Demystifying Student Learning Behaviors: An Explainable AI Perspective in Educational Analytics

[1] Zen Tiger, [2] James Smith

[1] University of Oxford, Oxford, UK, zen126745@gmail.com

[2] University of Edinburgh, Scotland, UK, jamessmith126745@gmail.com

Abstract

Understanding student learning behaviors is crucial for improving educational outcomes, enabling personalized instruction, and fostering effective learning environments. This study presents an approach to decode and analyze student behaviors using Explainable Artificial Intelligence (XAI) methods integrated into educational analytics systems. Traditional black-box machine learning models often provide high predictive accuracy but lack transparency, making it difficult for educators to trust and act upon the insights generated. To address this challenge, we propose a framework that employs explainable models, including SHAP, LIME, and attention-based neural networks, to identify key cognitive and behavioral indicators influencing student performance. Using real-world educational datasets, this research demonstrates how XAI enhances interpretability while maintaining competitive prediction accuracy. The findings indicate that explainable models not only uncover hidden patterns in student learning processes but also empower educators with actionable insights to design personalized interventions and adaptive learning pathways.

Keywords: Explainable AI, Educational Analytics, Student Learning Behaviors, XAI Models, Cognitive Skill Prediction, SHAP, LIME, Attention Networks.

Introduction

The field of educational analytics has witnessed a significant transformation with the integration of artificial intelligence (AI) and machine learning (ML) techniques. These technologies have provided institutions with the capability to analyze vast amounts of learning data to predict student performance, identify at-risk learners, and recommend personalized learning interventions. However, most of these AI-based solutions have traditionally relied on complex, opaque models often referred to as black-box systems. While these models can achieve high levels of predictive accuracy, their lack of transparency raises concerns among educators, policymakers, and learners. Without understanding the reasoning behind predictions, stakeholders struggle to trust the outcomes or utilize them effectively for decision-making[1].

Explainable Artificial Intelligence (XAI) has emerged as a promising solution to bridge this gap by enabling transparency, interpretability, and accountability in AI-driven systems. XAI provides insights into how algorithms reach their predictions and decisions, making them comprehensible to non-technical users such as educators[2]. In the context of educational analytics, explainability becomes even more crucial because decisions directly influence students' academic trajectories, resource allocation, and learning strategies. For example, identifying why a student is predicted to underperform can help instructors implement targeted interventions rather than applying generalized solutions[3, 4].

This study focuses on demystifying student learning behaviors through the integration of XAI techniques within predictive models used in educational analytics. Student learning behaviors are multifaceted, encompassing cognitive engagement, interaction patterns with digital platforms, participation in collaborative tasks, and self-regulated learning activities. By leveraging explainable models, we can uncover the factors that most significantly contribute to performance variations, thus creating opportunities for personalized and equitable education[5, 6].

The research employs a combination of traditional machine learning classifiers and modern explainability tools, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), alongside attention-based deep learning models[7]. The objective is to provide not only accurate predictions but also human-understandable explanations for these predictions. Real-world datasets from online learning management systems (LMS) and digital classrooms were analyzed to validate the proposed framework. The outcomes demonstrate how XAI enhances the interpretability of cognitive skill prediction while maintaining performance levels comparable to black-box models[8, 9].

Ultimately, this study aims to shift the paradigm from opaque predictions to transparent and actionable insights in educational analytics. By decoding the underlying mechanisms of student learning behaviors, educators can foster more inclusive and effective learning environments, aligning with the broader goals of data-driven education and ethical AI adoption in academia[10, 11].

Explainable AI in Educational Analytics

Explainable AI serves as a crucial element in transforming how educational institutions utilize learning analytics. Traditional analytics pipelines typically involve feature extraction from students' interactions with learning platforms, followed by training predictive models that forecast academic performance, dropout likelihood, or knowledge mastery[12]. However, when educators receive only the output—such as a risk score or predicted grade—without knowing which features drove the decision, the actionable value of such predictions remains limited[13, 14].

