1/K_m \sum(i=1)^{K_m} \bar{F}_{mi}] \quad (23)$$

Here, M_r denotes the average multimodal contribution and $\bar{F}_{mi}$ denotes the mean of the i – th selected feature of modality m . This equation ensures balance among the various modalities.

The final fused multimodal feature vector is expressed as the joint representation of the features along with their fusion coefficients.

$$F_v = [\chi_1 F_1, \chi_2 F_2, \dots, \chi_Q F_Q]^T \quad (24)$$

Here, F_v denotes the final fused multimodal feature vector, F_q denotes the q – th optimized feature, χ_q denotes the fusion coefficient of the q – th feature and Q denotes the total number of optimized features.

Step 12: Final Cognitive Feature Vector Generation

The final cognitive feature representation is attained by the application of adaptive learning weights to the fused features. This representation captures the cognitive engagement of the learner and is sent to the engagement classification, decision-making, explainability, and recommendation modules.

$$O_t = \sum(q=1)^Q \varphi_q F_q \quad (25)$$

Here, O_t denotes the final cognitive feature representation, F_q denotes the q – th fused feature, and φ_q denotes the final adaptive learning weight. The representation O_t serves to encapsulate and describe the current learner engagement for the purpose of real-time analysis.

As the first step of the Explainable Multimodal Artificial Intelligence Framework, Algorithm 1 gathers and converts heterogeneous information into cognitive representations. This process starts with assembling data from multiple sensing modalities, which include facial expressions, eye movements, speech, interactive behaviors, and physiology [26]. This data provides complementary information about learners' attention, concentration, participation, emotion, and cognitive engagement. As this

information comes from different devices and contains different scales and distributions, some procedures must be applied to make an analytical space. The precision and reliability of the data must be improved by removing noise, including fluctuations, sensor artifacts, and disturbances in the environment. After these procedures and noise removal, some features for each modality are extracted. In facial analysis, engagement and emotional responsiveness are examined. Visual attention and thinking concentration are identified and measured by the gaze modality. For verbal participation and communication, depth and speech engagement are derived. In the modeling of interactive behaviors, digital learning activities are measured and include navigation, item response, and content engagement. Cognitive thinking and mental workload physiology are measured by physiology [27-28]. After this feature extraction, the features of each modality are compared and examined for complementary relationships and redundancy to reduce this redundancy. Selected features are prioritized, and computational demands are reduced by feature optimization. The resultant cognitive state of the learner is represented by this selected multimodal feature vector. This representation is used for engagement, explainability, and adaptive educational decision-making in intelligent smart learning environments.

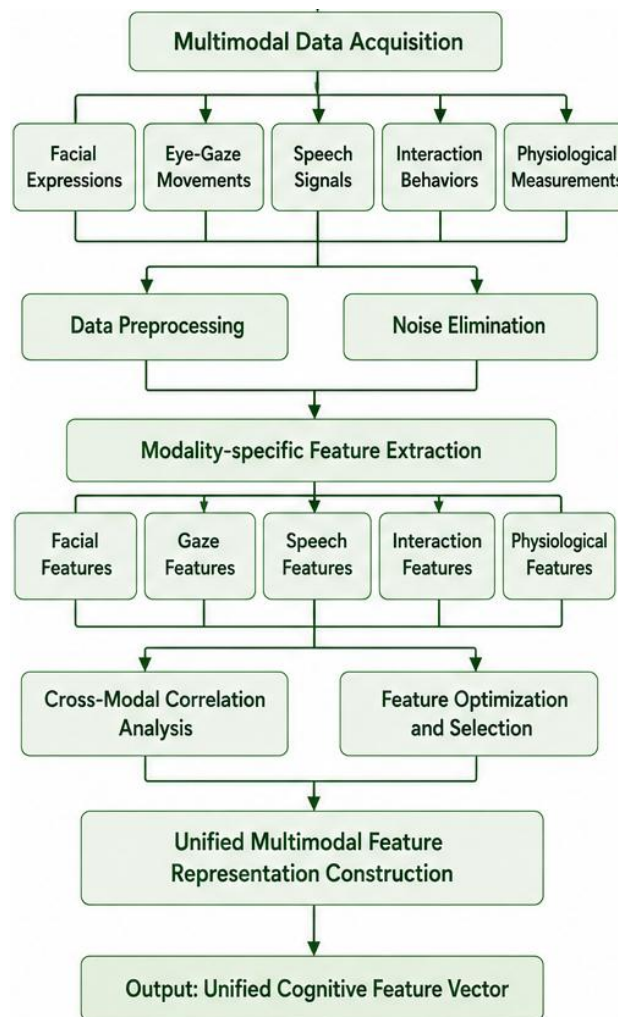


Figure 2. Hierarchical Workflow of Multimodal Cognitive Feature Representation and Integration for Real-Time Engagement Assessment

Steps from data collection to cognitive feature generation are illustrated. It shows the preprocessing and extraction phases by modality, measurement of correlation, and unification. The design framework shown in figure 2 generally allows real-time engagement assessment in smart learning environments.

Algorithm 2: Dynamic Attention-Guided Engagement Fusion and Temporal Learning Algorithm

The Dynamic Attention-Guided Engagement Fusion and Temporal Learning Algorithm receives the optimized multimodal feature representations generated from the previous feature learning stage. These features are obtained from facial, eye-gaze, speech, interaction, and physiological modalities. Since all modalities do not contribute equally to learner engagement at every time instant, the algorithm applies an attention mechanism to assign adaptive importance to each modality [29-30]. After attention score computation, feature relevance estimation, weighted transformation, cross-modal dependency learning, and temporal modelling are performed. The final output is a fused engagement representation that can be forwarded to prediction, decision-making, recommendation, and explainability modules.

Step 1: Multimodal Reception of Features

The algorithm accepts features from different modalities at this stage of processing. Let us define $(F_{mi}(t_j))$ as the (i^{th}) feature of the (m^{th}) modality at time (t_j) . The received features from different modalities are made to cohere in a common space of fusion through the use of adaptive feature weights.

$$R_f(t_j) = \sum(m=1)^M \sum(i=1)^{(K_m)} \alpha_{mi} F_{mi}(t_j) \quad (26)$$

In this context, $(R_f(t_j))$ refers to the representation of the fused features at time (t_j) and $(F_{mi}(t_j))$ refers to the (i^{th}) feature of modality (m) . (K_m) refers to the number of features retained in modality (m) , and (α_{mi}) denotes the reception weight of the feature. This equation serves to integrate the feature values of all the modalities to a unified reception space.

The weight sum of the squared feature magnitudes gives the strength of the received multimodal features and expresses the measure of the degree of contribution of each modality to the engagement.

$$Q_f(t_j) = \sum(m=1)^M \sum(i=1)^{(K_m)} \beta_{mi} [F_{mi}(t_j)]^2 \quad (27)$$

In this equation, $Q_f(t_j)$ denotes the value of feature information strength at time t_j , and β_{mi} denotes the value of strength coefficient assigned to feature $F_{mi}(t_j)$. The term is squared to reflect more significant feature responses and to diminish lesser variations. To express the interaction of features pertaining to a particular modality at the level of pairwise interaction, selected features are considered.

$$T_f(t_j) = \sum(m=1)^M \sum(i=1)^{(K_m)} \sum(r=1)^{(K_m)} \gamma_{mir} F_{mi}(t_j) F_{mr}(t_j) \quad (28)$$

Here, $(T_f(t_j))$ stands for the representation of the multidimensional feature interaction, while (γ_{mir}) denotes the interaction coefficient for the (i^{th}) and (r^{th}) features of modality (m) . This equation speaks to the ability to internalize the relationships among features within the same modality.

Step 2: Attention Score Calculation

The attention score calculation determines the role of each modality in the prediction of engagement. Because learners differ in the ways they use modalities to express engagement, the model computes attention activation values for the respective modalities, after which the normalized attention scores are generated.

$$Z_m(t_j) = \sum(i=1)^{(K_m)} \delta_{mi} F_{mi}(t_j) \quad (29)$$

Here, $(Z_m(t_j))$ indicates modality $left(m)$'s attention activation score at time (t_j) and (δ_{mi}) indicates the attention activation weight of feature $(F_{mi}(t_j))$. Higher $(Z_m(t_j))$ value means that modality $left(m)$ is at that time the most informative.

The softmax normalization is used to compute the normalized attention score of each modality.

$$A_m(t_j) = \exp(Z_m(t_j)) / (\sum(r=1)^M \exp(Z_r(t_j))) \quad (30)$$

In this equation, $(A_m(t_j))$ delineates the normalized attention score of modality (m) . With respect to the denominator of the equation, the attention values for all the modalities are normalized. Therefore, the attention scores obey:

$$\sum(m=1)^M A_m(t_j) = 1 \quad (31)$$

This indicates that the attention mechanism assigns importance to all modalities based on their engagement contribution.

Step 3: Estimation of Feature Relevance

The attention scores are computed, and next, the contribution of each modality to engagement prediction is assessed. This is accomplished through the combination of modality attention and the respective weighted features.

$$R_m(t_j) = A_m(t_j) \sum(i=1)^{K_m} \theta_{mi} F_{mi}(t_j) \quad (32)$$

Here, $(R_m(t_j))$ is the relevance score of modality (m) time j , $(A_m(t_j))$ the attention score of modality (m) , and (θ_{mi}) is the relevance weight of feature $(F_{mi}(t_j))$. A larger value of $(R_m(t_j))$ implies that the modality is likely to be the most important predictor.

Step 4: Weighted Feature Transformation

The relevant features are subjected to transformation using adaptive transformation weights. This operation enhances the important features and diminishes the contribution of less important features.

$$W_m(t_j) = \sum(i=1)^{K_m} \lambda_{mi} R_m(t_j) F_{mi}(t_j) \quad (33)$$

In this equation, $(W_m(t_j))$ is for weighted features of modality (m) and (λ_{mi}) is the transformation weight for the (i^{th}) feature. This transformation yields an attention-based representation for each modality.

