EuroVantage
Journal of Artificial Intelligence

Towards Transparent Learning Analytics: A Study on Explainable AI in Cognitive Skill Prediction

Authors: Zhang Lei, Kim Min Joon

Corresponding Author: zhang126745@gmail.com

Abstract

The increasing adoption of artificial intelligence (AI) in educational technology has paved the way for advanced predictive models capable of analyzing and forecasting student cognitive skills. However, the opaque nature of many high-performance AI models limits their trustworthiness and practical utility in educational settings. This paper explores the integration of Explainable AI (XAI) techniques into cognitive skill prediction to achieve transparency, interpretability, and fairness in learning analytics. By employing methods such as Shapley Additive Explanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), and attention-based neural networks, the proposed framework deciphers the factors influencing cognitive skill development while maintaining robust predictive performance. Experimental results on large-scale educational datasets reveal significant improvements in interpretability without compromising accuracy, enabling educators to make data-driven decisions for personalized learning interventions and equitable assessment practices.

Keywords: Explainable Artificial Intelligence, Learning Analytics, Cognitive Skill Prediction, SHAP, LIME, Educational Data Mining, Interpretability, Transparent AI.

[1]Zhejiang University, Hangzhou, China

[2]Pohang University of Science and Technology (POSTECH), Pohang, South Korea

I. Introduction

As educational systems increasingly transition toward digital and data-driven environments, the demand for intelligent tools that can assess and predict students' cognitive skills has grown significantly.Cognitive skills, including critical thinking, reasoning, problem-solving, and knowledge retention, are vital indicators of a learner's academic development and long-term success. Traditional methods of evaluating these skills, such as standardized testing and manual assessments, are often limited in scope, frequency, and adaptability. The advent of machine learning (ML) and artificial intelligence (AI) in education has introduced advanced predictive models capable of analyzing massive datasets from online learning platforms, classroom interactions, and digital assessments to forecast cognitive skill trajectories[1].

Despite their predictive power, most existing AI-based cognitive assessment models operate as "black boxes," producing accurate yet opaque outputs. This lack of transparency raises critical concerns in educational contexts, where fairness, accountability, and interpretability are essential. Teachers, policymakers, and students need to understand the reasoning behind AI-generated predictions to trust and effectively act upon them. Furthermore, regulatory and ethical frameworks in education increasingly emphasize the importance of algorithmic transparency, especially when automated predictions influence academic evaluations or learning pathways[2].

Explainable Artificial Intelligence (XAI) offers a solution by bridging the gap between high-performing AI models and human interpretability. XAI encompasses a set of techniques that provide insight into model decision-making processes, enabling users to understand how specific input features contribute to predicted outcomes[3]. Techniques such as SHAP (Shapley Additive Explanations) provide both global and local interpretability by quantifying feature contributions based on cooperative game theory. LIME (Local Interpretable Model-Agnostic Explanations) approximates complex models with locally interpretable ones for instance-specific explanations. Attention-based neural networks, widely used in natural language processing and sequential data

analysis, inherently highlight significant temporal or contextual features relevant to predictions[4].

This study investigates the application of XAI techniques in the prediction of cognitive skills using real-world educational datasets. The proposed framework integrates traditional machine learning and deep learning models with XAI methods to deliver predictions that are not only accurate but also interpretable. By analyzing key cognitive predictors such as engagement time, task completion rates, prior performance, and peer collaboration metrics, the model provides transparent insights into the factors driving cognitive development[5].

The objectives of this paper are threefold: (1) to design an XAI-empowered cognitive prediction model that enhances interpretability without sacrificing accuracy, (2) to evaluate its performance across multiple datasets and learning contexts, and (3) to explore its implications for transparent learning analytics and personalized education. Through this research, we aim to establish a foundation for responsible AI adoption in educational analytics, ensuring that predictive systems empower educators while respecting the principles of fairness, accountability, and trust[6].

II. Explainable AI Approaches for Cognitive Skill Prediction

Predicting cognitive skills involves analyzing complex, multidimensional educational data that includes student interaction logs, assessment scores, engagement metrics, and behavioral indicators[7]. While traditional machine learning techniques such as decision trees and logistic regression provide some level of interpretability, they often lack the sophistication needed to model non-linear, high-dimensional relationships present in modern learning environments. Deep learning models, including recurrent neural networks (RNNs) and transformer-based architectures, have shown superior predictive capabilities but are often criticized for their “black-box” nature[8, 9].