XAI techniques address this limitation by elucidating the inner workings of machine learning algorithms. Among the most widely used techniques are SHAP and LIME. SHAP values quantify the contribution of each input feature to a model's output, allowing educators to understand whether study time, assignment submission patterns, forum interactions, or quiz attempts are the primary drivers of a particular prediction[15]. LIME, on the other hand, creates interpretable local approximations of complex models, revealing the decision logic for individual students rather than entire datasets. Such individualized insights are highly valuable in educational settings where each learner's pathway is unique[16, 17].

In addition to these post-hoc explanation methods, attention mechanisms within deep learning architectures have gained traction in education-focused XAI research. Attention-based neural networks dynamically highlight the most important elements within sequential learning data, such as clickstream logs or temporal engagement metrics, thereby offering an intrinsic form of interpretability. This is particularly relevant in environments where students' learning behaviors evolve over time[18, 19].

The benefits of integrating XAI into educational analytics extend beyond interpretability. They promote fairness by exposing potential biases within predictive systems, such as over-reliance on demographic variables. They also foster trust among stakeholders, as educators and students alike can scrutinize the reasoning behind algorithmic outputs[20]. For administrators and policymakers, XAI can guide the formulation of evidence-based interventions and resource allocation strategies, ultimately contributing to more effective learning ecosystems[6, 21, 22].

This section underscores that while predictive accuracy remains important, explainability has emerged as an equally critical metric in the deployment of AI for education. By providing a window into the decision-making process, XAI ensures that technological advancements are not only powerful but also ethically aligned with educational values[14, 23].

Decoding Student Learning Behaviors with XAI

Student learning behavior is a complex construct influenced by cognitive, emotional, and behavioral factors. Decoding these behaviors using explainable AI allows educators to move from generalized interventions to precise, data-driven strategies tailored to individual learners[24, 25]. In this study, learning behaviors were analyzed across multiple dimensions, including engagement frequency, time-on-task, assessment performance, interaction with peers, and adaptability to feedback. These variables were extracted from diverse educational datasets, including online courses, blended learning environments, and traditional classrooms enhanced with digital tracking systems[26, 27].

The application of SHAP values revealed several key patterns. For instance, timely submission of assignments and consistent participation in formative assessments were found to be among the strongest predictors of high academic achievement. Conversely, irregular login frequencies and abrupt drops in engagement often signaled potential learning difficulties. LIME explanations provided localized, student-specific interpretations, enabling educators to pinpoint the exact factors leading to a prediction of risk or success[6, 28, 29].

Attention-based deep learning models further enriched the analysis by dynamically identifying how students' behaviors changed over time. For example, a student initially disengaged might show improved performance after receiving targeted feedback, a transition captured effectively by the attention weights. This temporal interpretability is particularly valuable for designing adaptive learning interventions that evolve with students' needs[30, 31].

Moreover, the insights derived from explainable models can be fed back into instructional design processes. Educators can restructure course materials to emphasize high-impact behaviors, while learning platforms can integrate adaptive triggers, such as personalized reminders or supplementary resources, based on explainable risk predictions[32, 33].

By demystifying the learning process, XAI transforms raw educational data into actionable intelligence. It empowers teachers to become proactive facilitators rather than reactive observers and promotes a culture of transparency in data-driven education. Importantly, the study also highlights the ethical dimension of explainable learning analytics—students are more likely to accept algorithmic recommendations when they understand the rationale behind them[34, 35].

Conclusion

This research demonstrates that explainable AI provides a powerful lens for understanding and enhancing student learning behaviors in educational analytics. By combining post-hoc explanation techniques such as SHAP and LIME with inherently interpretable models like attention-based networks, the study bridges the gap between predictive power and actionable insights. The findings confirm that XAI not only predicts student performance effectively but also reveals the factors shaping these outcomes, enabling educators to design personalized, fair, and transparent learning interventions. Moving forward, integrating explainability as a standard practice in educational AI systems can transform opaque data pipelines into meaningful tools for improving teaching and learning processes.

References:

[1] A. Abulibdeh, E. Zaidan, and R. Abulibdeh, "Navigating the confluence of artificial intelligence and education for sustainable development in the era of industry 4.0: Challenges, opportunities, and ethical dimensions," *Journal of Cleaner Production,* p. 140527, 2024.

