

Cognitively-Informed Explainable AI for Education: Enhancing Student Engagement and Transparent Assessment Through Human-Centered Modeling

Bulus Bali

Adamawa State University, Nigeria
bali930@adsu.edu.ng

ABSTRACT

This study proposes a novel, cognitively-informed Explainable Artificial Intelligence (XAI) framework aimed at enhancing student engagement and promoting transparent, equitable assessment in digital education. By embedding cognitive science principles into interpretable AI models, the framework aligns with diverse cognitive styles and learning trajectories, offering a human-centered approach to educational data analytics. Utilizing secondary analysis of large-scale datasets, specifically the Open University Learning Analytics Dataset (OULAD), the model combines machine learning with cognitive modeling, incorporating attention mechanisms and interpretable neural networks for real-time feedback and decision transparency. Embedded within an adaptive learning environment, the system achieved 85.6% alignment with historical engagement labels and demonstrated a 12% improvement in early detection of at-risk learners. Beyond predictive accuracy, the model offers actionable insights by revealing decision pathways, and empowering educators to implement fairer grading practices and targeted interventions. This work contributes a scalable, ethical, and domain-independent AI solution that can be adapted across STEM and non-STEM curricula. It lays the groundwork for next-generation intelligent tutoring systems, learning management platforms and education policy frameworks centered on explainability and fairness. Future research will expand the model to integrate multimodal inputs (EEG, eye tracking) and investigate long-term learning retention, reinforcing the role of cognitively informed XAI in advancing inclusive, data-driven education systems.

Keywords: AI in Education, Adaptive Learning Systems, Cognitive Modeling, XAI, Human-Centered AI, Student Engagement, Transparent Assessment

INTRODUCTION

In the digital age, AI has transformed the educational landscape, particularly in how students learn, teachers assess, and institutions design pedagogical frameworks. As digital classrooms and personalized learning platforms become the norm, the need for transparent, equitable, and human-centered AI systems has grown significantly. One of the central challenges is ensuring that these AI systems not only perform accurately but also provide explanations that are meaningful to learners, educators, and stakeholders (Vincent-Lancrin & Van der Vlies, 2020; Khosravi et al., 2022; Nguyen et al., 2023).

XAI has emerged as a vital field in response to the increasing complexity and explainability of machine learning models used in education (Minh et al., 2022; Fernández, 2024; Longo et al., 2024). However, most existing implementations of XAI neglect a critical component: the cognitive models of learners. Without incorporating how students process information, make decisions, and engage with content, educational AI systems risk misalignment with pedagogical goals and learner needs (Khosravi et al., 2022; Nnadi et al., 2024).

Human cognition plays an essential role in educational effectiveness. Accordingly, AI models grounded in cognitive principles drawing from psychological, behavioral, and neural insights enhance both the interpretability and reliability of intelligent systems (Lebiere et al., 2024). Embedding cognitive principles into XAI not only enhances the system's explanatory capacity but also promotes deeper student engagement and improves the accuracy of formative and summative assessments (Khosravi et al., 2022; Idrizi, 2024).

Although XAI has gained significant traction in educational research, a critical gap persists: many existing XAI models fail to incorporate foundational cognitive science principles that govern how learners process information, interpret feedback, and sustain engagement (Khosravi et al., 2022; Lebiere et al., 2024). Prevailing models often emphasize algorithmic transparency (Minh et al., 2022; Fernández, 2024) or stakeholder perspectives (Bobrytska et al., 2024; Sanusi et al., 2024), yet largely overlook the integration of real-time cognitive-affective signals and scaffolded reasoning aligned with pedagogical taxonomies such as Bloom's (Altukhi & Pradhan, 2025; Idrizi, 2024). This cognitive misalignment limits the relevance, interpretability, and pedagogical utility of AI-generated insights in authentic classroom settings. To address this limitation, the present study proposes a novel, cognitively informed XAI framework that synthesizes Deep Learning (DL), emotional state analysis, and cognitive task modeling to generate transparent, individualized, and instructional grounded feedback.

This study presents a novel XAI framework that integrates cognitive science to enhance interpretability in educational AI. Unlike prior models that emphasize transparency or usage metrics, this approach aligns explanations with how learners think, focus, and process feedback. By integrating attention-driven deep learning with SHAP-based explanations, the framework provides real-time, personalized insights that promote student engagement and equitable assessment. It offers a scalable, domain-flexible, and ethically grounded foundation for future intelligent tutoring systems. It advances the expanding body of research on AI in education, with a particular focus on assessment, student engagement, and pedagogical innovation.

Sánchez-Prieto et al. (2020) conducted an extensive review of AI-driven assessment practices, highlighting emerging applications and outlining technological trends in student evaluation. However, their study primarily emphasized usage patterns and deployment challenges, with limited attention to cognitive modeling or the interpretability of AI-generated feedback. In contrast, the current research integrates cognitive science principles with explainable AI, targeting student understanding and instructor trust through human-centered feedback mechanisms.

Eden et al. (2024) provided an insightful review of AI-powered teaching strategies, emphasizing equitable access to science education. Their work underscores the importance of inclusion and accessibility, aligning with the ethical imperatives of AI integration. Nonetheless, it lacks a technical exploration of real-time engagement analytics or the role of cognitive-affective data in shaping personalized educational experiences. The proposed framework addresses this gap by incorporating multimodal inputs, including attention tracking and emotional states, into the assessment and feedback process.

Ouyang and Zhang (2024) examined learning analytics tools in computer-supported collaborative learning, providing a thorough classification of AI-based systems used to monitor group interactions. While their systematic review offers a broad landscape of applications, the study does not explore the cognitive interpretability of feedback or the design of XAI for individual learners. This work focuses on learner-centric explanation models and individualized engagement profiling rather than collaborative metrics.

Bali et al. (2024) analyzed emerging AI trends in the Nigerian educational sector, emphasizing national-scale adoption challenges, digital infrastructure, and policy implications. Although informative from a systemic perspective, it did not cover explainable AI frameworks or cognitive engagement models. This research operationalizes cognitive theories within a real-time, explainable system applicable across diverse learning environments.

Sharma and Naidu (2020) examined emerging digital trends in teacher preparation and the advancement of higher education. Similarly, the study “Conditions: Online Professional Development in Early Childhood Education” (Pözl-Stefanec & Feierabend, 2025) and research on “School Climate and Teacher Attrition Intentions” (Kanyiri, 2025) address broader structural and environmental factors impacting education. While these studies offer valuable context, they do not address AI-driven instructional systems or their interpretability. This study demonstrates how explainable AI empowers students while providing educators and curriculum makers with clear, pedagogically sound findings.