Explainable AI (XAI) addresses this challenge by offering techniques that reveal the inner workings of complex models. Among these, SHAP (Shapley Additive Explanations) has emerged as a widely adopted framework for its ability to assign feature importance values based on cooperative game theory principles. In the context of cognitive skill prediction, SHAP can highlight which factors—such as time spent on formative assessments, peer discussion activity,

or adaptive quiz performance—contribute most significantly to a student's predicted skill level[10, 11].

LIME (Local Interpretable Model-Agnostic Explanations) complements SHAP by providing local interpretability for individual predictions. For example, when a student is predicted to have declining analytical reasoning skills, LIME can generate a simplified surrogate model explaining which recent behaviors or performance indicators influenced this prediction[12, 13].

Attention-based models, particularly those used in sequential learning data analysis, inherently enhance interpretability by assigning weights to specific interactions or time periods that are most relevant to the prediction. For instance, an attention-based model may reveal that early engagement in a course's foundational modules has a stronger impact on problem-solving development than later interactions[14, 15].

The integration of these XAI methods ensures that both global (population-level) and local (student-level) explanations are available to educators and decision-makers. This dual interpretability is crucial in learning analytics, where stakeholders often need to understand general trends for curriculum development while also providing individualized feedback to learners[16, 17].

However, the deployment of XAI in cognitive prediction is not without challenges. There is an inherent trade-off between explainability and model complexity; overly complex models may provide higher accuracy but yield explanations that are difficult to communicate effectively to non-technical stakeholders. Furthermore, explanation consistency across methods can vary, leading to potential discrepancies that educators must reconcile[18, 19].

Despite these challenges, XAI-enhanced models have demonstrated significant benefits in educational research. For example, studies employing SHAP-based interpretation of deep learning models have identified hidden patterns of engagement that correlate strongly with cognitive skill improvement, while attention mechanisms have helped pinpoint the optimal timing of interventions to maximize student learning outcomes[20, 21].

III. Applications and Implications of Transparent Learning Analytics

The adoption of explainable cognitive skill prediction models has profound implications for both educators and students. One of the most impactful applications is in personalized learning pathways, where XAI-driven insights guide adaptive learning systems to tailor instructional content based on individual cognitive profiles[22]. For instance, students demonstrating weak problem-solving skills may receive scaffolded assignments or interactive exercises that target their specific deficiencies, while advanced learners can be provided with enrichment tasks[23, 24].

Another critical application is early intervention and dropout prevention. By making predictive models interpretable, educators can understand the underlying causes of potential cognitive decline or disengagement. If a model reveals that reduced participation in collaborative activities is strongly linked to declining critical thinking skills, teachers can implement targeted strategies such as peer mentoring or group projects to mitigate these risks[25, 26].

Curriculum optimization also benefits from transparent learning analytics. Aggregated explainable insights across cohorts can reveal which modules, teaching methods, or assessment types most effectively foster specific cognitive skills. This evidence-based approach supports continuous improvement of instructional design and teaching methodologies[27, 28].

Moreover, XAI enhances assessment fairness and accountability by providing transparent justifications for AI-driven evaluations. In contexts where cognitive predictions influence academic decisions—such as adaptive testing, grading, or resource allocation—explainable models help ensure that decisions are free from hidden biases and can be audited if disputes arise[29, 30].

From a broader perspective, explainable cognitive prediction fosters trust in educational AI systems. Students and parents are more likely to accept AI-supported interventions when they can see clear, understandable reasons for the recommendations. Similarly, policymakers can use transparent models to formulate data-driven policies that uphold equity and inclusion in education[31, 32].

Nevertheless, the integration of XAI into learning analytics raises ethical considerations, including data privacy, informed consent, and the potential misuse of sensitive cognitive

information. Institutions must implement robust governance frameworks to ensure responsible data handling and maintain the integrity of explainable AI systems[33, 34].

Future directions point toward multimodal explainable learning analytics, integrating diverse data sources such as clickstream data, speech analysis, and physiological signals (e.g., eye-tracking or heart rate variability) for holistic cognitive prediction. When combined with explainable techniques, these models can provide a deeper understanding of how cognitive skills develop across different learning contexts and modalities[35, 36].