[2] J. Shao, J. Dong, D. Wang, K. Shih, D. Li, and C. Zhou, "Deep Learning Model Acceleration and Optimization Strategies for Real-Time Recommendation Systems," *arXiv preprint arXiv:2506.11421,* 2025.

[3] H. Allam, J. Dempere, V. Akre, D. Parakash, N. Mazher, and J. Ahamed, "Artificial intelligence in education: an argument of Chat-GPT use in education," in *2023 9th International Conference on Information Technology Trends (ITT)*, 2023: IEEE, pp. 151-156.

[4] T. Niu, T. Liu, Y. T. Luo, P. C.-I. Pang, S. Huang, and A. Xiang, "Decoding student cognitive abilities: a comparative study of explainable AI algorithms in educational data mining," *Scientific Reports,* vol. 15, no. 1, p. 26862, 2025.

[5] L. E. Alvarez-Dionisi, M. Mittra, and R. Balza, "Teaching artificial intelligence and robotics to undergraduate systems engineering students," *International Journal of Modern Education and Computer Science,* vol. 11, no. 7, pp. 54-63, 2019.

[6] J. Shen, W. Wu, and Q. Xu, "Accurate prediction of temperature indicators in eastern china using a multi-scale cnn-lstm-attention model," *arXiv preprint arXiv:2412.07997,* 2024.

[7] Y. Zhao, H. Shen, D. Li, L. Chang, C. Zhou, and Y. Yang, "Meta-Learning for Cold-Start Personalization in Prompt-Tuned LLMs," *arXiv preprint arXiv:2507.16672,* 2025.

[8] C. Becker, G. Lauterbach, S. Spengler, U. Dettweiler, and F. Mess, "Effects of regular classes in outdoor education settings: A systematic review on students' learning, social and health dimensions," *International journal of environmental research and public health,* vol. 14, no. 5, p. 485, 2017.

[9] S. Diao, C. Wei, J. Wang, and Y. Li, "Ventilator pressure prediction using recurrent neural network," *arXiv preprint arXiv:2410.06552,* 2024.

[10] T. Buser, M. Niederle, and H. Oosterbeek, "Can competitiveness predict education and labor market outcomes? Evidence from incentivized choice and survey measures," *Review of Economics and Statistics,* pp. 1-45, 2024.

[11] X. Shi, Y. Tao, and S.-C. Lin, "Deep neural network-based prediction of B-cell epitopes for SARS-CoV and SARS-CoV-2: Enhancing vaccine design through machine learning," in *2024 4th International Signal Processing, Communications and Engineering Management Conference (ISPCEM)*, 2024: IEEE, pp. 259-263.

[12] H. Yang, L. Wang, J. Zhang, Y. Cheng, and A. Xiang, "Research on edge detection of LiDAR images based on artificial intelligence technology," *arXiv preprint arXiv:2406.09773,* 2024.

[13] S. S. Gill *et al.*, "Transformative effects of ChatGPT on modern education: Emerging Era of AI Chatbots," *Internet of Things and Cyber-Physical Systems,* vol. 4, pp. 19-23, 2024.

[14] K. Mo *et al.*, "Dral: Deep reinforcement adaptive learning for multi-uavs navigation in unknown indoor environment," *arXiv preprint arXiv:2409.03930,* 2024.

[15] Y. Zhao, H. Lyu, Y. Peng, A. Sun, F. Jiang, and X. Han, "Research on Low-Latency Inference and Training Efficiency Optimization for Graph Neural Network and Large Language Model-Based Recommendation Systems," *arXiv preprint arXiv:2507.01035,* 2025.

[16] E. Kasneci *et al.*, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and individual differences,* vol. 103, p. 102274, 2023.

[17] X. Han, "Optimizing Cloud Computing Energy Consumption Prediction Using Convolutional Neural Networks with Bidirectional Gated Cycle Unit," in *2025 4th International Symposium on Computer Applications and Information Technology (ISCAIT)*, 2025: IEEE, pp. 173-177.

[18] M. Khan, "Ethics of Assessment in Higher Education–an Analysis of AI and Contemporary Teaching," EasyChair, 2516-2314, 2023.