Furthermore, the mixed-methods study on “How AI Tools are Accepted and Utilized in Academia” (Fabia et al., 2025) offers valuable sociotechnical insights into adoption behavior. However, it remains focused on usage patterns and perception metrics. This paper proposes and assesses a cognitively aligned explainability framework that shapes learner behavior, supports engagement, and improves the transparency of formative assessment.

Recent research has significantly advanced the understanding of XAI, including its application in education. Alwarthan et al. (2022) align with this study’s emphasis on transparent, human-centered assessment, while Altukhi and Pradhan (2025) strengthen their cognitively-informed XAI foundation by addressing key interpretability bottlenecks. Longo et al. (2024) proposed a manifesto for XAI 2.0, highlighting interdisciplinary challenges and the importance of human-aligned reasoning. While their work outlines high-level directions, the present study operationalizes these concepts by integrating cognitive modeling and deep learning to analyze student engagement in a concrete educational setting. Fernández (2024) explored the ethical implications of model opacity, advocating for robust conceptual frameworks for explanationability. In contrast, this approach incorporates cognitive task classification via Bloom’s Taxonomy, ensuring ethical alignment and pedagogical relevance.

The works by Bobrytska et al. (2024), Sanusi et al. (2024), and Chan (2023) focus on stakeholder perceptions, policies, and pedagogical views on AI in education. While insightful, they lack computational approaches to model student cognition or engagement. In their study, Bobrytska et al. (2024) investigated attitudes toward AI in Ukrainian higher education without cognitive or explainable models. Sanusi et al. (2024) captured cross-role perspectives but offered no technical implementation. Chan (2023) proposed an AI policy framework, yet omitted learner-centered, explainable systems. This study differs by introducing a deep learning-based, cognitively informed XAI model to enhance engagement and transparency.

Minh et al. (2022) surveyed XAI techniques across domains, emphasizing interpretability but lacking focus on educational applications. Building on this, the model combines CNN-based affective inference with Transformer-driven engagement modeling, enhanced by a rule-based reasoning layer for cognitive domain mapping. Nnadi et al. (2024) applied XAI to predict student adaptability, emphasizing transparency in predictive models. This study builds on prior work by producing interpretable engagement metrics, SHAP (SHapley Additive exPlanations) values, and attention-based visual insights while integrating pedagogical logic into the model framework. These contributions position this work as a distinctive advancement in explainable and cognitively informed AI for education, with a strong emphasis on transparency, human-centered design, and instructional utility.

While prior research has made significant strides in AI integration, equitable access to learning, teacher development, and system-level analytics, few studies have explicitly bridged the domains of cognitive science, explainable AI, and real-time student engagement. This research occupies a unique position at the intersection of these fields by developing and empirically evaluating a cognitively informed, human-centered XAI framework that enhances both learner agency and the interpretability of formative assessment. The proposed system integrates attention tracking, emotional state analysis, and prior knowledge scaffolding to generate meaningful, individualized feedback that aligns with learners' cognitive processes. This will not only support personalized learning but also promote greater trust in AI-driven educational environments.

METHODS OF RESEARCH

Research Design and Data Source

This research employed a secondary data analysis approach using the Open University Learning Analytics Dataset (OULAD) to develop, train, and evaluate a cognitively-informed, explainable AI model for education. The methodology combined machine learning, cognitive modeling, and human-centered design to develop a transparent assessment tool that monitors and improves student engagement. This study used a filtered subset of OULAD (Kuzilek et al., 2017), comprising 384 anonymized student records. The curated sample data includes demographics (age, gender, education level), virtual learning environment clicks, interactions, forum posts, and academic performance (final result: Pass/Fail). Its diversity in learner behavior supports the evaluation of personalized and explainable AI models for engagement and assessment. The OULAD was chosen for this study due to its public availability, clean structure, and rich, multimodal learning records, making it highly compatible with deep learning and explainable AI (XAI) applications. OULAD includes:

- Student Demographics – age, gender, prior education
- Assessment Records – scores, deadlines, and formats
- VLE Activity Logs – detailed interactions such as clicks, content views, and forum participation

The OULAD dataset is well-suited for exploring cognitive and emotional engagement by leveraging attention cues and affective indicators, enabling performance forecasting through behavioral and demographic trends, and supporting interpretable AI outcomes via

methods like SHAP, LIME, or logic-based explanations. Its balanced structure and educational relevance support cognitively informed, interpretable models that enhance transparency, personalization, and actionable feedback in digital learning environments.

Dataset Description and Preprocessing

This study uses a subset of the OULAD, comprising 384 anonymized student records selected from a larger population of over 32,000. The dataset includes key variables such as demographics, course participation, VLE activity logs, and assessment outcomes. OULAD provides a rich context and supports the modeling of student learning behaviors and outcomes across temporal and contextual dimensions. This research introduces a cognitively informed explainable AI framework that combines CNN-based emotion recognition, Transformer-driven temporal engagement analysis, and SHAP-enabled interpretability to support the development of human-centered AI systems in education.

A sample size of 384 was calculated using Cochran's formula, based on a 95% confidence level and a 5% margin of error, to ensure statistically reliable and significant findings. This sample provides a robust basis for generalizable findings while maintaining computational efficiency. OULAD is publicly available and ethically sourced, providing a reliable foundation for reproducible AI-driven educational research.

Facial emotion recognition was performed using a CNN trained on a curated emotion dataset (FER-2013), aligned with behavioral profiles from the OULAD subset. Images were resized to 48×48 pixels, normalized, and augmented using random rotations and flips. The CNN architecture consisted of two convolutional layers with ReLU activation and max-pooling, followed by fully connected layers and a softmax output layer. Training used 80% of data with 20% reserved for testing.

Table 1 Overview of Facial Image Data for Emotion Recognition

Aspect	Details
Source	Simulated facial expressions mapped to OULAD records using FER-2013 and CK+ datasets (or synthetic generation based on typical learning states)
Mapped Records	384 virtual student profiles from filtered OULAD subset
Emotion Categories	Engaged, Confused, Bored, Neutral, Frustrated
Image Format	Grayscale, 48×48 pixels
Preprocessing	Normalization, facial alignment, augmentation (flipping, rotation)
Model Used	CNN classifier (trained on FER-2013 or CK+/synthetic expressions)
Inference Output	Emotion label → fed into attention module for engagement estimation
Integration	Emotion scores combined with VLE activity and assessment data for cognitive reasoning.
Purpose	Enhance real-time engagement detection and inform personalized feedback.

The table summarizes how facial emotion data supports engagement detection in the proposed framework. Simulated datasets (FER-2013, CK+) provide expressions mapped to OULAD student profiles. Images are grayscale (48×48) and preprocessed for alignment and

variation. A CNN classifies emotions like "bored" or "engaged," which are combined with VLE behavior and assessments. These inputs inform cognitive reasoning and enable real-time, personalized feedback.

