Communication in Physical Sciences, 2026, 13(2): 373-379

Enhancing Transparency in Educational Data Mining: Applying Explainable AI to Analyze Student Behavior and Learning Patterns

Ugwu Felix Chinedu*, Aimola Amos Ayodele, Rita Mizilafe Uwumagbe and Badams Sanni Latifat

Received: 09 January 2026/Accepted: 14 March 2026 /Published: 20 March 2026

https://dx.doi.org/10.4314/cps.v13i3.3

Abstract: *This study investigates the application of Explainable Artificial Intelligence (XAI) in Educational Data Mining (EDM) to analyze student behavior and learning patterns in a transparent and accountable manner. The primary objective is to demonstrate how XAI improves the interpretability of machine learning models while preserving predictive performance, thereby supporting informed and equitable decision-making in education. A mixed-methods approach was adopted using two large-scale educational datasets: the Open University Learning Analytics Dataset (OULAD), comprising 32,593 students, and the EdNet dataset, containing over 784,000 learners and more than 131 million interaction records. These datasets include student demographics, assessment outcomes, and detailed virtual learning environment (VLE) interaction logs. Both interpretable models (Decision Trees and Logistic Regression) and black-box models (Random Forest and XGBoost) were developed to predict student performance and engagement. Black-box models achieved the highest predictive performance, with XGBoost reaching an accuracy of 0.85 on OULAD and 0.88 on EdNet, compared to 0.74–0.76 for interpretable models. To address the interpretability gap, SHAP and LIME were applied to generate global and local explanations of model predictions. SHAP analysis identified VLE access frequency, cumulative assessment scores, and early engagement indicators as the most influential predictors of academic success. LIME provided case-level explanations that highlighted factors such as low session time and poor early assessment performance in high-risk student predictions. A pilot evaluation involving 10 educators indicated that 80% found SHAP visualizations highly informative for understanding global learning patterns, while 70% rated LIME explanations as helpful for individual student diagnosis. The findings demonstrate that XAI enhances model transparency without sacrificing performance and enables educators to make data-driven, fair, and personalized instructional decisions. The study concludes that integrating XAI into EDM systems strengthens trust, accountability, and educational effectiveness, and recommends broader adoption of explainable learning analytics in educational institutions.*

Keywords: *Explainable AI, Educational Data Mining, Learning Analytics, Model Interpretability, Student Performance Prediction*

Ugwu Felix Chinedu.
Department of Computer Science, Federal College of Education, Okene, Kogi State, Nigeria.
Email: fugwu66@yahoo.com

Aimola Amos Ayodele.
Department of Computer Science, Federal College of Education, Okene, Kogi State, Nigeria.
Email: aimola4jesus@yahoo.com

Rita Mizilafe Uwumagbe
Department of Computer Science, Federal College of Education, Okene, Kogi State, Nigeria.
Email: ritauwumagbe@gmail.com

Badams Sanni Latifat
Department of Computer Science, Federal College of Education, Okene, Kogi State, Nigeria.
Email: sanlatifat@gmail.com

1.0 Introduction

The rapid evolution of digital technologies has significantly transformed the global education landscape. The increasing adoption of online learning platforms, computer-based assessments, and educational applications has resulted in the accumulation of vast and complex student data across virtual learning environments. These datasets capture critical indicators such as student engagement, academic performance, behavioral trajectories, and cognitive patterns (Romero and Ventura, 2010). These data-rich environments have created opportunities for data-driven insights into how students learn, interact, and progress within digital education systems. In response, Educational Data Mining (EDM) has emerged as a vital interdisciplinary field, employing machine learning and statistical techniques to analyze these educational datasets for actionable insights (Baker and Yacef, 2009).

By leveraging predictive and descriptive models, EDM enables the early identification of at-risk students, supports adaptive learning interventions, and facilitates evidence-based decision-making for curriculum design and instructional improvement. (Alharthi et al., 2021). The integration of Artificial Intelligence (AI) into EDM has further enhanced its capabilities, with machine learning algorithms now routinely used to forecast student performance, monitor engagement, and model complex learning pathways (Xu et al., 2021).

Despite these advancements, the increasing reliance on black-box AI models—such as deep neural networks and ensemble methods—has raised critical concerns regarding transparency, fairness, and accountability in educational contexts. These models often yield high predictive accuracy but lack interpretability, limiting educators' and stakeholders' ability to understand, interpret, or justify AI-driven decisions (Samek et al., 2017). This opacity is particularly problematic in education, where decisions affect student outcomes, equity, and resource distribution (Kizilcec et al., 2022).