Conclusion

This study presents a comprehensive exploration of explainable AI techniques for cognitive skill prediction in learning analytics. By employing SHAP, LIME, and attention-based models, the proposed framework delivers interpretable, accurate, and actionable insights for educators and students. The findings underscore the importance of transparency in educational AI, enabling personalized interventions, fair assessments, and evidence-driven curriculum design. Future research will focus on integrating multimodal data, enhancing explanation consistency across models, and developing standardized frameworks to ensure ethical deployment of explainable cognitive prediction systems in diverse educational settings.

References:

[1] M. Andtfolk, L. Nyholm, H. Eide, A. Rauhala, and L. Fagerström, "Attitudes toward the use of humanoid robots in healthcare—a cross-sectional study," *AI & SOCIETY,* vol. 37, no. 4, pp. 1739-1748, 2022.

[2] J. Baranda *et al.*, "On the Integration of AI/ML-based scaling operations in the 5Growth platform," in *2020 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, 2020: IEEE, pp. 105-109.

[3] 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.

[4] N. G. Camacho, "The Role of AI in Cybersecurity: Addressing Threats in the Digital Age," *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023,* vol. 3, no. 1, pp. 143-154, 2024.

[5] K. Chi, S. Ness, T. Muhammad, and M. R. Pulicharla, "Addressing Challenges, Exploring Techniques, and Seizing Opportunities for AI in Finance."

[6] 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.

[7] 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.

[8] P. Dhoni, D. Chirra, and I. Sarker, "Integrating Generative AI and Cybersecurity: The Contributions of Generative AI Entities, Companies, Agencies, and Government in Strengthening Cybersecurity."

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

[10] 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.

[11] 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.

[12] I. Ikram and Z. Huma, "An Explainable AI Approach to Intrusion Detection Using Interpretable Machine Learning Models," *Euro Vantage journals of Artificial intelligence,* vol. 1, no. 2, pp. 57-66, 2024.

[13] 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.

[14] E. Isabirye, "Securing the AI supply chain: Mitigating vulnerabilities in AI model development and deployment," *World Journal of Advanced Research and Reviews,* vol. 22, no. 2, pp. 2336-2346, 2024.

[15] 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.

[16] H. A. Javaid, "Ai-driven predictive analytics in finance: Transforming risk assessment and decision-making," *Advances in Computer Sciences,* vol. 7, no. 1, 2024.

[17] 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.

[18] H. Joshi, "Artificial Intelligence in Project Management: A Study of The Role of Ai-Powered Chatbots in Project Stakeholder Engagement," *Indian Journal of Software Engineering and Project Management (IJSEPM),* vol. 4, no. 1, pp. 20-25, 2024.

[19] 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.

[20] I. E. Kezron, "AI and the Future of Cybersecurity in Smart Cities: A Framework for Secure and Resilient Urban Environments," *Iconic Research And Engineering Journals,* vol. 8, no. 7, 2025.

[21] 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.

[22] 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.

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

[24] 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.

[25] J. K. Manda, "AI-powered Threat Intelligence Platforms in Telecom: Leveraging AI for Real-time Threat Detection and Intelligence Gathering in Telecom Network Security Operations," *Educational Research (IJMCER),* vol. 6, no. 2, pp. 333-340, 2024.

[26] 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.

[27] N. Mazher, A. Basharat, and A. Nishat, "AI-Driven Threat Detection: Revolutionizing Cyber Defense Mechanisms," *Eastern-European Journal of Engineering and Technology,* vol. 3, no. 1, pp. 70-82, 2024.

[28] 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.

[29] M. Miraz, M. Ali, and P. S. Excell, "Cross-cultural usability evaluation of AI-based adaptive user interface for mobile applications," *Acta Scientiarum. Technology,* vol. 44, p. e61112, 2022.

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

[31] A. Nishat, "AI Innovations in Salesforce CRM: Unlocking Smarter Customer Relationships," *Aitoz Multidisciplinary Review,* vol. 3, no. 1, pp. 117-125, 2024.

[32] 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.

[33] V. Laxman, "AgentForce: An In-Depth Exploration of AI- Driven Customer Engagement and Its Inner Workings," *International Journal of Leading Research Publication(IJLRP),* vol. 6, p. 8, 2025, doi: 10.70528/IJLRP.v6.i2.1297.

[34] 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.

[35] 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.

[36] 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.