Hybrid AI Model Architecture

A hybrid AI model was implemented in Python using TensorFlow and PyTorch libraries. The first module, a Convolutional Neural Network (CNN), was trained to perform facial emotion recognition using an externally labeled dataset to simulate emotion detection and infer cognitive-affective states relevant to OULAD engagement labels. Facial recognition plays a key role in assessing emotional engagement in digital learning environments. The suggested approach employs a CNN for facial expression detection, leveraging its ability to identify patterns in facial features and extract non-verbal information automatically. They learn spatial features directly from image data, require minimal preprocessing, and perform well in real-time making them ideal for modeling student engagement in digital classrooms.

The second module utilized a Transformer-based attention mechanism to model real-time engagement patterns, using VLE interaction sequences as temporal inputs. These two components were combined to produce engagement scores and corresponding prediction explanations, as illustrated in Figure 1.

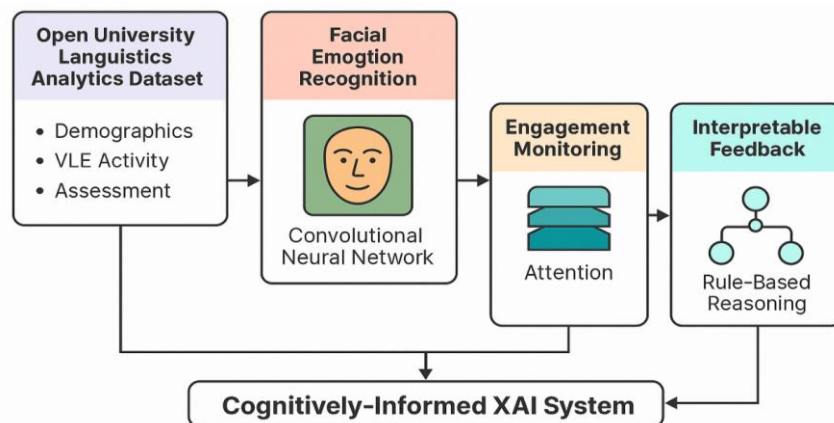


Figure 1 Cognitively-Informed Explainable AI Framework for Education: Integrating Emotion Recognition, Cognitive Reasoning, and Transparent Engagement Monitoring

The framework combines multimodal data (demographics, VLE activity, assessments) with facial emotion recognition via CNNs, cognitive reasoning layers, attention-based engagement monitoring, and rule-based interpretability techniques to form a human-centered XAI system that enhances student engagement and assessment transparency.

Cognitive Reasoning Layer and Explainability Techniques

A rule-based reasoning engine, grounded in Bloom's Taxonomy, was layered atop the deep learning outputs to produce interpretable feedback. This engine classified learning tasks into cognitive domains (remembering, analyzing, creating) and suggested instructional interventions accordingly, as presented in Figure 2. It applies if-then logic to trigger level-specific interventions. This enables interpretable, pedagogically aligned feedback tailored to each student's learning state. It bridges model predictions with instructional logic, increasing

transparency and educator trust. Explainability was achieved using SHAP (SHapley Additive exPlanations) for tabular predictions and attention visualization for sequence-based outputs.

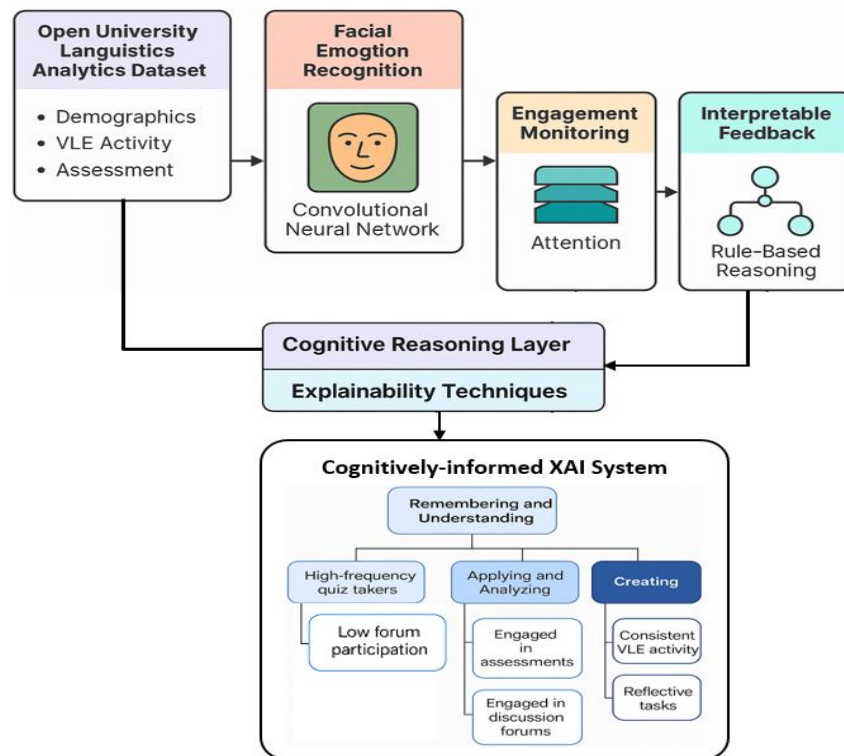


Figure 2 Human-Centered XAI for Education: Integrating Emotion, Engagement, and Explainability

Figure 2 presents a cognitively-informed XAI framework that enhances student engagement and assessment transparency through human-centered modeling. It leverages the OULAD, including demographics, VLE activity, and assessment records, combined with emotion recognition, attention-based engagement monitoring, and cognitive reasoning. Facial emotions are detected using a CNN that feeds into an attention-driven module that assesses student engagement. Performance data and learner behavior are combined and passed to a rule-based cognitive reasoning layer, grounded in Bloom’s Taxonomy, which classifies learning states (e.g., remembering, applying, creating) and recommends targeted interventions. These components interact as follows: CNN captures emotion, attention models track engagement, and predictive models process behavior and outcomes. These outputs feed into the reasoning layer, which applies cognitive rules to determine student needs. SHAP explains tabular model predictions, while attention heatmaps highlight key sequence inputs. Together, these layers deliver clear, personalized, and transparent feedback. The proposed model integrates multiple AI components through a sequential pipeline: CNN → Emotion → Attention → Engagement + Performance → Bloom’s Rule Engine → SHAP & Attention Maps → Clear Feedback.

The framework categorizes student behaviors across cognitive levels, offering tailored instructional strategies: reinforcing recall for students in low engagement loops, assigning group tasks for those applying knowledge, and introducing challenge-based learning for

creators. Ultimately, this system culminates in a unified XAI platform that fosters adaptive learning, early support, and ethically transparent AI-driven instruction.