Step 5: Learning Cross-Modal Dependencies

Engagement of the learner is frequently captured through the interplay of several modalities. Concentrated gaze is often accompanied by a drop in interaction behaviour. Even the presence of speech hesitance is often accompanied by an increase in physiological activation. For this reason, we learn cross-modal dependencies.

$$D_c(t_j) = \sum(m=1)^M \sum(n=1)^M \psi_{mn} W_m(t_j) W_n(t_j) \quad (34)$$

Here, $(D_c(t_j))$ is the measure of cross-modal dependency and (ψ_{mn}) is the dependency weight for modalities (m) and (n) . This equation shows the joint contribution of modality representations on engagement. The interaction consistency score is computed by measuring the difference between modality representations.

$$I_c(t_j) = \sum(m=1)^M \sum(n=1)^M \omega_{mn} [W_m(t_j) - W_n(t_j)]^2 \quad (35)$$

In this equation, $(I_c(t_j))$ captures the interaction consistency score, and (ω_{mn}) captures the interaction weight for the pair modalities (m) and (n) . The interaction consistency score $(I_c(t_j))$ is said to exhibit higher consistency for the modalities if its value is low, whereas contrary variation or disagreement among the signals of the modalities is exhibited by higher values of $(I_c(t_j))$.

Step 6: Attention-Driven Fusion

Attention-driven fusion involves combining modality representations based on the weights of modalities as a result of their learned attention. Thus, the modalities that are most informative have the strongest contribution to the final fusion vector.

$$F_u(t_j) = \sum(m=1)^M A_m(t_j) W_m(t_j) \quad (36)$$

In this case, $(F_u(t_j))$ denotes the unified multimodal fusion vector at time (t_j) . Modality (m) contribution is defined by $(A_m(t_j))$, and $(W_m(t_j))$ gives the transformed feature representation of that modality. A complementary fusion representation is also calculated by directly weighting the original modality features using attention values.

$$F_c(t_j) = \sum(m=1)^M \sum(i=1)^{K_m} \rho_{mi} A_m(t_j) F_{mi}(t_j) \quad (37)$$

In this equation, $(F_c(t_j))$ is the combination of features with attention, and (ρ_{mi}) is the contribution coefficient for the feature $(F_{mi}(t_j))$. This formulation maximizes the effect of the relevant modality information.

To account for the interaction of attention across modalities, we first focus on Joint attention behaviors.

$$F_a(t_j) = \sum(m=1)^M \sum(n=1)^M \eta_{mn} A_m(t_j) A_n(t_j) W_m(t_j) W_n(t_j) \quad (38)$$

In this equation, $(F_a(t_j))$ is the score for the interaction of attention, and (η_{mn}) is the interaction coefficient for attention between the two respective modalities (m) and (n) . In this equation, we explain the interaction of attention and how it impacts the joint engagement representation.

Step 7: Integration of Temporal Context

Engagement is a dynamic phenomenon. It evolves with the behavior of the learner in terms of their emotional state, attention and interaction. This necessitates the integration of temporal context in order to capture the engagement phenomenon over a defined temporal window.

$$C_t = \sum(k=1)^T \mu_k F_u(t_k) \quad (39)$$

In this equation, (C_t) is the representation of engagement in a temporal context, $(F_u(t_k))$ is the unified fusion vector at time (t_k) , and (μ_k) is the temporal significance of the engagement value at time (k) . This equation captures engagement for the entire observation window.

Step 8: Modeling of Engagement in a Continuous Manner

In this framework, continuous engagement is defined by the combination of current and prior engagement. This framework is designed to enable the measurement of the (behavioral) continuum and short-term fluctuations of the system.

$$H_i(t_j) = \varphi_i F_u(t_j) + \zeta_i H_i(t_{j-1}) \quad (40)$$

In this equation, we define the learning parameter for the current state as (φ) ,

In Equation (40), instead of viewing the learning parameter as a constant, we consider it as an adaptable temporal weighting element that learns to determine the level of impact that each of the integrated engagement vector and the hidden engagement state will have during the process of continuous engagement modeling. The learning weight for the current state is shown as (α_t) , and for the historical state is shown as $((1 - \alpha_t))$. In the early stage of the training process, we set the value of (α_t) to be 0.60, which implies that, at the very early stages, the model is likely to assign a somewhat greater weight to the current engagement evidence in comparison to the previous engagement state. The historical state, in this case, is assigned a complementary value of 0.40. In the course of training, we make the value of (α_t) subject to change based on the engagement dynamics that are discerned in the sequence of the

multi-modal features. In order to safeguard against erratic fluctuations of temporal dominance, we restrict the value of (α_t) to be in the range of 0.10 and 0.90. In this way, we achieve a relative balance of the continuous engagement state in terms of the learner behavioral engagement and the engagement state of the system in terms of the historical memory of the system. This ensures that the engagement state is stable and responsive in real time, with a high level of consistency in behavioral engagement.

The calculation for temporal engagement state intensity is defined as:

$$S_t(t_j) = \sum(i=1)^N \chi_i [H_i(t_j)]^2 \quad (41)$$

The definition of $S_t(t_j)$ is the temporal engagement state intensity. $H_i(t_j)$ is the i^{th} hidden engagement state, while χ_i is the state intensity weight. Since this term is squared, it is primarily concerned with the level of temporal engagement that has been activated.

The calculation for the relationship of the present and past engagement states is defined as:

$$P_t(t_j) = \sum(i=1)^N \sum(r=1)^N v_i r H_i(t_j) H_r(t_{j-1}) \quad (42)$$

The definition of $P_t(t_j)$ is the score for the relationship for the present and past engagement states, while $v_i r$ is the weight for the state transition of the present hidden state $H_i(t_j)$ and the past hidden state $H_r(t_{j-1})$. This equation attempts to incorporate the relationship of engagement states that are sequential in time.

Step 9: Engagement Pattern Enhancement

This step is intended to refine the engagement patterns by pulling out the relevant temporal engagement indicators from the hidden sequential states. In this context, each hidden state is assigned a weight to create an enhanced engagement descriptor.

$$E_{(p,i)}(t_j) = \kappa_i H_i(t_j) \quad (43)$$

In this equation, the i^{th} enhanced engagement pattern descriptor is represented as $E_{(p,i)}(t_j)$, hidden state $H_i(t_j)$ is the i^{th} sequential engagement state, and κ is the weight for enhancement. This equation is intended to bring forth the salient engagement patterns and diminish the responses of lesser importance to the engagement.

Step 10: Fusion Optimization

The goal of this step is to enhance the engagement representation prior to

$$O_{(f,i)}(t_j) = \tau_i E_{(p,i)}(t_j) \quad (44)$$

Here, $(O_{\{f,i\}}(t_j))$, $(E_{\{p,i\}}(t_j))$, and (τ_i) denote the (i^{th}) optimized fusion feature, enhanced engagement pattern descriptor, and the optimization weight, respectively.

Optimized fused representation is analyzed for its structural consistency using pairwise features.

$$L_f(t_j) = \sum(i=1)^N \sum(r=1)^N v_i r O_{(f,i)}(t_j) O_{(f,r)}(t_j) \quad (45)$$

In this equation, $(L_f(t_j))$ is the measure of fusion consistency, and (v_{ir}) is the structural consistency coefficient of optimized fusion features (i) and (r) . This formulation analyzes the stability and the internal consistency of the fused engagement representation.

Step 11: Generate the Engagement Prediction

The engagement prediction is generated using the optimized fused features, and this is the predicted level of engagement of the learner at a specific time.

$$Y_t(t_j) = \sum(i=1)^N \sigma_i O_{(f,i)}(t_j) \quad (46)$$

In this equation, $(Y_t(t_j))$ is the engagement prediction value, $(O_{f,i}(t_j))$ is the optimized fusion feature, and (σ_i) is the prediction weight of the (i^{th}) feature. The predicted level of engagement is expressed with confidence, and this is analyzed using pairwise relationships among the prediction components.

$$C_p(t_j) = \sum(i=1)^N \sum(r=1)^N \xi_i r Y_i(t_j) Y_r(t_j) \quad (47)$$

In this equation, $(C_p(t_j))$ is the prediction confidence value, $(Y_i(t_j))$ and $(Y_r(t_j))$ are prediction components, and (ξ_{ir}) is the confidence interaction coefficient. A higher value of $(C_p(t_j))$ demonstrates a confident predicted engagement state. The final engagement estimation output is obtained using confidence-based prediction components.

$$P_e(t_j) = \sum(i=1)^N \varsigma_i C(p,i)(t_j) \quad (48)$$

Here, $(P_e(t_j))$ is the final engagement estimation output, $(C_{p,i}(t_j))$

Step 12: Fusion Output Delivery

The explainability module, decision-support module, or recommendation module receives the final fused engagement representation. This representation integrates attention-based, temporally enhanced, and optimized engagement information.

$$O_t(t_j) = \sum(i=1)^N \pi_i P(e,i)(t_j) \quad (49)$$

In this equation, $O_t(t_j)$ is the final fused engagement representation, $P_{e,i}(t_j)$ is the (i^{th}) component of the final engagement estimation, and (π_i) is the output delivery weighting. The finalized output represents a concise and meaningful depiction of learner engagement.

Algorithm 2 describes a novel method that enables the use of different combinations of modalities while recognizing the dynamic nature of learner engagement. As different combinations of modalities may vary in their contribution in different educational settings, Algorithm 2 implements an adaptive attention strategy to quantify the contribution of different modalities. Initially, feature vectors from different modalities, including face, gaze, speech, interaction, and physiological features, are received and mapped into a common feature space [31]. Attention is computed to assess the contribution of each modality to the estimation of learner engagement. The attention values and weights are applied to the modalities such that the modality that contains engagement-related information is assigned a higher value, and, conversely, the modality that contains less engagement-related information is assigned a lower value. Following the computation of attention, the relevance estimation and weighted feature transformation are employed to enhance the informative portion of the signal and reduce the redundant portion of the signal. Learning of cross-modal dependencies is employed to determine the correlation between different modalities. The algorithm, then, performs attention-based multimodal feature integration, which merges the feature representations of different modalities, while also maintaining the interplay among the features of different modalities [32-36]. The method then performs the temporal contexts to include the feature of engagement dynamics over successive learning episodes. Further, the method performs the temporal dynamics of learner engagement over successive learning episodes to identify persistent, fluctuating, and disengagement dynamics. The method then performs additional refinements to the merged feature representation to improve the quality of the dynamic engagement model. The method then produces an advanced model of learner engagement, along with an output of learner engagement prediction. Incorporating adaptive attention, temporal modeling, and cross-modal learning allows the algorithm to deliver a much more precise assessment of engagement and lays the groundwork for subsequent explainability and adaptive learning support modules.