[19] K. Shih, Y. Han, and L. Tan, "Recommendation system in advertising and streaming media: Unsupervised data enhancement sequence suggestions," *arXiv preprint arXiv:2504.08740,* 2025.

[20] H. Yang, Z. Cheng, Z. Zhang, Y. Luo, S. Huang, and A. Xiang, "Analysis of Financial Risk Behavior Prediction Using Deep Learning and Big Data Algorithms," *arXiv preprint arXiv:2410.19394,* 2024.

[21] F. Ni, H. Zang, and Y. Qiao, "Smartfix: Leveraging machine learning for proactive equipment maintenance in industry 4.0," in *The 2nd International scientific and practical conference "Innovations in education: prospects and challenges of today"(January 16-19, 2024) Sofia, Bulgaria. International Science Group. 2024. 389 p.*, 2024, p. 313.

[22] H. Yang, Z. Shen, J. Shao, L. Men, X. Han, and J. Dong, "LLM-Augmented Symptom Analysis for Cardiovascular Disease Risk Prediction: A Clinical NLP," *arXiv preprint arXiv:2507.11052,* 2025.

[23] L. Yan *et al.*, "Practical and ethical challenges of large language models in education: A systematic scoping review," *British Journal of Educational Technology,* vol. 55, no. 1, pp. 90-112, 2024.

[24] L. Min, Q. Yu, Y. Zhang, K. Zhang, and Y. Hu, "Financial prediction using DeepFM: Loan repayment with attention and hybrid loss," in *2024 5th International Conference on Machine Learning and Computer Application (ICMLCA)*, 2024: IEEE, pp. 440-443.

[25] Y. Zhao, Y. Peng, L. Zhang, Q. Sun, Z. Zhang, and Y. Zhuang, "Multimodal Foundation Model-Driven User Interest Modeling and Behavior Analysis on Short Video Platforms," *arXiv preprint arXiv:2509.04751,* 2025.

[26] T. Shehzadi, A. Safer, and S. Hussain, "A Comprehensive Survey on Artificial Intelligence in sustainable education," *Authorea Preprints,* 2022.

[27] H. Yang, H. Lyu, T. Zhang, D. Wang, and Y. Zhao, "LLM-Driven E-Commerce Marketing Content Optimization: Balancing Creativity and Conversion," *arXiv preprint arXiv:2505.23809,* 2025.

[28] J. Weinberg *et al.*, "A multidisciplinary model for using robotics in engineering education," in *2001 Annual Conference*, 2001, pp. 6.59. 1-6.59. 9.

[29] H. Lyu, J. Dong, Y. Tian, D. Wang, L. Men, and Z. Zhang, "Self-Supervised User Embedding Alignment for Cross-Domain Recommendations via Multi-LLM Co-Training," *Authorea Preprints,* 2025.

[30] L. W. Anderson, "Objectives, evaluation, and the improvement of education," *Studies in educational evaluation,* vol. 31, no. 2-3, pp. 102-113, 2005.

[31] Z. Yang, A. Sun, Y. Zhao, Y. Yang, D. Li, and C. Zhou, "RLHF Fine-Tuning of LLMs for Alignment with Implicit User Feedback in Conversational Recommenders," *arXiv preprint arXiv:2508.05289,* 2025.

[32] M. Byram, *From foreign language education to education for intercultural citizenship: Essays and reflections*. Multilingual Matters, 2008.

[33] X. Lin, Y. Tu, Q. Lu, J. Cao, and H. Yang, "Research on Content Detection Algorithms and Bypass Mechanisms for Large Language Models," *Academic Journal of Compufing & Informafion Science,* vol. 8, no. 1, pp. 48-56, 2025.

[34] M. Z. Samsuri, S. A. Ariffin, and N. S. Fathil, "Incorporating cultural design elements in mobile applications creative industries in Malaysia: A conceptual study," *Journal of ICT in Education,* vol. 8, no. 2, pp. 110-117, 2021.

[35] H. Guo, Y. Zhang, L. Chen, and A. A. Khan, "Research on vehicle detection based on improved YOLOv8 network," *arXiv preprint arXiv:2501.00300,* 2024.