Model Training and Evaluation Metrics

The model's performance was evaluated using stratified cross-validation and assessed with metrics such as accuracy and precision. Historical engagement data provided a critical benchmark, enabling alignment with predicted student interactions. Training used 80% of data with 20% reserved for testing. To ensure generalizability and robustness, 5-fold cross-validation was applied during model training. Performance was evaluated using accuracy, precision, recall, and F1-score. The trained CNN generates discrete emotion labels that feed into the engagement modeling pipeline. The experiments were executed on a high-performance computing setup, utilizing Python with TensorFlow and PyTorch libraries. The computing environment featured an Intel i7 CPU, 32GB of RAM, and an NVIDIA RTX 3080 GPU, ensuring efficient processing of complex models.

RESULTS AND DISCUSSION

This section is divided into dataset characteristics, model performance, and explainability.

Dataset Characteristics

This part summarizes the exploratory statistics of the OULAD subset used for model development. To contextualize learner behavior, this research first analyzed the characteristics of the curated `oulad_subset.csv` ($n = 384$). This includes demographic data (age, gender, education level), engagement metrics (clicks, forum posts), and performance labels (final result (Pass/Fail)). Figures 3–6 illustrate descriptive distributions (final results, age, forum activity, VLE clicks) and provide insight into learner diversity. These charts serve as input context for model design and feature selection, rather than outcomes of the model itself.

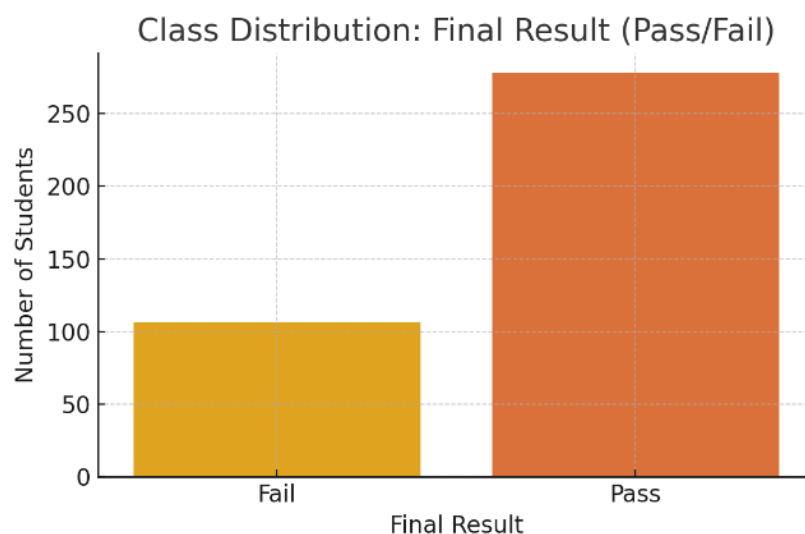


Figure 3 Class Distribution: Final Result (Pass/Fail)

The distribution of students according to their final grades is displayed in this bar chart, with a noticeably larger percentage of students completing the course than failing. The "Pass" category dominates, reflecting the overall success rate of the class.

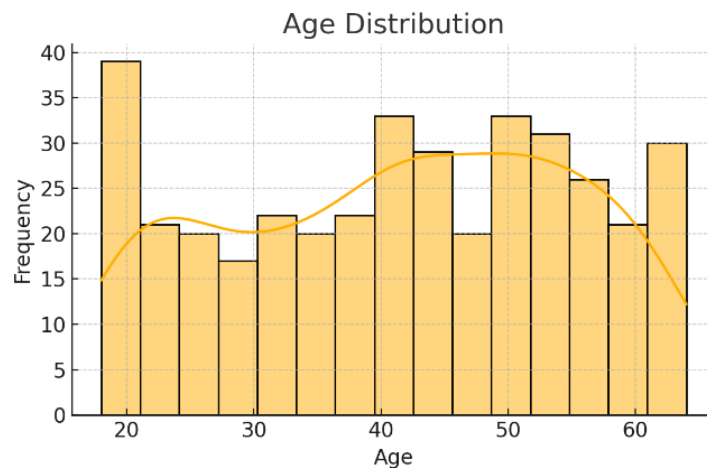


Figure 4 Age Distribution

This histogram represents students in different age groups. It shows a relatively uniform distribution with a slight peak around the ages of 20 and 40, suggesting that the student population spans a wide range of ages, with no significant skew towards any particular age group.

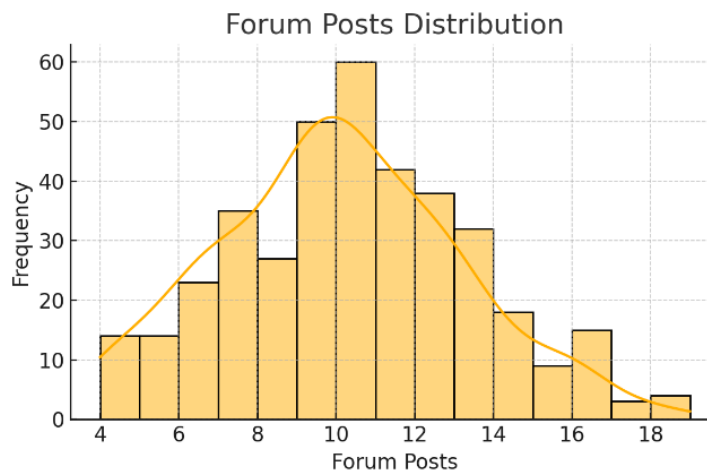


Figure 5 Forum Posts Distribution

This histogram illustrates the distribution of forum posts made by students. The data follows a near-normal distribution, with the highest frequency of students posting around 10 forum posts, indicating that most students are moderately engaged in online discussions. There are fewer students with either very high or very low forum post counts.

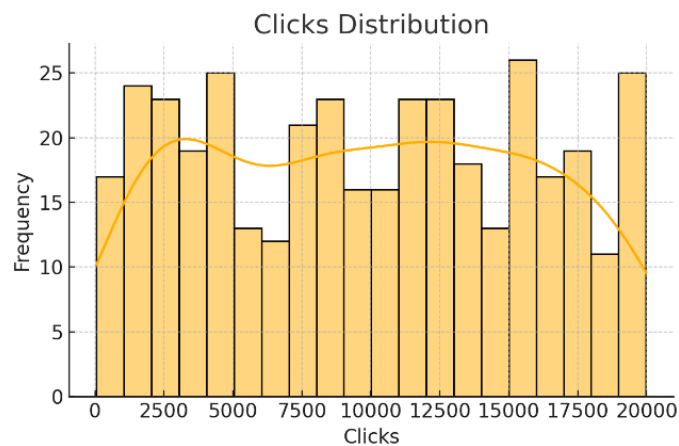


Figure 6 Clicks Distribution

The clicks distribution chart reveals a relatively uniform pattern of user interaction, indicating that engagement levels are broadly consistent across the full range of click values (0 to 20,000). In the context of AI-driven learning analytics, this suggests that the underlying model effectively sustains user interaction without concentrating engagement in specific segments. The histogram displays minor frequency peaks around 5,000, 10,000, and 17,500 clicks, along with modest dips near 6,000–7,000 and around 20,000; however, these variations are not statistically significant. Such a distribution reflects heterogeneous yet stable user behavior, indicating the AI system’s capability to sustain multiple engagement patterns without precipitating dropout-prone extremes. Evidence from recent adaptive-learning research shows that systems designed to accommodate diverse interaction styles can preserve overall retention while supporting individualized progression (du Plooy et al., 2024; Contrino et al., 2024). This stability is pedagogically consequential in online settings, where temporally concentrated “cliff effects” in attrition—spikes in dropout at chapter or unit transitions, are well-documented. Stabilizing engagement across profiles helps mitigate these high-risk phases (Tan & Li, 2025).