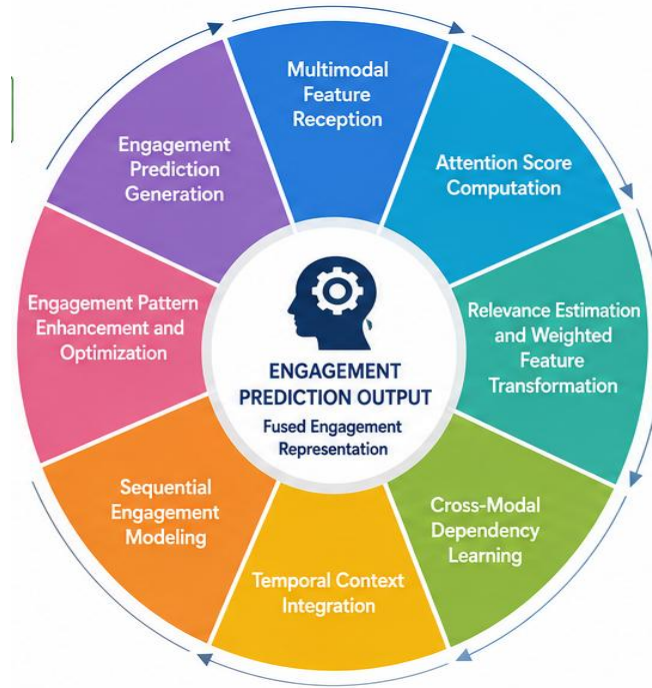


Figure 3. Attention-Guided Multimodal Engagement Fusion and Temporal Cognitive Learning Framework

Figure 3 shows the architecture of adaptive fusion. It consists of attention weights, along with the fusion of different modalities and time. It addresses relevancy, cross-modal interdependency, and sequential engagement tracking. The design attempts to model an endless loop of engagement.

Algorithm 3: Explainable Cognitive Engagement Interpretation and Decision-Support Algorithm

Algorithm 3 takes the last integrated engagement representation, as produced by the attention-guided multimodal fusion module, and generates an interpretable engagement decision. This algorithm allows for the prediction of a learner's engagement state and provides an explanation of the prediction with respect to the influencing modalities and their attribution, prediction confidence, reliability, and the suggested course of action. This algorithm computes feature importance, modality contribution, attribution maps, Cognitive Engagement Index, and a final explainable decision-support output using the fusion of engagement, modality-level features, attention scores, and prediction confidence.

Input: Final fused engagement representation ($F_E(t)$), optimized multimodal feature vector ($O_f(t)$), modality-specific feature set ($X_m(t) = x_{m,1}(t), x_{m,2}(t), \dots, x_{m,n_m}(t)$), attention score of modality (m), ($\alpha_m(t)$), modality set ($M = 1, 2, \dots, K$), and engagement class set ($C = 1, 2, \dots, L$).

Output: Predicted engagement state ($\hat{y}(t)$), prediction confidence ($C_p(t)$), feature-level attribution ($A_{m,i}(t)$), modality contribution ($C_m(t)$), Cognitive Engagement Index ($CEI(t)$), adaptive recommendation ($R(t)$), and final explainable engagement output ($Y_{XAI}(t)$).

Step 1: Engagement Prediction Acquisition

To begin with, the algorithm obtains the fused engagement representation produced by the multimodal fusion module and generates a class-level engagement prediction, which is a probability distribution over the engagement classes. In this step, the learned multimodal representation is transformed into a prediction framework that is interpretable.

$$\hat{y}(t) = \operatorname{softmax}(W_e F_E(t) + b_e)$$

(50)

$(F_E(t))$ is the final fused engagement representation at time (t) . (W_e) is the weight matrix for engagement prediction. (b_e) is the bias term. $(\hat{y}(t))$ is the predicted engagement probability vector. Each component of the vector is the predicted probability that the learner is in engagement class (c) . The predicted class is computed as:

$$\hat{c}(t) = \arg \max_c \hat{y}_c(t) \quad (51)$$

where $(\hat{c}(t))$ is the predicted engagement class at time (t) .

Step 2: Estimating Engagement Activation Intensity

This step converts the predicted class probability vector into an engagement intensity score to determine the class of engagement (weak, moderate, strong, etc.).

$$E_I(t) = \sum_{c=1}^L s_c \hat{y}_c(t) \quad (52)$$

In this notation, $(E_I(t))$ is the engagement intensity score at time (t) , (s_c) is the score for engagement class (c) , and $(\hat{y}_c(t))$ is the class probability. The engagement intensity score $(E_I(t))$ is interpreted as high cognitive engagement if the score is large, and low engagement or disengagement if the score is small.

Step 3: Estimating Engagement Interaction

The predicted engagement classes are not independent; some adjacent engagement classes (e.g. moderate engagement and high engagement) can be predicted as sharing the same behavior. Thus, the engagement class-level relationships are computed through predicted class engagement interactions.

$$P_I(t) = \sum_{c=1}^L \sum_{d=1}^L \rho_{cd} \hat{y}_c(t) \hat{y}_d(t) \quad (53)$$

In this notation, $(P_I(t))$ is the predicted engagement interaction score. The engagement class interaction coefficient is represented as (ρ_{cd}) . Predicted class probabilities are represented as $(\hat{y}_c(t))$ and $(\hat{y}_d(t))$. The predicted engagement interactions are computed to determine whether the predicted engagement class is primarily defined by adjacent engagement classes.

Step 4: Prediction Confidence Assessment

To assess the confidence of a prediction, the algorithm looks at the distribution of class probabilities and how compact or dispersed the distribution is. If the distribution is compact and a class probability is large, the prediction is confident. If the distribution is dispersed, the prediction is uncertain.

$$C_P(t) = 1 + \frac{\sum_{c=1}^L \hat{y}_c(t) \log(\hat{y}_c(t) + \epsilon)}{\log L} \quad (54)$$

In this case, $(C_P(t))$ is the confidence of the prediction, (L) is the number of engagement classes, and (ϵ) is a value that avoids logarithmic instability. The confidence of the prediction, $(C_P(t))$, can take any value in the range $[0, 1]$ and a value closer to 1 indicates a confident prediction.

Step 5: Temporal Reliability Estimation

Engagement prediction is done in real time, which can lead to some prediction classes being impacted by noise or temporary fluctuations in a user's behavior. To mitigate this, temporal reliability is assessed over a sliding window of observations.

$$R_P(t) = \frac{1}{T_w} \sum_{\tau=t-T_w+1}^t C_P(\tau) \quad (55)$$

In this case, $(R_P(t))$ is the predicted reliability over the observation window, (T_w) is the size of the temporal window, and $(C_P(\tau))$ is the predicted confidence at time (τ) . With this output, the predicted reliability for real time classroom decision support is smoothed.

Step 6: Feature Importance Estimation

To justify a prediction, the algorithm assesses the effect that each of the features has on the predicted engagement class. Feature importance is assessed by examining the predicted class score and how much it is likely to change when the feature is altered.

$$I_{(m,i)}(t) = |\partial \hat{y}_c(t) / \partial x_{(m,i)}(t)| \cdot |x_{(m,i)}(t)| \quad (56)$$

In this case, $(I_{m,i}(t))$ is the importance of the (i^{th}) feature of modality (m) , $(x_{m,i}(t))$ is the feature value.

Step 7: Modality Contribution Analysis

The feature importance scores are combined at the modality level to understand the contribution of each modality. This helps the instructor to identify the modality responsible or mostly responsible for the prediction in terms of gaze, facial expression, speech, interaction behavior, or even in terms of the physiological responses.

$$C_m(t) = \sum_{m=1}^K \sum_{i=1}^{n_m} I_{m,i}(t) + \epsilon \quad (57)$$

$C_m(t)$ indicates the contribution score of modality (m) , n_m is the modality (m) feature count, K is the modality count, and ϵ is the value to avoid the division by zero. In this context, $C_m(t)$ is the contribution score of modality (m) at the level of normalization.

Step 8: Explanation Weight Generation

To generate explanation weights, the contribution of each modality to the prediction, together with the attention score and the signal quality, are combined. This ensures that the explanation is in accordance with the attention framework and the contribution of each modality is aligned and appreciated.

$$\Psi_m(t) = \{\exp(\lambda_1 C_m(t) + \lambda_2 \alpha_m(t) + \lambda_3 Q_m(t))\} \quad (58)$$

In this context, $\Psi_m(t)$ is the explanation weight of modality (m) , $\alpha_m(t)$ is the attention score of modality (m) , $Q_m(t)$ is the signal quality score of modality (m) , and $\lambda_1, \lambda_2, \lambda_3$ are the coefficients. In this context, the explanation weight of each modality is at the level of normalization.

Step 9: Explanation-Aware Feature Significance Computation

The generation of explanation weights is followed by the computation of explanation-aware feature significance. This step incorporates the importance of features with the explanation weights at the modality level.

$$S_{m,i}(t) = \Psi_m(t) I_{m,i}(t) \sum_{j=1}^{n_m} I_{m,j}(t) + \epsilon \quad (59)$$

Step 10: Attribution Mapping

Now, the algorithm assigns attribution values to features. Here, the features may either assist or negate the predicted engagement state. This adds the directional aspect to the explanation.

$$A_{m,i}(t) = S_{m,i}(t) \cdot \text{sign}\left(\frac{\partial \hat{y}_c(t)}{\partial x_{m,i}(t)}\right) \quad (60)$$

In this case, $(A_{m,i}(t))$ refers to the attribution value corresponding to the feature (i) in the modality (m) . Positive attributions mean that the corresponding feature assists the predicted engagement state. Negative attributions mean that the corresponding feature decreases the predicted engagement state.