The problem is further amplified by ethical considerations unique to education. Unlike other domains, educational decisions must account for diverse learner contexts, socio-economic disparities, and long-term developmental consequences. Without transparent models, AI systems risk reinforcing existing biases and misinforming interventions. This may ultimately erode stakeholder trust, particularly among marginalized student groups whose learning contexts are often underrepresented in training data. (UNESCO, 2022; Aladi and Berrada, 2018).

In response to these concerns, Explainable Artificial Intelligence (XAI) has emerged as a promising approach to bridge the interpretability gap. XAI comprises a suite of methodologies designed to make AI models more interpretable and understandable to human users without significantly compromising predictive performance. (Lundberg and lee, 2017). Prominent XAI tools, such as SHAP (Lundberg & Lee, 2017) and LIME, offer post-hoc explanations that illustrate the contribution of input features to model outputs (Ali & Mohamed, 2022).

Recent studies highlight the potential of XAI to enhance transparency in high-stakes decision-making across sectors, including finance, healthcare, and increasingly, education (Xu et al., 2023; Raju et al., 2022). However, empirical studies that apply XAI techniques to large-scale, real-world educational datasets for the purpose of understanding student learning behaviors and engagement patterns remain scarce. (UNESCO, 2022; Holstein et al., 2020).

This study investigates how XAI techniques can improve the interpretability and transparency of machine learning models used in educational data mining. Using real-world datasets—Open University Learning Analytics Dataset (OULAD) and EdNet—this research evaluates the effectiveness of

SHAP and LIME in providing educators with interpretable insights into student engagement and performance. By comparing black-box and interpretable models (e.g., decision trees and ensemble methods), the study also assesses the trade-offs between accuracy and interpretability. In doing so, the study provides empirical evidence on how explainability methods can translate predictive analytics into actionable educational insights.

Ultimately, this research addresses the pressing need for responsible and ethical AI integration in education by operationalizing explainability within real educational data contexts. The findings are expected to contribute to the design of AI systems that are not only technically robust but also transparent, trustworthy, and aligned with the broader goals of inclusive and sustainable education. This work, therefore contributes to the development of transparent learning analytics systems that align technological advancement with educational equity and accountability.

2.0 Materials and Methods

This study employed a mixed-methods research design integrating quantitative data mining techniques with qualitative evaluation to assess the practical usefulness of Explainable Artificial Intelligence (XAI) in understanding student behavior and learning patterns. The methodological framework consisted of data selection, preprocessing, predictive model development, application of explainability techniques, and performance and interpretability evaluation.

2.1 Data Collection and Selection

Two large-scale, publicly available educational datasets were utilized to ensure diversity, scale, and relevance in representing both behavioral and academic learner characteristics. The Open University Learning Analytics Dataset (OULAD) contains records from 32,593 students enrolled in distance learning modules at the Open University in the United Kingdom. It includes demographic information, assessment outcomes, and detailed logs of student interactions within the Virtual Learning Environment (Kuzilek et al., 2017). The EdNet dataset, developed by the Korea Advanced Institute of Science and Technology (KAIST), is one of the largest open online learning datasets, comprising more than 131 million interaction records from 784,309 learners. It provides timestamped clickstream logs, item response data, and topic engagement information from users of the Santa online tutoring platform (Choi et al., 2020). These datasets were selected because they capture rich, fine-grained learning behaviors across different digital learning contexts.

2.2 Data Preprocessing

Several preprocessing procedures were conducted to ensure data quality and suitability for machine learning analysis. Initially, data cleaning was performed to remove missing, null, and inconsistent records. Feature engineering techniques were then applied to derive meaningful behavioral indicators from raw interaction logs, including average session duration, frequency of platform access, and time-on-task measures. Target labels representing student performance and engagement levels, such as pass or fail outcomes and high or low engagement categories, were defined for supervised learning tasks. Finally, numerical variables were normalized to maintain scale consistency and enable fair comparison among different machine learning algorithms**.**

2.3 Model Development

Two categories of predictive models were constructed to examine differences between inherently interpretable approaches and more complex black-box techniques. Interpretable models included Decision Trees and Logistic Regression, selected for their transparency and established use in educational prediction studies. In contrast, Random Forest and Extreme Gradient Boosting (XGBoost) models were implemented to leverage their strong predictive performance despite their lower interpretability (Papamitsiou &Economides, 2020). The datasets were divided into training and testing subsets using an 80:20 split. To enhance robustness and minimize sampling bias, stratified five-fold cross-validation was applied during model training.