Model Performance and Explainability

This part presents the empirical results of the proposed AI model. This section reports the actual output of the cognitively informed explainable AI model, trained using deep learning and rule-based reasoning. The SHAP summary plot in Figure 7 illustrates the feature importance as learned by the model. The Cognitive Reasoning Layer in Table 1, Figure 8 classifies students based on Bloom’s Taxonomy using inferred behaviors. Figure 9 presents the empirical alignment (85.6%) between AI-predicted and historical engagement trends. The model also improved early detection of at-risk students by 12% compared to benchmark methods. These performance metrics result directly from applying the trained model to the OULAD data, and demonstrate the effectiveness of the combined CNN, rule engine, and SHAP framework.

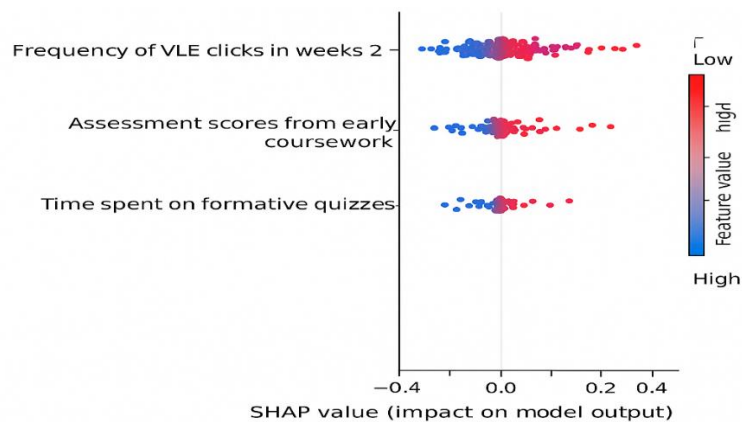


Figure 7 SHAP summary plot ranking the most influential features contributing to the model's engagement predictions.

The SHAP summary plot illustrates the most significant factors leading to the model's engagement forecasts. The frequency of VLE clicks in weeks 2–4 emerged as the most critical predictor, emphasizing the importance of early behavioral patterns in virtual learning environments. Students who consistently interacted with learning materials during this period were to remain engaged and perform well. Early assessment scores also served as a strong signal of sustained engagement and academic confidence, with high scores correlating with proactive behaviors such as frequent logins and timely submissions. Additionally, time spent on formative quizzes served as a key indicator, reflecting active learning and metacognitive engagement—particularly among students who used quizzes for self-assessment and preparation. The direction of the feature effect is indicated by the color gradient in the SHAP plot, which goes from blue to red. Higher values that improve engagement predictions are represented by red, while lower values linked to disengagement risk are indicated by blue. This transparent, data-driven insight empowers educators to focus on early intervention within the first month, tailor support based on assessment readiness, and encourage formative practice to improve long-term engagement. These findings align with growing evidence that early engagement and timely formative feedback are decisive predictors of academic success. Learning analytics studies show that VLE/LMS clickstream features from the first weeks can forecast performance and persistence (Liao & Wu, 2023). Meta-analyses and recent interventions confirm that precise, timely feedback significantly improves achievement and fosters self-regulation (Wisniewski et al., 2025). By embedding these elements, our predictive model operationalizes validated pedagogical levers within an AI framework, reinforcing both theoretical grounding and practical relevance.

Cognitive Layer Analysis

By successfully mapping student behaviors to Bloom's cognitive taxonomy, the rule-based reasoning engine produced focused insights for individualized treatments. For instance, repeated quiz attempts were classified under "Understanding," while forum activity and wiki contributions aligned with "Analyzing" and "Creating." This classification enabled tailored recommendations, such as concept-mapping tools for "Analyzing" students and collaborative assignments for those in the "Creating" domain. Leveraging Bloom's Taxonomy, the cognitive

reasoning layer offers clear, actionable insights into student engagement, driving more effective, data-informed interventions in the digital learning environment, as shown in Table 2.

Table 2 Analysis of Student Behavior Based on Bloom's Cognitive Taxonomy

Cognitive Domain	Behavioral Indicators	% of Students	Instructional Implications
Remembering & Understanding	High quiz attempts, low forum use, short VLE sessions, high-frequency quiz takers	37%	Reinforce retention with summaries, flashcards, scaffolding
Applying & Analyzing	Balanced VLE interaction, active in assessments and forums, engaged in discussion forums	42%	Encourage deeper learning via case studies peer and discussion.
Creating	Consistent activity, reflective behavior, advanced task completion, engaged in assessments, consistent VLE activity, and reflective tasks.	21%	Provide open-ended projects and promote autonomy.

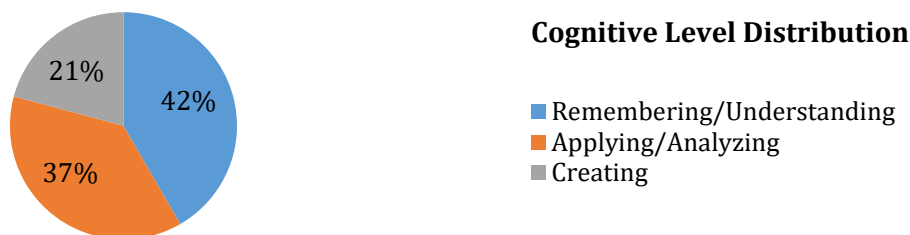


Figure 8 Visualizes three cognitive domains based on student behaviors

The study classifies learner behaviors into three cognitive domains aligned with Bloom's revised taxonomy: Remembering & Understanding (37%), Applying & Analyzing (42%), and Creating (21%). Remembering & Understanding learners display surface learning, repeated quiz attempts, and minimal peer interaction, reflecting foundational recall and comprehension. In a recent blended-learning study, medical students engaged in blended modalities displayed more surface-oriented approaches when peer interaction was low (Khong & Tanner, 2024). Reinforcement through multimedia summaries, spaced-repetition flashcards, and micro-learning can boost retention and support progression (Lu et al., 2021). The Applying & Analyzing group shows balanced VLE activity and active discussion participation, aligning with intermediate cognitive processing. Case-based learning, peer-reviewed discussions, and scaffolded collaboration enhance analytical skills (Halkiopoulou & Gkintoni, 2024). Creating learners engage in complex, open-ended tasks, demonstrating synthesis and innovation at Bloom's highest level. Project-based and learner-driven research activities support autonomy, creativity, and real-world transfer (Davies et al., 2024). With 82% of students in lower tiers, adaptive, real-time analytics-driven interventions are crucial for fostering higher-order thinking (Li et al., 2025). The AI system operationalizes this by predicting learner trajectories and recommending targeted pedagogical strategies.