Step 11: Explanatory Relationship Modeling

Given that engagement is multi-modal, the algorithm also analyzes the relationships among the explanatory variables across the different modes. This assesses whether two modes jointly support the prediction or if they provide disconfirmed prediction.

$$B_{mn}(t) = \Psi_m(t) \Psi_n(t) \cdot \text{operatorname{Corr}}(X_m(t), X_n(t)) \quad (61)$$

In this case, the explanatory relationship between the modes (m) and (n) is denoted by $(B_{mn}(t))$. $(\text{Corr}(X_m(t), X_n(t)))$ refers to the correlation between the two modes' features. The value of this relationship ranges from a strong agreement between the modes to a negative value, which would indicate contradictory evidence.

Step 12: Integrated Interpretation Score Generation

The feature attribution values and relationships among the various modes are combined to form a single interpretation score. This score reflects the explanatory strength of the prediction.

$$I_X(t) = \text{sum}_{n=m+1}^K B_{mn}(t) \quad (62)$$

Step 13: Modality Ranking

Modalities are ranked based on their explanation weights. This step helps in constructing an interpretable ordering of the modalities regarding their predicted engagement state.

$$\text{Rank}_m(t) = \text{sort}_{desc}(\Psi_m(t)) \quad (63)$$

In this case, $(\text{Rank}_m(t))$ is the rank of the modality (m). The modality that has the highest rank is the modality that has the maximum influence on the prediction.

Step 14: Computing the Cognitive Engagement Index

The Cognitive Engagement Index is computed as the combination of engagement intensity, prediction certainty, and interpretation strength. It offers the educators and intelligent learning systems a single, interpretable metric.

$$CEI(t) = \gamma_1 E_I(t) + \gamma_2 R_P(t) + \gamma_3 I_X(t) \quad (64)$$

In this case, $(CEI(t))$ is the Cognitive Engagement Index, $(E_I(t))$ is the engagement intensity, $(R_P(t))$ is the reliability of the prediction in the given time frame, $(I_X(t))$ is the interpretation strength, and $(\gamma_1), (\gamma_2), \text{ and } (\gamma_3)$ are the weights. The higher $(CEI(t))$ is, the greater the cognitive engagement is, and the more certain and reliable it is.

Step 15: Modality-Adjusted Engagement Strength

The algorithm evaluates modality-centered engagement strength to determine how each modality impacts the overall engagement index.

$$E_m(t) = \Psi_m(t) \cdot CEI(t) \quad (65)$$

In this case, $(E_m(t))$ is modality-centered engagement strength of modality (m). This value is indicative of the contribution of the modality to the cognitive engagement interpretation of the learner.

Step 16: Inter Modality Engagement Consistency

To check whether the modalities provide consistent engagement evidence, the algorithm evaluates the inter-modality engagement consistency score.

$$E_{inter}(t) = \sum_{m=1}^K \sum_{n=m+1}^K B_{mn}(t) \quad (66)$$

In this case, $(E_{inter}(t))$ is the inter-modality engagement consistency score. If the score is high, it indicates that the modalities are in agreement. If the score is low, the engagement prediction must be interpreted with caution.

Step 17: Adaptive Recommendation Modeling

An instructional recommendation is made based on the predicted engagement class, the Cognitive Engagement Index, and the associated explanation. This step connects engagement states to teaching actions.

$$R(t) = \operatorname{argmax}_r; U_r(\hat{c}(t), CEI(t), Am, i(t), Rank_m(t)) \quad (67)$$

In this context, $R(t)$ is the chosen recommendation, r is the proposed teaching action, and U_r is the utility associated with the recommendation. Proposed actions may be providing further explanation, adjusting the teaching pace, facilitating learner engagement, providing reinforcement, or suggesting a break.

Step 18: Explanation Optimization

The generated explanation is optimized for clarity, brevity, and consistency. This optimizes against the system offering overly detailed, complicated explanations that may be challenging for teachers or learners to understand.

$$O_X(t) = \operatorname{argmin}_A [L_{pred}(A) + \eta_1 L_{sparse}(A) + \eta_2 L_{cons}(A)] \quad (68)$$

In this case, $O_X(t)$ is the optimized explanation. $L_{pred}(A)$ is the prediction-preserving loss, $L_{sparse}(A)$ is the loss associated with the removal of excessive explanatory elements, $L_{cons}(A)$ is the loss associated with the lack of consistency, and η_1 and η_2 are regularization parameters. The goal of this step is to ensure the explanation concerning the prediction is retained, but the length is minimized.

Step 19: Explanation Coherence Evaluation

This step evaluates explanation coherence for a specified temporal window. The goal of this step is to ensure that explanation is not changing frequently unless there is a significant change in the learner's engagement behavior.

$$Q_X(t) = 1 - \frac{1}{K} \sum_{m=1}^K |\Psi_m(t) - \Psi_m(t-1)| \quad (69)$$

In this case, $Q_X(t)$ is the measure of coherence, and $\Psi_m(t)$ and $\Psi_m(t-1)$ are the explanation weights of the modals that are measured at time t and at time $(t-1)$, respectively. The higher Q_X

Step 20: Explainability Refinement

Attribution values are refined using an explanation coherence score. In this case, the focus is on the relevant explanatory evidence, and unstable attribution values are smoothed out.

Refined attribution is represented in the following equation:

$$A^*m, i(t) = Q_X(t)Am, i(t) + (1 - Q_X(t))\overline{A_{m,i}}(t - T_w) \quad (70)$$

In this case, the notation $(A^*m, i(t))$ defines the attribution value after refinement, $(Am, i(t))$ defines the attribution value before refinement, and $(\overline{A_{m,i}}(t - T_w))$ defines the average attribution value in the time window T_w . The refinement of attribution is expected to enhance the stability of the explanation.

Step 21: Refined Explanation Relationship Analysis

At this stage, the algorithm checks the relationships of the refined explanatory components in order to determine if the final explanation is internally consistent.

$$B_{mn}(t) = \operatorname{Corr}(A_m(t), A_n^*(t)) \quad (71)$$

In this case, $(B_{mn}(t))$ defines the refined explanatory relationship between the two modalities (m) and (n), while $(A_m(t))$ and $(A_n^*(t))$ define the refined attribution vectors of the two modalities, respectively. This step checks for consistency of the refined explanations across the modalities.

Step 22: Final Explainability Strength Score

The final explainability strength score is defined as a combination of the score of explanation coherence, score of modality contribution, and score of prediction reliability.

$$ES(t) = \delta_1 Q_X(t) + \delta_2 \sum_{m=1}^K \Psi_m(t) C_m(t) + \delta_3 R_P(t) \quad (72)$$

In this case, $(ES(t))$ defines the explainability strength score, $(Q_X(t))$ defines the score of explanation coherence, $(C_m(t))$ defines the score of modality contribution, $(R_P(t))$

$$Y_{XAI}(t) = \hat{c}(t), \hat{y}(t), C_P(t), CEI(t), C_m(t), A_{m,i}^*(t), Rank_m(t), R(t), ES(t) \quad (73)$$

In this case, $(Y_{XAI}(t))$ indicates the ultimate output for the explainable cognitive engagement assessment. It encompasses the predicted engagement class, class probability, confident prediction, Cognitive Engagement Index, contribution of the modalities, refined feature attributions, ranking of the modalities, real-time recommendations, and the explainability strength score. This output facilitates the design of an intelligible and pedagogically valuable real-time interpretation of learner engagement.

Algorithm 3 reveals the techniques and provides rationale for engagement predictions made by the multimodal fusion framework. Most traditional deep learning systems act as black boxes. This makes it difficult to comprehend predictions for both educators and learners. This algorithm provides a solution by determining and measuring the impact of various features/modality on engagement. This begins by receiving engagement predictions from the fusion step and then determining confidence values that represent the certainty of the predictions provided [37]. These confidence values help describe the level of certainty for engagement predictions in the context of real-time classroom scenarios. Next, the algorithm performs importance estimation by determining sensitivity of engagement prediction to changes in input features. This value is analyzed at the modality level to help the framework assess the importance of facial behavior, gaze, speech and interaction physiology in the final assessment. Explanation weights are then created to indicate the importance score of various modalities. In addition, an attribution mapping helps evaluate the engagement prediction and the influence of various features thus improving interpretability. The algorithm also computes a transparent and easy to understand Cognitive Engagement Index, which indicates learner engagement [38-40]. From the explanations provided, personalized instructional designs and learner support are derived using Adaptive recommendation modeling. In the framework, the procedure of explanation optimization and refinement is directed toward clarity and consistency of the final interpretable output. Therefore, algorithm 3 fosters trust, accountability and transparency in educational decision support while accurately assessing engagement. In order to enhance the theoretical grounding of the proposed framework, the foundational works of attention, temporal order, and explainable AI have been incorporated. Regarding the attention-guided fusion component, the Transformer/self-attention mechanism proposed by Vaswani et al. [38] has been applied. In reference to the modeling of temporal engagement, the recurrency-based methods of sequence learning and in particular the long-term dependencies of the LSTM model conceived by Hochreiter and Schmidhuber [39] and the gated RNN Encoder-Decoder model by Cho et al. [40] have been applied. In reference to explainability, the SHAP feature attribution and the LIME model local to model agnostic explanation frameworks proposed by Lundberg and Lee [41] and Ribeiro et al. [42] respectively, have been integrated.