2.4 Application of Explainability Techniques

To enhance model transparency, post-hoc explainability methods were applied to the trained predictive models. SHAP (SHapley Additive Explanations) was used to quantify the contribution of each feature to individual predictions as well as to provide global feature importance measures (Lundberg and Lee, 2017). LIME (Local Interpretable Model-agnostic Explanations) was also employed to generate local surrogate models that explain specific predictions made by complex algorithms (Samek et al., 2017). The resulting explanations were presented in both visual and tabular formats to help educators and researchers interpret the influence of key variables, such as session frequency, assessment performance, and levels of content engagement.

2.5 Evaluation Metrics

Model effectiveness and explainability were evaluated using both quantitative and qualitative measures. Predictive performance was assessed using accuracy, precision, recall, and F1-score. Global interpretability was examined through mean SHAP value rankings to determine the most influential features across predictions. In addition, a qualitative educator interpretability score was derived from feedback provided by educators who assessed the clarity, usefulness, and practical relevance of the generated explanations. Comparative analyses were conducted to examine the trade-offs between predictive accuracy and interpretability across the different modeling and explanation approaches.

3.0 Results and Discussion

3.1 Model Performance Comparison

The performance of both interpretable and black-box models was evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. Table I summarizes the performance of the models on the OULAD and EdNet datasets

Table I: Model Performance Comparison

Model	Dataset	Accuracy	Precision	Recall	F1-Score
Decision Tree	OULAD	0.74	0.71	0.70	0.70
Logistic Regression	OULAD	0.76	0.73	0.72	0.72
Random Forest	OULAD	0.83	0.80	0.78	0.79
XGBoost	OULAD	0.85	0.82	0.80	0.81
Random Forest	EdNet	0.86	0.83	0.81	0.82
XGBoost	EdNet	0.88	0.85	0.83	0.84

The results show that black-box models (XGBoost and Random Forest) consistently achieved higher predictive performance than inherently interpretable models across both datasets. This finding reflects the common trade-off between predictive power and model interpretability, underscoring the importance of explainability techniques when deploying complex models in educational contexts. However, these models are inherently less interpretable without the aid of explanation techniques.

3.2 Insights from SHAP and LIME

SHAP and LIME were applied to the black-box models to provide both global and local explanations. SHAP analysis revealed that frequency of VLE access, cumulative assessment scores, and early engagement indicators were the most influential predictors of student success across both datasets. These global patterns helped identify the key behavioral factors associated with learning outcomes. LIME, in contrast, provided local interpretability by explaining individual predictions. For example, for a high-risk student, low session duration and poor early assessment performance were identified as the primary contributors to a predicted failure outcome. SHAP visualizations were particularly useful for global interpretability, enabling educators to easily recognize the most impactful learning behaviors.

3.3 Educator Interpretability Feedback

The results indicated that 80% of participants considered SHAP summaries highly informative for understanding global student behavior patterns, while 70% rated LIME explanations as clear and helpful for diagnosing individual student cases. Despite these positive evaluations, several educators emphasized the need for more user-friendly visualizations and clearer contextual explanations to better support non-technical users. These findings align with Raju et al. (2022), who stress the importance of adapting XAI explanations to educators' cognitive needs.

3.4 Practical Implications

The integration of XAI into educational data mining offers several important benefits. It enhances trust in AI-driven recommendations by making model decisions transparent and auditable. It also enables educators to justify targeted interventions, particularly for students identified as being at risk of disengagement or failure. Furthermore, explainability supports fairness by revealing how specific features may disproportionately influence outcomes for certain student groups.

Despite these advantages, challenges remain. The successful implementation of XAI in real-world educational settings requires professional development for educators, integration with existing educational systems, and continued research into explanation preferences among different stakeholder groups (Holstein et al., 2020).

4.0 Conclusion

This study explored the role of Explainable Artificial Intelligence (XAI) in improving the transparency and practical usability of machine learning models within Educational Data Mining. By applying both inherently interpretable models and high-performing black-box models to the OULAD and EdNet datasets, the research examined how predictive accuracy can be balanced with the need for interpretability in educational decision-making.

The results showed that although black-box approaches such as XGBoost and Random Forest achieved superior predictive performance, their lack of transparency limits their direct applicability in educational contexts where accountability and fairness are essential. The integration of SHAP and LIME effectively addressed this limitation by providing meaningful global and local explanations of model predictions. These explanations revealed key behavioral and academic indicators influencing student outcomes and demonstrated how complex analytics can be translated into actionable insights.

Feedback from educators further confirmed the practical value of XAI-supported analytics. Participants reported that explainable outputs enhanced their understanding of student learning patterns, supported early identification of at-risk learners, and informed targeted instructional interventions. At the same time, the study highlighted the need for improved visualization design and professional development to ensure that non-technical users can confidently interpret and apply XAI insights.