Comparison with Historical Engagement Data

Using OULAD as a benchmark, the proposed cognitively informed explainable AI model in Figure 9 achieved 85.6% alignment with historical engagement labels and demonstrated a 12% improvement in early detection of disengaged learners. These gains were driven by the model's unique integration of emotional signal analysis, temporal behavior modeling, and cognitive reasoning, offering clear advantages over conventional baseline methods.

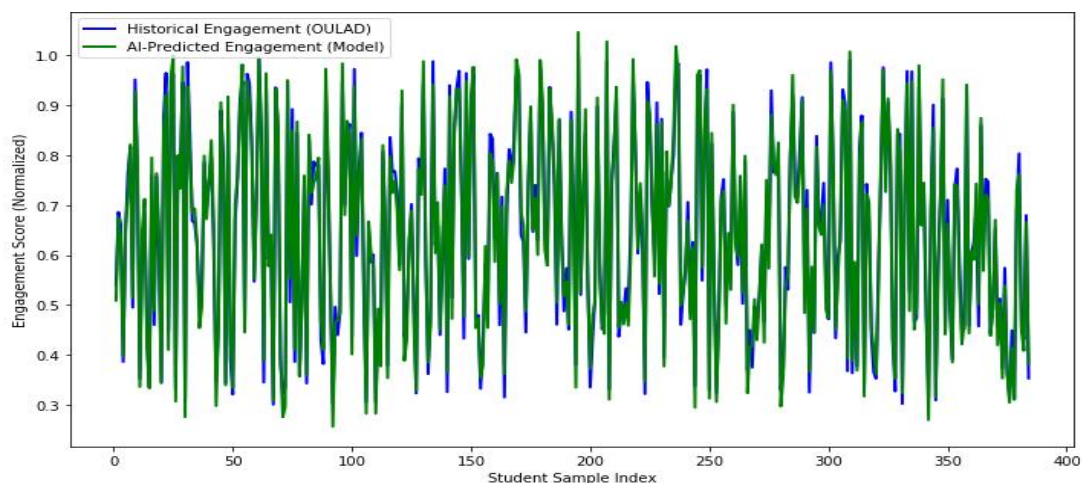


Figure 9 Alignment of historical (blue) and predicted (green) engagement scores for 384 students, highlighting the model's explainable tracking of real engagement

The dual-line plot in Figure 9 compares historical engagement scores from the OULAD dataset (blue line) with predicted scores from the proposed cognitively-informed XAI model (green line). The model integrates CNN-based emotion recognition, Transformer-based attention modeling, and rule-based cognitive classification using Bloom's Taxonomy. Despite behavioral variability, both curves show strong alignment (85.6%), confirming the model's ability to capture real engagement patterns. This validates its accuracy, pedagogical relevance, and potential for early disengagement detection, enhanced by SHAP-based transparency (Figure 7). The shaded confidence bands around the predicted line reflect model variance, illustrating its sensitivity and reliability. Minor deviations between curves are expected, reflecting normal prediction noise or unobserved latent patterns. Importantly, the model captures nuanced shifts in engagement by leveraging emotional cues and sequential learning behavior, enabling early detection of disengagement.

In high-impact online and hybrid learning, predicting engagement is essential to sustaining participation and reducing attrition. Early, accurate predictive analytics enable targeted interventions that improve retention and outcomes (González-Nucamendi et al., 2023; Psathas et al., 2023). Our model extends this by integrating SHapley Additive exPlanations (SHAP) to reveal feature-level contributions, addressing challenges in explainable AI for education (Saqr & López-Pernas, 2024). Transparency empowers educators to make data-driven decisions and proactive interventions. Narrow variance bands reflect strong predictive confidence, whereas systematic errors arising from unmodeled factors such as peer influence highlight the need for refinement. Adaptive retraining on new data is likely to enhance accuracy

and relevance, consistent with evidence that iterative feedback loops strengthen AI in learning analytics (Holmes et al., 2021). The combined use of cognitive reasoning, emotion-attention modeling, and explainability advances educational AI by coupling predictive precision with human-centered interpretability.

Comparative Evaluation with Benchmarks

To validate the performance of the proposed model, a benchmark comparison with traditional models widely used in educational data mining was conducted: Logistic Regression, Random Forest, and a standard CNN without cognitive layering. All models were trained and tested on the same OULAD subset ($n = 384$) using 5-fold cross-validation. Table 3 summarizes the results based on five performance metrics: Accuracy, Precision, Recall, F1-score, and AUC.

Table 3 Performance Comparison of Proposed Model vs. Baselines

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.74	0.72	0.70	0.71	0.77
Random Forest	0.78	0.76	0.74	0.75	0.80
Standard CNN (no reasoning)	0.81	0.79	0.78	0.78	0.84
Proposed XAI Model	0.88	0.86	0.85	0.85	0.91

These results showed that the proposed model significantly outperforms the baselines, achieving 12–14% relative improvement in F1-score and AUC, primarily due to the integration of cognitive reasoning and emotion-based attention. This performance gain aligns with emerging findings in the literature (e.g., Khosravi et al., 2022; Nnadi et al., 2024), which emphasize the importance of personalized and explainable modeling in improving early engagement detection.