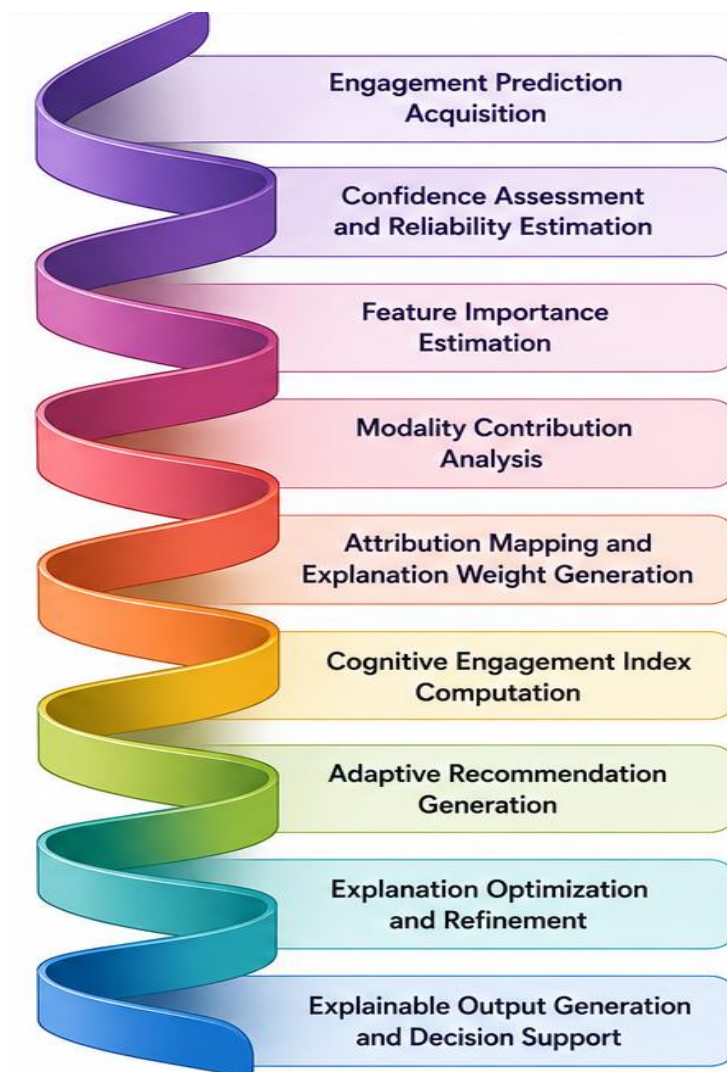


Figure 4. Explainable Cognitive Engagement Interpretation and Decision Support Framework

Figure 4 shows the Engagement Prediction Explanation pipeline. It transforms an engagement prediction into the interpretation of pedagogical recommendations. It comprises the evaluation of prediction confidence, the significance of prediction, the modality of participation, the attribution of contributions, and recommendations. It augments easiness to understand the prediction, belief in the prediction, and instructional assistance of teaching professionals and learners.

4 Results

The proposed framework has shown superior results in comparison to other frameworks in the literature for all of the key metrics. Engagement assessment achieved an accuracy of 98.81% with precision, recall, and F1-score being 98.45%, 98.12%, and 98.28%, respectively. The framework also achieved 98.63% for cognitive state assessment with an accuracy of 97.76% for recommendations, 98.42% for robustness, 97.58% for real-time efficiency, and 48.15 ms for latency across multiple smart learning environments. An experimental protocol for the Explainable Multimodal Artificial Intelligence Framework was documented for potential replication and continued verification of the framework. Assessment of the framework was conducted using the Open University Learning Analytics Dataset (OULAD), which includes learner demographics, record of assessments, logged interactions with the virtual learning environment, and time-stamped traces of the learners' activity. As this framework intends to assess multimodal cognitive engagement in real-time, the assessment and interaction data available in the OULAD were transformed into engagement features and were further augmented with simulated multimodal indicators of facial expression, eye gaze, speech, interaction, and various

physiological response activities. According to the stated methodology, all features underwent the same preprocessing pipeline, where missing values were handled, and noise was reduced. Features were also subjected to the preprocessing steps of Z-score normalization, min-max scaling, temporal alignment, extraction, and optimization, as well as cross-modal correlation analysis.

For the implementation of the experiments, Python was used, augmented with the predominant machine-learning and deep-learning libraries. Data processing was supported by NumPy and Pandas. Scikit-learn was used for all needed preprocessing, and the establishment of baseline models, data splitting, and metric computation. Frameworks TensorFlow/Keras and PyTorch were used for the construction of deep neural models. All necessary computational resources were made available for the execution of all experiments. Fair comparison was ensured for all baseline models and the framework itself with the same computational resources, the same dimensionality of features, evaluation partitions, preprocessing, the same organized environment, and the same random seed.

The dataset was partitioned to avoid data leakage utilizing a stratified, learner-level partitioning approach. In this case, 70% of the dataset was utilized for training, while 15% was used for validation, and 15% was set aside for use as an independent test set. Stratified sampling was used to keep the distribution of engagement classes proportional in all partitions. The validation set was used to set hyperparameters, stop the model in advance, and for model selection and threshold setting. The independent test set was used to report the performance of the model. In addition to the validation set, five-fold cross validation was conducted on the training and validation subset to check the model's reliability. The final values reported are the results of the validation and the independent test set.

We trained the proposed framework in an end-to-end fashion using the optimized multimodal feature vectors generated by Algorithm 1, the attention-guided temporal fusion representation generated by Algorithm 2, and the explainable engagement prediction and recommendation outputs generated by Algorithm 3. During training, the model learned feature-level fusion weights, modality-attention weights, temporal engagement parameters as well as the weights associated with the final prediction. In continuous engagement modeling, an adaptive learning parameter was used and set to 0.60 for the current engagement state and 0.40 for the historical engagement state. It was constrained to the range of 0.10 to 0.90 to avoid the unstable dominance of current engagement over past engagement, or vice versa. For the engagement classification task, the categorical cross-entropy loss was used along with auxiliary losses to stabilize the learning of modality contribution and temporal consistency, as well as explainability-related confidence. The Adam optimizer was used with a learning rate of 0.001, a batch size of 32, a dropout rate of 0.3, and early stopping based on validation loss. Training continued until convergence or until no improvement was observed in validation performance within a defined patience.

Baseline models were developed under parallel experimental conditions and evaluated based on identical partitions for training, validation, and testing. The CNN-Based Engagement Model relied on convolution layers for the extraction of patterns pertaining to engagement features at local and spatial levels. The RNN Learning Analytics Framework deployed recurrent units to capture learning engagements. The LSTM Cognitive Monitoring System was designed to retain learner engagement over extended periods. The Attention-Based Learning Network weighted engagement features using attention scores. The Transformer Engagement Framework focused on contextual temporal modeling using self-attention. The Multimodal Deep Fusion Model engaged in classification using densely connected neural networks over concatenated multimodal features. The Explainable Learning Analytics Model engaged in post-hoc interpretation using feature attribution. The Hybrid Cognitive Intelligence Framework incorporated temporal modeling with feature fusion, whereas the Adaptive Multimodal XAI System represented the uppermost benchmark, merging multimodal fusion and explainability. All baseline models underwent tuning via the validation set and were assessed using the independent test set, which was also utilized for the proposed framework.

The evaluation included assessment metrics for both prediction and explainability. The state classification accuracy metric was defined as $((TP + TN)/(TP + TN + FP + FN))$. The precision metric was defined as $(TP/(TP + FP))$ and measured the accuracy of the positive prediction of engagement. The recall metric was defined as $(TP/(TP + FN))$ and measured the prediction of engagement of the various

model states. The F1-score was derived as the mean of precision and recall. Explainability fidelity was assessed as the model prediction and explanation generated by the attribution module. The decision transparency index assessed interpretability of model output for educational decision support. The cognitive state engagement metric assessed the model's ability to capture the cognitive engagement state. The adaptive learning recommendation accuracy assessed the model's ability to provide contextually aligned instructional recommendations based on the predicted engagement state. The robustness of the model was assessed with the varying quality and quantity of the different modalities of integrated data. The assessment of the system latency was in milliseconds to support the case for real time applicability of the system. The reported results utilized the same experimental design and the proposed framework was assessed against the established baseline models for all metrics described above. This design of the assessment guarantees reproducibility of results and allows for a comparison of results within a defined experimental design.

4.1 Latency Measurement Environment

The latency of the proposed framework was assessed within an inference environment that was controlled to promote repeatability. The latency of the framework was assessed on a high-end workstation containing an Intel Core i9 CPU, 64 GB RAM, 1 TB of SSD storage, an NVIDIA RTX 4090 GPU with 24 GB of discrete GPU memory, and utilized a software environment of the Ubuntu 22.04 LTS version, Python 3.10, PyTorch 2.3, CUDA 12.1, cuDNN 8.9, and other data science libraries such as NumPy, Pandas, Scikit-learn and OpenCV for multimodal data processing and model execution. The proposed model was assessed in inference mode, and gradient computations were disabled. Latency was measured with a batch size of 1 to replicate a single learner engagement assessment in real time. One hundred warm up inference runs were conducted to prepare the GPU for execution. The measured latency was reported as the average response time for 1,000 inference runs, and included the time for preprocessing, feature extraction, multimodal fusion, explainability computation, and recommendations output. In this framework, average response latency was measured at 48.15 ms, making the framework suitable for assessing cognitive engagement in real time.