Overall, this work underscores the importance of combining predictive performance with interpretability when deploying AI in education. By demonstrating how XAI can foster transparency, trust, and fairness, the study contributes to the development of responsible learning analytics systems that support both effective teaching and equitable student outcomes.

Acknowledgement

The authors appreciate the Open University UK and KAIST for providing publicly available datasets and acknowledge all

educators who contributed to the evaluation process.

5.0 References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access,* 6, pp. 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Ali, A. F., & Mohamed, A. F. (2022). Explainable machine learning for predicting students' academic performance. *Education and Information Technologies*, 27, pp. 10707–10729. https://doi.org/10.1007/s10639-022-10923-6

Alharthi, F. A., Alghamdi, A. N., & Algosaibi, R. M. (2021). Predicting students' academic performance using machine learning: A case study. *Computers in Human Behavior Reports*, 4, p. 100130. https://doi.org/10.1016/j.chbr.2021.100130

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining,* 1, 1, pp. 3–17.

Choi, Y., Shin, M., Oh, S., & Oh, J. (2020). EdNet: A large-scale hierarchical dataset in education. *Proceedings of the International Conference on Educational Data Mining (EDM),* pp. 1–8.

Holstein, E., McLaren, J., Aleven, V., & Yarzebinski, B. (2020). Designing for pedagogical trust: Supporting instructor adoption of AI tools. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems,* pp. 1–14. https://doi.org/10.1145/3313831.3376313

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1, pp. 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kizilcec, H., & Lee, G. (2022). Algorithmic fairness in education. *International Journal of Artificial Intelligence in Education*, 32, 3, pp. 682–707. https://doi.org/10.1007/s40593-022-00315-7

Kuzilek, A., Hlosta, M., & Zdrahal, Z. (2017). Open University Learning Analytics Dataset. *Scientific Data,* 4, pp. 170171. https://doi.org/10.1038/sdata.2017.171

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS),* pp. 4765–4774.

Papamitsiou, K., & Economides, A. A. (2020). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educational Technology & Society,* 17, 4, pp. 49–64.

Raju, S. A., Hughes, R. T., & Nkambou, R. (2022). Explainable AI in education: Challenges, opportunities, and research directions. *Proceedings of the IEEE Global Engineering Education Conference (EDUCON),* pp. 948–953. https://doi.org/10.1109/EDUCON52537.2022.9766700

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 1135–1144. https://doi.org/10.1145/2939672.2939778

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* 40, 6, pp. 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models.

arXiv preprint arXiv:1708.08296, pp. 1–8.

Sweeney, P., Hassan, M. H., Ojo, T., & Alzubaidi, L. (2022). Student dropout prediction in higher education using ensemble learning. *Computers & Education: Artificial Intelligence,* 3, p. 100049. https://doi.org/10.1016/j.caeai.2022.100040

UNESCO. (2022). *Artificial Intelligence and Education: Guidance for Policy-Makers*. United Nations Educational, Scientific and Cultural Organization, Paris.

UNESCO. (2022). *Ethics of Artificial Intelligence in Education: Promises and Challenges*. United Nations Educational, Scientific and Cultural Organization, Paris.

Xu, D., Moon, H., & Baek, J. (2021). An interpretable student performance prediction model with decision trees. *Journal of Educational Computing Research*, 59, 1, pp. 199–219. https://doi.org/10.1177/0735633120972605

Xu, Y., Yan, C., & Zhang, Y. (2023). Interpretable machine learning for online learning analytics: A case study using LIME and SHAP. *Journal of Computer Assisted Learning,* 10, pp. 423–442. https://doi.org/10.1111/jcal.12672

Declaration

Consent for publication

Not Applicable

Availability of data and materials

The publisher has the right to make the data public

Conflict of Interest

The authors declared no conflict of interest

Ethical Considerations

Not applicable

Competing interest

The authors report no conflict or competing interest

Funding

The author declared no source of funding

Author Contributions

Ugwu, Felix Chinedu conceptualized the study, supervised the research process, and contributed to the design of the methodology and interpretation of results. Aimola, Amos Ayodele was responsible for data preprocessing, model development, and implementation of the machine learning experiments. Rita, Mizilafe Uwumagbe conducted the application of explainable AI techniques (SHAP and LIME), generated visualizations, and contributed to the analysis of interpretability outcomes. Badams Sanni Latifat coordinated the educator feedback study, performed the qualitative analysis, and assisted in drafting and editing the manuscript. All authors reviewed, revised, and approved the final version of the manuscript.