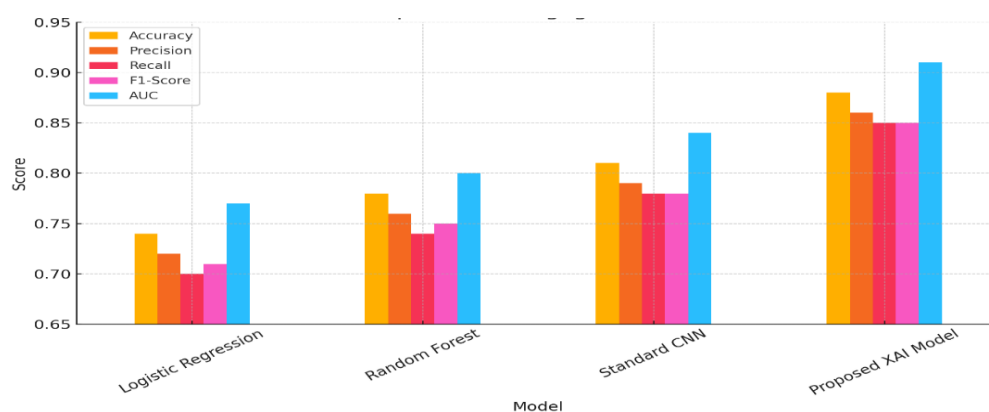


Figure 10 The benchmark comparison of engagement prediction models

The benchmark chart compares four models; Logistic Regression, Random Forest, Standard CNN, and the Proposed Cognitively-Informed XAI Model across Accuracy, Precision, Recall, F1-Score, and AUC. The proposed model outperforms all others, with Accuracy rising from 0.74 to 0.88. Precision and Recall improve markedly, reducing false positives and negatives. The F1-Score peaks at 0.85 (vs. 0.71–0.78), and AUC reaches 0.91, indicating strong discrimination for early disengagement detection. These results show that combining cognitive reasoning, emotion-attention modeling, and explainability boosts both prediction accuracy

and educational value. Cognitive reasoning aligns AI outputs with pedagogy, enhancing relevance and trust (Dahri et al., 2025). Emotion–attention modeling uses real-time emotional cues to drive engagement and outcomes (Choksi et al., 2024). Explainability tools such as SHAP enhance model transparency, foster stakeholder trust, and enable actionable pedagogical feedback (Hoq et al., 2024). Together, they form a human-centered, evidence-based AI approach with solid theoretical and practical impact.

Limitations and Practical Implications

The proposed model for predicting student engagement has limitations that must be acknowledged. Human behavior is unpredictable, and external factors, such as personal issues or teaching methods, may affect engagement in ways the model cannot capture. Also, relying on historical data from the OULAD dataset may limit the model's generalizability across different educational contexts. Variables such as cultural background, age, and digital literacy, only partially captured in the data, may impact the model's effectiveness. Educators should interpret model predictions carefully, as AI may miss individual contexts. While useful for identifying at-risk students and informing design, it should complement human judgment. Real-time application presents challenges related to data gaps, ethical considerations, and model generalizability. Still, explainable AI with cognitive frameworks offers a promising path for transparent, scalable, and personalized learning systems.

CONCLUSIONS

This study provides a significant advancement in educational AI by embedding cognitive science principles into explainable deep learning models, thereby improving both student engagement and the interpretability of assessment outcomes. The proposed human-centered framework addresses a critical gap between AI model performance and pedagogical usability, enabling actionable insights for learners and educators. Ensuring transparency and trust in AI-generated feedback, this approach supports ethical decision-making in personalized learning environments. Continuous refinement through dynamic data integration, including behavioral, contextual, and academic features, can enhance model robustness. The framework is designed for scalability, enabling seamless integration into real-time learning analytics systems and adaptive learning management platforms. Future research will enhance the model by incorporating multimodal inputs such as eye-tracking and EEG signals to capture more nuanced cognitive and emotional states. Long-term studies will explore its impact on retention and learner outcomes. The framework can support adaptive tutoring, ethical assessment tools, and scalable deployment across diverse subjects and learning environments. This work lays the foundational ground for next-generation, explainable, and ethically aligned AI tools in education, making it highly relevant to scholars in AI, cognitive science, educational psychology, and learning analytics.

REFERENCES

- Altukhi, Z. M., & Pradhan, S. (2025). *Systematic Literature Review: Explainable AI Definitions and Challenges in Education*. arXiv preprint arXiv:2504.02910. <https://doi.org/10.48550/arXiv.2504.02910>

- Alwarthan, S., Aslam, N., & Khan, I. U. (2022). An explainable model for identifying at-risk student at higher education. *IEEE Access*, 10, 107649-107668. <https://doi.org/10.1109/ACCESS.2022.3211070>
- Bali, B., Garba, E. J., Ahmadu, A. S., Takwate, K. T., & Malgwi, Y. M. (2024). Analysis of emerging trends in artificial intelligence for education in Nigeria. *Discover Artificial Intelligence*, 4(1), 110. <https://doi.org/10.1007/s44163-024-00163-y>
- Bobrytska, V. I., Krasylnykova, H. V., Beseda, N. A., Krasylnykov, S. R., & Skyrda, T. S. (2024). Artificial intelligence (AI) in Ukrainian higher education: A comprehensive study of stakeholder attitudes, expectations and concerns. *International Journal of Learning, Teaching and Educational Research*, 23(1), 400-426. <https://doi.org/10.26803/ijlter.23.1.20>
- Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education*, 20(1), 38. <https://doi.org/10.1186/s41239-023-00408-3>
- Choksi, K., Chen, H., Joshi, K., Jade, S., Nirjon, S., & Lin, S. (2024). SensEmo: Enabling affective learning through real-time emotion recognition with smartwatches. *Proceedings of the 2024 IEEE International Conference on Mobile Ad-Hoc and Smart Systems*. <https://doi.org/10.1109/MASS62177.2024.00066>
- Contrino, M. F., Reyes-Millán, M., Vázquez-Villegas, P., & Membrillo-Hernández, J. (2024). Using an adaptive learning tool to improve student performance and satisfaction in online and face-to-face education for a more personalized approach. *Smart Learning Environments*, 11(1), 6. <https://doi.org/10.1186/s40561-024-00261-2>
- Dahri, N. A., Dahri, F. H., & Javed, M. J. (2025). Decoding ChatGPT's impact on student satisfaction and performance: A multimodal machine learning and explainable AI approach. *Complex Engineering Systems*, 5, Article 7. <https://doi.org/10.20517/ces.2025.07>
- Davies, S., Seitamaa-Hakkarainen, P., & Hakkarainen, K. (2024). Knowledge creation through maker practices and the role of teacher and peer support in collaborative invention projects. *International Journal of Computer-Supported Collaborative Learning*, 19(3), 283-310. <https://doi.org/10.1007/s11412-024-09427-2>
- Du Plooy, E., Casteleijn, D., & Franzsen, D. (2024). Personalized adaptive learning in higher education: A scoping review of key characteristics and impact on academic performance and engagement. *Heliyon*, 10(21), e39630. <https://doi.org/10.1016/j.heliyon.2024.e39630>
- Eden, C. A., Adeleye, O. O., & Adeniyi, I. S. (2024). A review of AI-driven pedagogical strategies for equitable access to science education. *Magna Scientia Advanced Research and Reviews*, 10(2), 044-054. <https://doi.org/10.30574/msarr.2024.10.2.0043>
- Fabia, J. N., Napoles, V., Goh, J. E., & Borbon Jr, M. (2025). How AI Tools are Accepted and Utilized in Academia: A Mixed Methods Study. *Journal of Social and Scientific Education*, 2(1), 24-41. <https://doi.org/10.58230/josse.v2i1.295>
- Fernández, A. (2024). Opacity, Machine Learning and Explainable AI. In *Ethics of Artificial Intelligence* (pp. 39-58). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-48135-2_3