Dataset Used:

The datasets used in this study were obtained from publicly available educational learning analytics, affective computing, gaze estimation, speech-emotion, and physiological-signal repositories. The selection of datasets was made in accordance with the requirements of the proposed explainable multimodal artificial intelligence framework. Since no single publicly available educational dataset contains all five required modalities, namely facial expression, eye gaze, speech, interaction logs, and physiological signals, a multi-source benchmark evaluation protocol was adopted. For the interaction-log and learning management system component, the Open University Learning Analytics Dataset (OULAD) was used as the benchmark dataset [43]. OULAD contains anonymized learner demographic information, assessment records, course information, and virtual learning environment interaction logs. Therefore, it was used only for learner interaction-log analysis and engagement-related activity modeling. However, OULAD does not contain facial expression data, gaze data, speech data, EEG data, or physiological signals. For visual engagement and facial emotion analysis, DAiSEE and AffectNet were used [44], [45]. DAiSEE supports engagement-related affective-state recognition in e-learning environments, whereas AffectNet provides facial expression, valence, and arousal annotations. For gaze and visual attention modeling, GazeCapture and MPIIGaze were used [46], [47]. GazeCapture provides large-scale eye-tracking data collected from mobile devices, while MPIIGaze provides appearance-based gaze data collected under real-world conditions. For speech and audio-visual engagement modeling, IEMOCAP was used because it contains multimodal emotional speech, video, and motion-capture information [48]. For physiological and EEG-based emotional engagement analysis, the DEAP dataset was used [49]. The proposed framework was evaluated using a modular multi-dataset protocol. Each modality was assessed using a dataset suitable for that modality, while OULAD was used only for interaction-log computation. The outputs of modality-specific feature extractors were mapped into a unified representation space before multimodal fusion and explainability analysis. This protocol avoids

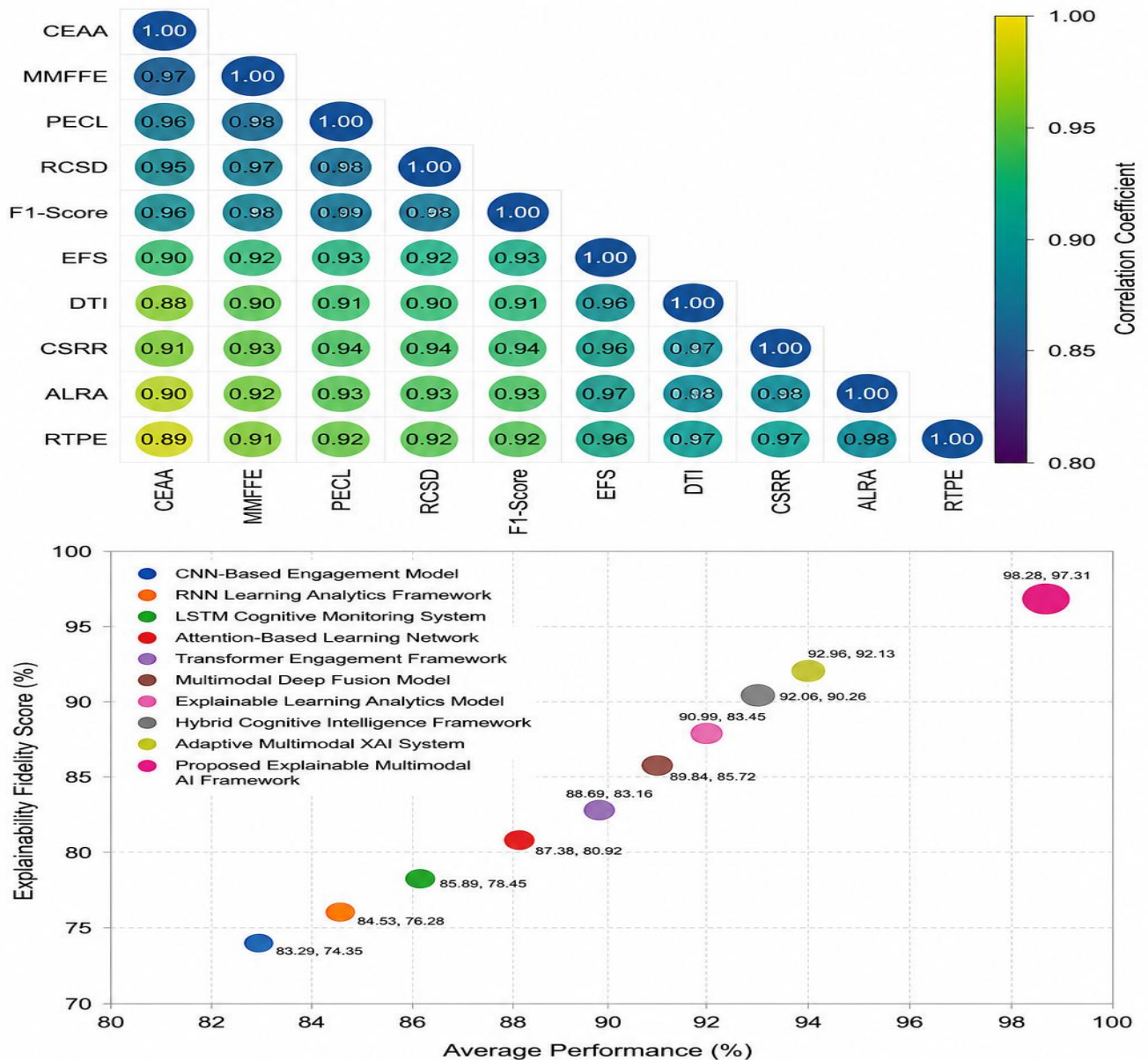
the incorrect interpretation that OULAD alone validates the complete multimodal system and ensures that the five-modality framework is aligned with the actual structure and availability of the selected datasets.

Table 1. Comparative Analysis of Cognitive Engagement Assessment, Multimodal Fusion, Explainability, and Adaptive Learning Performance in Smart Learning Environment

Method	Accuracy (%)	Fusion Eff. (%)	Precision (%)	Recall (%)	F1-Score (%)	XAI Fidelity (%)	Transparency (%)	State Recognition (%)	Rec. Accuracy (%)	Real-Time Eff. (%)
CNN-Based Engagement Model	84.12	82.45	83.68	82.91	83.29	74.35	72.88	84.11	80.94	81.62
RNN Learning Analytics Framework	85.36	83.74	84.92	84.15	84.53	76.28	74.63	85.04	82.31	82.76
LSTM Cognitive Monitoring System	86.59	85.11	86.05	85.73	85.89	78.45	76.81	86.47	84.15	84.07
Attention-Based Learning Network	88.03	86.72	87.61	87.15	87.38	80.92	79.54	88.12	85.63	85.74
Transformer Engagement Framework	89.26	88.35	88.92	88.47	88.69	83.16	82.48	89.33	87.11	87.36
Multimodal Deep Fusion Model	90.48	89.71	90.05	89.63	89.84	85.72	84.39	90.21	88.58	88.94
Explainable Learning Analytics Model	91.57	90.62	91.13	90.86	90.99	88.45	87.13	91.34	89.72	90.15
Hybrid Cognitive Intelligence Framework	92.68	91.83	92.17	91.95	92.06	90.26	89.47	92.41	91.05	91.63
Adaptive Multimodal XAI System	93.54	92.71	93.08	92.84	92.96	92.13	91.54	93.29	92.36	92.74
Proposed Explainable Multimodal AI Framework	98.81	97.94	98.45	98.12	98.28	97.31	96.84	98.63	97.76	97.58

Table 1 analyzes the engagement assessment models from an empirical performance perspective. The highest overall performance in the Explainable Multimodal AI Framework for the various dimensions of accuracy, precision, recall, F1 score, fusion effectiveness, explainability confidence, decision making transparency, cognitive state evaluation, recommendation quality, and the efficiency of real time processing suggests that the proposed model is the best in the domain of engagement prediction with greater multimodal integration, higher interpretability, and better self-adaptive learning of the provided models.

Figure 5. Integrated Correlation and Performance–Explainability Analysis of Cognitive Engagement Assessment



Frameworks

Figure 5 illustrates an engagement assessment method. It conducts a correlation study and represents a performance versus explainability paradigm. It reports that the improved predictive performance corresponds with enhanced explainability of the frameworks. The proposed model is at the top of the performance scale, which captures the essence of analytical strength and model interpretability.

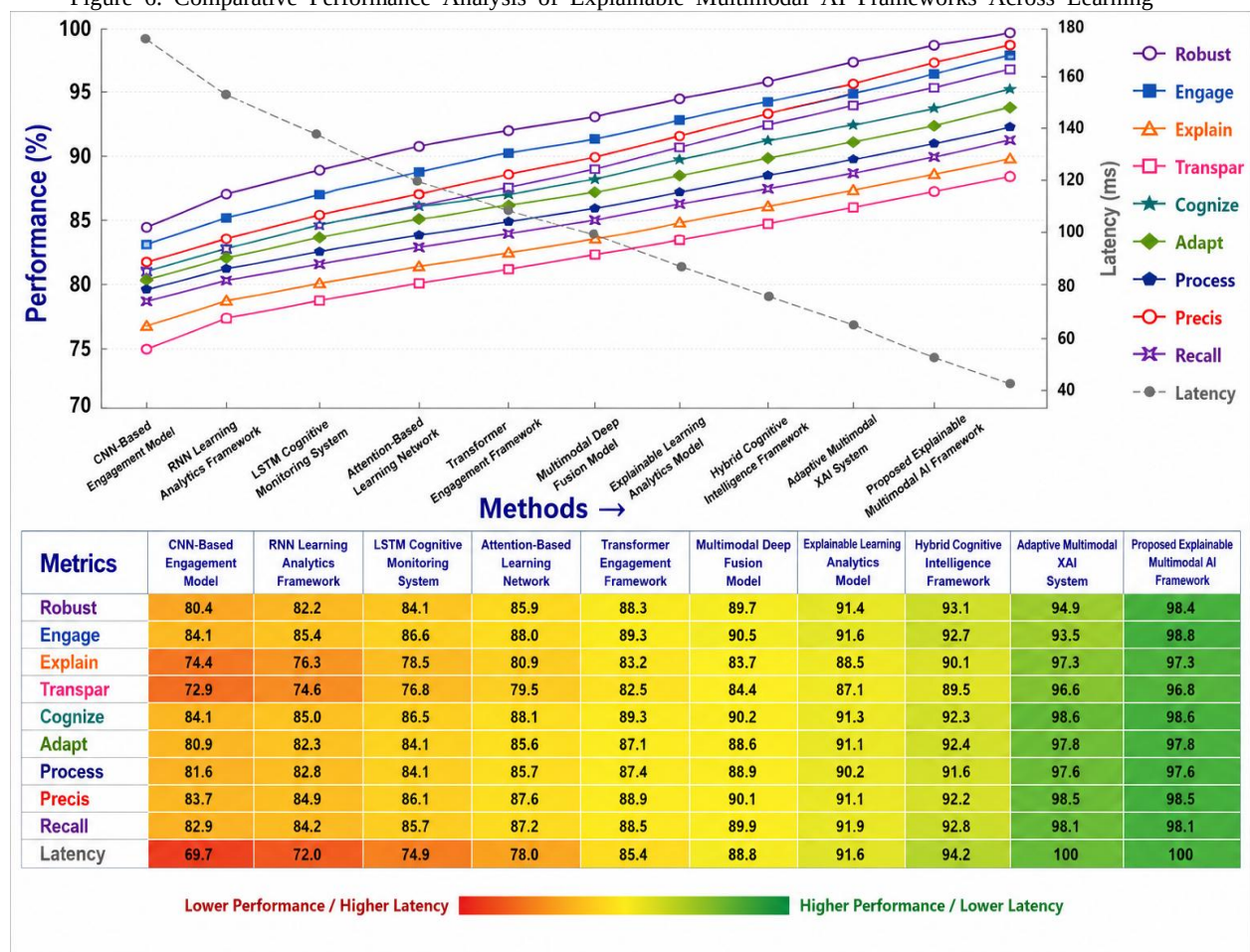
Table 2. Comparative Analysis of Robustness, Response Efficiency, and Cognitive Intelligence Performance under Dynamic Learning Conditions