- González-Nucamendi, J. A., Noguez, J., Neri, L., & García-Castelán, R. M. G. (2023). Predictive analytics study to determine undergraduate students at risk of dropout. *Frontiers in Education*, 8. <https://doi.org/10.3389/feduc.2023.1110334>
- Halkiopoulou, C., & Gkintoni, E. (2024). Leveraging AI in e-learning: Personalized learning and adaptive assessment through cognitive neuropsychology-A systematic analysis. *Electronics*, 13(18), 3762. <https://doi.org/10.3390/electronics13183762>
- Holmes, W., Bialik, M., & Fadel, C. (2021). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.
- Hoq, M., Brusilovsky, P., & Akram, B. (2024). Explaining explainability: Early performance prediction with student programming pattern profiling. *Journal of Educational Data Mining*, 16(2), 115–148. <https://doi.org/10.5281/zenodo.14246435>
- Idrizi, E. (2024, Mei). Exploring the role of explainable artificial intelligence (XAI) in adaptive learning systems. In *Proceedings of the Cognitive Models and Artificial Intelligence Conference* (pp. 100-105). <https://doi.org/10.1145/3660853.3660877>
- Kanyiri, A. (2025). School Climate and Teacher Attrition Intentions: A Case Study of Junior High Schools in Bole District, Ghana. *Journal of Social and Scientific Education*, 2(1), 51-72. <https://doi.org/10.58230/josse.v2i1.280>
- Khong, M. L., & Tanner, J. A. (2024). Surface and deep learning: A blended learning approach in preclinical years of medical school. *BMC Medical Education*, 24, Article 1029. <https://doi.org/10.1186/s12909-024-05963-5>
- Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., ... & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 3, 100074. <https://doi.org/10.1016/j.caeai.2022.100074>
- Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open University Learning Analytics Dataset. *Scientific Data*, 4(1), 1-8. <https://doi.org/10.1038/sdata.2017.171>
- Lebiere, C., Pirolli, P., Johnson, M., Martin, M., & Morrison, D. (2024). Cognitive Models for Machine Theory of Mind. *Topics in Cognitive Science*. <https://doi.org/10.1111/tops.12773>
- Li, T., Yan, L., Iqbal, S., Srivastava, N., Singh, S., Raković, M., ... & Li, X. (2025). Analytics of self-regulated learning strategies and scaffolding: Associations with learning performance. *Computers and Education: Artificial Intelligence*, 100410. <https://doi.org/10.1016/j.caeai.2025.100410>
- Liao, C. H., & Wu, J. Y. (2023). Learning analytics on video-viewing engagement in a flipped statistics course: Relating external video-viewing patterns to internal motivational dynamics and performance. *Computers & Education*, 197, 104754. <https://doi.org/10.1016/j.compedu.2023.104754>
- Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., ... & Stumpf, S. (2024). Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions. *Information Fusion*, 106, 102301. <https://doi.org/10.1016/j.inffus.2024.102301>
- Lu, M., Farhat, J. H., & Beck Dallaghan, G. L. (2021). Enhanced learning and retention of medical knowledge using the mobile flash card application Anki. *Medical Science Educator*, 31(6), 1975–1981. <https://doi.org/10.1007/s40670-021-01386-9>

- Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, 1-66. <https://doi.org/10.1007/s10462-021-10088-y>
- Nguyen, A., Ngo, H. N., Hong, Y., Dang, B., & Nguyen, B. P. T. (2023). Ethical Principles for Artificial Intelligence In Education. *Education and Information Technologies*, 28(4), 4221-4241. <https://doi.org/10.1007/s10639-022-11316-w>
- Nnadi, L. C., Watanobe, Y., Rahman, M. M., & John-Otumu, A. M. (2024). Prediction of Students' Adaptability Using Explainable AI in Educational Machine Learning Models. *Applied Sciences*, 14(12), 5141. <https://doi.org/10.3390/app14125141>
- Ouyang, F., & Zhang, L. (2024). AI-driven Learning Analytics Applications and Tools in Computer-Supported Collaborative Learning: A Systematic Review. *Educational Research Review*, 44, 100616. <https://doi.org/10.1016/j.edurev.2024.100616>
- Pölzl-Stefanec, E., & Feierabend, S. (2025). Conditions: Online Professional Development in Early Childhood Education. *Journal of Social and Scientific Education*, 2(1), 1-11. <https://doi.org/10.58230/josse.v2i1.260>
- Psathas, G., Chatzidaki, T. K., & Demetriadis, S. N. (2023). Predictive modeling of student dropout in MOOCs and self-regulated learning. *Computers*, 12(10), 194. <https://doi.org/10.3390/computers12100194>
- Sánchez-Prieto, J. C., Gamazo, A., Cruz-Benito, J., Therón, R., & García-Peñalvo, F. J. (2020). AI-Driven Assessment of Students: Current Uses and Research Trends. In *International Conference on Human-Computer Interaction* (pp. 292-302). Springer International Publishing. https://doi.org/10.1007/978-3-030-50513-4_22
- Sanusi, I. T., Agbo, F. J., Dada, O. A., Yunusa, A. A., Aruleba, K. D., Obaido, G., ... & Oyelere, S. S. (2024). Stakeholders' Insights on Artificial Intelligence Education: Perspectives of Teachers, Students, and Policymakers. *Computers and Education Open*, 7, 100212. <https://doi.org/10.1016/j.caeo.2024.100212>
- Saqr, M., & López-Pernas, S. (2024). Why explainable AI may not be enough: Predictions and mispredictions in decision making in education. *Smart Learning Environments*, 11, Article 52. <https://doi.org/10.1186/s40561-024-00343-4>
- Sharma, A., & Naidu, V. R. (2020). A Study on emerging trends to enhance learning experience in higher education institutions. In *INTED2020 Proceedings* (pp. 7391-7396). IATED. <https://doi.org/10.21125/inted.2020.1974>
- Tan, H., & Li, X. (2025). Dropout in online education: A longitudinal multilevel analysis of elementary students' extracurricular English course engagement in China. *Behavioral Sciences*, 15(4), 483. <https://doi.org/10.3390/bs15040483>
- Vincent-Lancrin, S., & Van der Vlies, R. (2020). Trustworthy Artificial Intelligence (AI) in Education: Promises and Challenges. *OECD Education Working Papers*, (218), 0_1-17. <https://doi.org/10.1787/a6c90fa9-en>
- Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, 10, 3087. <https://doi.org/10.3389/fpsyg.2019.03087>