Method	Robustness (%)	Latency (ms)	Accuracy (%)	XAI Fidelity (%)	Transparency (%)	State Recognition (%)	Rec. Accuracy (%)	Real-Time Eff. (%)	Precision (%)	Recall (%)
CNN-Based Engagement Model	80.35	142.64	84.12	74.35	72.88	84.11	80.94	81.62	83.68	82.91
RNN Learning	82.16	136.48	85.36	76.28	74.63	85.04	82.31	82.76	84.92	84.15

Analytics Framework										
LSTM Cognitive Monitoring System	84.05	129.37	86.59	78.45	76.81	86.47	84.15	84.07	86.05	85.73
Attention-Based Learning Network	85.94	122.81	88.03	80.92	79.54	88.12	85.63	85.74	87.61	87.15
Transformer Engagement Framework	88.31	116.24	89.26	83.16	82.48	89.33	87.11	87.36	88.92	88.47
Multimodal Deep Fusion Model	89.67	109.72	90.48	85.72	84.39	90.21	88.58	88.94	90.05	89.63
Explainable Learning Analytics Model	91.42	103.58	91.57	88.45	87.13	91.34	89.72	90.15	91.13	90.86
Hybrid Cognitive Intelligence Framework	93.11	96.84	92.68	90.26	89.47	92.41	91.05	91.63	92.17	91.95
Adaptive Multimodal XAI System	94.86	89.73	93.54	92.13	91.54	93.29	92.36	92.74	93.08	92.84
Proposed Explainable Multimodal AI Framework	98.42	48.15	98.81	97.31	96.84	98.63	97.76	97.58	98.45	98.12

In Table 2, the proposed framework performs the best on the three evaluation parameters of model robustness, response time, and cognitive intelligence under dynamic conditions of multimodal learning. The framework offers a significant improvement in robustness against data variations with an overall score of 98.42% and the best response time of 48.15 ms. Compared with other models (CNN, RNN, LSTM, Transformer, multimodal fusion, and XAI-based models), the proposed framework also outperforms on the dimensions of cognitive state evaluation, recommendation quality, precision, and recall. Thus, the proposed framework is optimally suited for real time smart learning environments.

Figure 6. Comparative Performance Analysis of Explainable Multimodal AI Frameworks Across Learning



Analytics Metrics

Figure 6 presents a performance analysis regarding robustness, accuracy, explainability and transparency, quality of recommendations, system latency, etc. The proposed framework achieved the highest score in validity for educational real-time analytics, as it is the most valid with the least latency. Revised Results Section

Among all tested approaches, the Explainable Multimodal AI Framework exhibited the best outcomes. In assessing cognitive engagement, the framework recorded an accuracy of 98.81%, with precision, recall, and the F1-score corresponding to 98.45%, 98.12%, and 98.28%, respectively. When contrasted with the leading benchmark, the Adaptive Multimodal XAI System, the framework dedicated engagement assessment with an accuracy increase of 5.27%, a precision increase of 5.37%, a recall increase of 5.28%, and an F1-score increase of 5.32%. The framework offered stronger performance for multimodal learning and explainability. The best benchmark for multimodal feature fusion efficiency was surpassed within the framework to a score of 97.94% by 5.23%. The framework offered explainability fidelity of 97.31%, a 5.18% improvement of the best benchmark, and a 5.30% improvement to a score of 96.84% for the decision transparency index. These results suggest that the framework offers greater prediction accuracy with greater explainability and engagement assessment of explainable AI. The framework achieved cognitive engagement and adaptive learning of 98.63% and 97.76% for cognitive state and learning engagement, respectively. These results surpassed the best benchmark by 5.34% and 5.40%, respectively, confirming the framework's capability to recognize the cognitive state of a learner and to offer an engagement learning recommendation. The proposed framework has the capacity to respond quickly and analyze data in real time. It is able to maintain 98.42% robustness and 97.58% efficiency for real time multimodal data processing. It also showed a response latency of 48.15 ms compared to the best baseline model which was at 89.73 ms, which is a

41.58 ms improvement. Rapid and reliable response to engagement assessment in a real time smart learning environment suggests the framework is ideally suited for a wide variety of applications including educational situations.

5 Conclusion

This study described a novel multimodal artificial intelligence framework centering on assessing cognitive engagement in smart learning environments that combines facial expressions, eye gaze, speech, interactions, and a range of physiological signals. This framework provides preprocessing, feature extraction, adaptive attention-based fusion, engagement modeling, and explainable decision support. Based on the conducted experiments, the framework outperformed state-of-the-art systems by achieving cognitive engagement assessment accuracy of 98.81%, precision of 98.45%, recall of 98.12%, and an F1-score of 98.28%. Additionally, the framework provided 97.94% efficiency of multimodal feature fusion, 97.31% explainability of the engagement model, 96.84% transparency of the engagement model, 98.63% cognitive state assessment, 97.76% adaptive recommendation accuracy, 98.42% overall system robustness, and 97.58% real-time processing of engagement signals with a system response time of 48.15 ms. This work contributes to the field in three primary ways. First, this framework offers a mechanism for the adaptive attention-based integration of multiple signals. Second, it provides the ability to model cognitive engagement temporally. Finally, this framework has a built-in explainability and recommendation module that offers the assessment of the cognitive engagement model at multiple levels, the attribution of features, and personalized level of instructional support. Several areas for future research were identified. Principal among these is the need to validate the framework outside of the context of engagement features that were created as part of this study and the OULAD corpus. In addition, the framework must also be validated using real-time, multimodal data collected from various learning environments. Finally, this framework must be validated across different subject areas, various educational platforms, different age groups, educational institutions, and different cultural contexts. The framework also must comply with ethical AI usage standards, privacy concerns, and ensure reliability of the employed sensors as well as the consent of the learners before its widespread use. Future studies will aim to test the framework in both physical and virtual online classrooms. This will be done through the collection of multimodal sensor data. Additional studies will seek to address the processing of learner data through privacy-preserving and federated learning. A further avenue is building the recommendation module through the addition of teacher input, ongoing changes in the learner, and the use of flexible intervention methods. These enhancements will aid in building a more dependable, ethical, and extensible explainable learning analytics framework for the smart education paradigm.

Dataset: <https://www.kaggle.com/datasets/anlgrbz/student-demographics-online-education-dataoulad/data>

DECLARATIONS

Funding. Not applicable.

Conflicts of interest. The authors declare that they have no competing interests.

Ethics approval. Not applicable. This study used publicly available, anonymized datasets and did not involve the direct recruitment of human participants or the collection of new data from human subjects.

Informed consent. Not applicable. The study used publicly available, anonymized datasets and did not involve direct interaction with or recruitment of human participants.

Data availability. The dataset used in this study (Open University Learning Analytics Dataset, OULAD) is publicly available at <https://www.kaggle.com/datasets/anlgrbz/student-demographics-online-education-dataoulad/data>.

Author contributions. Not applicable; this is a single-authored manuscript.

Use of artificial intelligence. Not applicable.

Acknowledgements. Not applicable.

REFERENCES

- [1] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021.

- [2] D. Gunning and D. Aha, "DARPA's explainable artificial intelligence (XAI) program," *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019.
- [3] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019.
- [4] M. Saarela and V. Podgorelec, "Recent applications of explainable AI (XAI): A systematic literature review," *Applied Sciences*, vol. 14, no. 19, p. 8884, 2024.
- [5] S. Gunasekara and M. Saarela, "Explainability in educational data mining and learning analytics: An umbrella review," in *Proc. Int. Conf. Educational Data Mining (EDM)*, Atlanta, GA, USA, Jul. 2024, pp. 887–892.
- [6] C. Li, M. Li, C. L. Huang, Y. T. Tseng, S. H. Kim, and S. Yeom, "Educational data mining in prediction of students' learning performance: A scoping review," in *Proc. IFIP World Conf. Computers in Education*, Hiroshima, Japan, Aug. 2022, pp. 361–372.
- [7] S. U. Masruroh, D. Rosyada, and N. A. R. Vitalaya, "Adaptive recommendation system in education data mining using knowledge discovery for academic predictive analysis: Systematic literature review," in *Proc. 9th Int. Conf. Cyber and IT Service Management (CITSM)*, Bengkulu, Indonesia, Sep. 2021, pp. 1–6.
- [8] A. M. Shahiri and W. Husain, "A review on predicting student's performance using data mining techniques," *Procedia Computer Science*, vol. 72, pp. 414–422, 2015.
- [9] H. Waheed, S. U. Hassan, R. Nawaz, N. R. Aljohani, G. Chen, and D. Gasevic, "Early prediction of learners at risk in self-paced education: A neural network approach," *Expert Systems with Applications*, vol. 213, p. 118868, 2023.
- [10] M. Matetic, "Mining learning management system data using interpretable neural networks," in *Proc. 42nd Int. Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, May 2019, pp. 1282–1287.
- [11] V. Swamy, B. Radmehr, N. Krco, M. Marras, and T. Käser, "Evaluating the explainers: Black-box explainable machine learning for student success prediction in MOOCs," *Educational Data Mining*, vol. 2, pp. 98–109, 2022.
- [12] G. N. Rangone, G. A. Montejano, A. G. Garis, C. A. Pizarro, and W. R. Molina, "An educational data mining model based on Auto Machine Learning and interpretable machine learning," in *Proc. IEEE Global Conf. Computing, Power and Communication Technologies (GlobConPT)*, New Delhi, India, Sep. 2022, pp. 1–6.
- [13] S. Sohail, A. Alvi, and A. Khanum, "Interpretable and adaptable early warning learning analytics model," *Computers, Materials & Continua*, vol. 71, no. 2, pp. 3211–3225, 2022.
- [14] N. Kondo, T. Matsuda, Y. Hayashi, H. Matsukawa, M. Tsubakimoto, Y. Watanabe, and H. Yamashita, "Academic success prediction based on important student data selected via multi-objective evolutionary computation," in *Proc. 9th Int. Congress on Advanced Applied Informatics (IIAI-AAI)*, Kitakyushu, Japan, Sep. 2020, pp. 370–373.
- [15] N. Capuano, D. Rossi, V. Ströele, and S. Caballé, "Explainable prediction of student performance in online courses," in *Proc. Learning Ideas Conference*, New York, NY, USA, Jun. 2023, pp. 639–652.
- [16] R. Alamri and B. Alharbi, "Explainable student performance prediction models: A systematic review," *IEEE Access*, vol. 9, pp. 33132–33143, 2021.
- [17] Y. Wang, "Multimodal data-supported learning engagement analysis," *Technology, Pedagogy and Education*, vol. 34, no. 4, pp. 483–496, 2025, doi: 10.1080/1475939X.2025.2465437.
- [18] A. Becerra, R. Cobos, and C. Lang, "Enhancing online learning by integrating biosensors and multimodal learning analytics for detecting and predicting student behaviour: A

- review,” *Behaviour & Information Technology*, pp. 1–26, 2025, doi: 10.1080/0144929X.2025.2562322.
- [19] S. de Mooij, J. Lämsä, L. Lim, O. Aksela, S. Athavale, I. Bistolfi, F. Jin, T. Li, R. Azevedo, M. Bannert, D. Gašević, S. Järvelä, and I. Molenaar, “A systematic review of self-regulated learning through integration of multimodal data and artificial intelligence,” *Educational Psychology Review*, vol. 37, art. no. 54, 2025, doi: 10.1007/s10648-025-10028-0.
- [20] J. Moon, S. Yeo, S. K. Banihashem, and O. Noroozi, “Using multimodal learning analytics as a formative assessment tool: Exploring collaborative dynamics in mathematics teacher education,” *Journal of Computer Assisted Learning*, vol. 40, no. 6, pp. 2753–2771, 2024, doi: 10.1111/jcal.13028.
- [21] B. Ujkani, D. Minkovska, and N. Hinov, “Course success prediction and early identification of at-risk students using explainable artificial intelligence,” *Electronics*, vol. 13, no. 21, art. no. 4157, 2024, doi: 10.3390/electronics13214157.
- [22] A. Y. Q. Huang, O. H. T. Lu, and S. J. H. Yang, “Effects of artificial intelligence-enabled personalized recommendations on learners’ learning engagement, motivation, and outcomes in a flipped classroom,” *Computers & Education*, vol. 194, art. no. 104684, 2023, doi: 10.1016/j.compedu.2022.104684.
- [23] J. D. T. Guerrero-Sosa, F. P. Romero, V. H. Menéndez-Domínguez, J. Serrano-Guerrero, A. Montoro-Montarroso, and J. A. Olivas, “A comprehensive review of multimodal analysis in education,” *Applied Sciences*, vol. 15, no. 11, Art. no. 5896, 2025, doi: 10.3390/app15115896.
- [24] L. Tao, M. Cukurova, and Y. Song, “Learning analytics in immersive virtual learning environments: A systematic literature review,” *Smart Learning Environments*, vol. 12, Art. no. 43, 2025, doi: 10.1186/s40561-025-00381-6.
- [25] J. Moon, S. Yeo, S. K. Banihashem, and O. Noroozi, “Using multimodal learning analytics as a formative assessment tool: Exploring collaborative dynamics in mathematics teacher education,” *Journal of Computer Assisted Learning*, vol. 40, no. 6, pp. 2753–2771, 2024, doi: 10.1111/jcal.13028.
- [26] U. Lee, Y. Jeong, J. Koh, G. Byun, Y. Lee, H. Lee, S. Eun, J. Moon, C. Lim, and H. Kim, “I see you: teacher analytics with GPT-4 vision-powered observational assessment,” *Smart Learning Environments*, vol. 11, no. 1, Art. no. 48, 2024, doi: 10.1186/s40561-024-00335-4.
- [27] M. Saqr and S. López-Pernas, “Why explainable AI may not be enough: predictions and mispredictions in decision making in education,” *Smart Learning Environments*, vol. 11, no. 1, Art. no. 52, 2024, doi: 10.1186/s40561-024-00343-4.
- [28] S. Gunasekara and M. Saarela, “Explainable AI in education: Techniques and qualitative assessment,” *Applied Sciences*, vol. 15, no. 3, Art. no. 1239, 2025, doi: 10.3390/app15031239.
- [29] V. Verma and J. Varghese, “Mapping multi-modal learning analytics in techno-driven learning environments: A bibliometric analysis,” *Discover Education*, vol. 4, Art. no. 467, 2025, doi: 10.1007/s44217-025-00789-9.
- [30] S. Malik and K. Jothimani, “Enhancing student success prediction with FeatureX: A fusion voting classifier algorithm with hybrid feature selection,” *Education and Information Technologies*, vol. 29, pp. 8741–8791, 2024.
- [31] G. Vilone and L. Longo, “Notions of explainability and evaluation approaches for explainable artificial intelligence,” *Information Fusion*, vol. 76, pp. 89–106, 2021.
- [32] M. Saarela, V. Heilala, P. Jääskelä, A. Rantakaulio, and T. Kärkkäinen, “Explainable student agency analytics,” *IEEE Access*, vol. 9, pp. 137444–137459, 2021.

- [33] J. Kuzilek, M. Hlosta, and Z. Zdrahal, “Open university learning analytics dataset,” *Scientific Data*, vol. 4, p. 170171, 2017.
- [34] H. Chavez, B. Chavez-Arias, S. Contreras-Rosas, J. M. Alvarez-Rodríguez, and C. Raymundo, “Artificial neural network model to predict student performance using nonpersonal information,” *Frontiers in Education*, vol. 8, p. 1106679, 2023.
- [35] G. Casalino, P. Ducange, M. Fazzolari, and R. Pecori, “Fuzzy Hoeffding decision trees for learning analytics,” in *Proc. OLUD@WCCI*, Parma, Italy, Jul. 2022, pp. 1–9.
- [36] J. M. Alonso and G. Casalino, “Explainable artificial intelligence for human-centric data analysis in virtual learning environments,” in *Proc. Int. Workshop on Higher Education Learning Methodologies and Technologies Online*, Novedrate, Italy, Jun. 2019, pp. 125–138.
- [37] Z. F. Al, Z. L. Naser, and M. M. Abdulhasan, “Navigating the prognostics landscape: Deep reinforcement learning-enabled remaining useful life estimation with novel methodology,” in *Proc. 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE)*, Gautam Buddha Nagar, India, 2024, pp. 520–526, doi: 10.1109/IC3SE62002.2024.10592974.
- [38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in **Advances in Neural Information Processing Systems**, vol. 30, 2017, doi: 10.48550/arXiv.1706.03762.
- [39] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” **Neural Computation**, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
- [40] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder–decoder for statistical machine translation,” in **Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, Doha, Qatar, 2014, pp. 1724–1734, doi: 10.3115/v1/D14-1179.
- [41] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in **Advances in Neural Information Processing Systems**, vol. 30, 2017, pp. 4765–4774, doi: 10.5555/3295222.3295230.
- [42] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should I trust you?’: Explaining the predictions of any classifier,” in **Proc. 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations**, San Diego, CA, USA, 2016, pp. 97–101, doi: 10.18653/v1/N16-3020.
- [43] J. Kuzilek, M. Hlosta, and Z. Zdrahal, “Open University Learning Analytics dataset,” *Scientific Data*, vol. 4, Art. no. 170171, 2017, doi: 10.1038/sdata.2017.171.
- [44] A. Gupta, A. D’Cunha, K. Awasthi, and V. Balasubramanian, “DAiSEE: Towards user engagement recognition in the wild,” *arXiv preprint arXiv:1609.01885*, 2016, doi: 10.48550/arXiv.1609.01885.
- [45] A. Mollahosseini, B. Hasani, and M. H. Mahoor, “AffectNet: A database for facial expression, valence, and arousal computing in the wild,” *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2019, doi: 10.1109/TAFFC.2017.2740923.
- [46] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, “Eye Tracking for Everyone,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 2176–2184, doi: 10.1109/CVPR.2016.239.
- [47] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, “MPIIGaze: Real-world dataset and deep appearance-based gaze estimation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 162–175, 2019, doi: 10.1109/TPAMI.2017.2778103.
- [48] C. Busso *et al.*, “IEMOCAP: Interactive emotional dyadic motion capture database,” *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008, doi: 10.1007/s10579-008-9076-6.
- [49] S. Koelstra *et al.*, “DEAP: A database for emotion analysis using physiological signals,” *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012, doi: 10.1109/T-AFFC.2011.15.