International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Explainable Artificial Intelligence in Higher Education: A Systematic Literature Review of Methods, Applications, Challenges and Research Gaps

Dr. Sasikala P [1], Dr. Asha N[2], Dr. Pushpalatha M[3], Dr. Santhosh Kuamr B N[4]

[1]Associate Professor, Dept of Computer Science, Lal Bahadur Shastri Govt. First Grade College, RT Nagar, Bengaluru, India
[2]Associate Professor, Department of computer science, Nrupathunga University(formerly government science college), Bangalore, India
[3,4]Associate Professor, Department of Computer Science, Govt. Women's College Hunsur, Karnataka, India

Abstract
Artificial Intelligence (AI) and Machine Learning (ML) have been increasingly applied in higher education with aims of student-performance prediction, dropout-risk identification, learning analytics, assessment, personalization, and academic decision support. While machine learning models can often identify patterns in student data, many models and their predictions remain opaque to teachers, administrators, and students. Explainable Artificial Intelligence (XAI) has therefore become an important research area that facilitates transparency, accountability, trust, and human-level understanding of educational decisions made by ML models.
This paper provides a systematic review of research on XAI in higher education with a focus on student-performance prediction, as well as educational decision support. In particular, the review explores the application areas, predictive models, explanation methods, types of explanations, and challenges in applying XAI in higher education. The review highlights that SHAP, feature importance, rule-based models, decision trees, and local explanation techniques have been commonly applied for model interpretation in education. However, most of the prior work focuses on simply providing explanations rather than turning the insights into educational interventions. The review also reveals that fairness, individual-level explanations, visualization, personalization, and educational relevance of model explanations have received insufficient attention in the literature. Based on the review, this paper proposes a conceptual framework for trustworthy XAI-enabled higher education analytics that integrates educational data, modelling, explanations, fairness, domain knowledge, human-level interpretation, and interventions. This paper also identifies research gaps and discusses potential research directions in relation to human-centric explanations, domain-knowledge-aware XAI, fairness-aware prediction, multimodal educational data, and evaluation of usefulness of model explanations. The review and

framework proposed in this paper provide a foundation for researchers and higher-education institutions that seek to build explanatory and trustworthy AI-driven systems for education.

Keywords: Explainable Artificial Intelligence; Higher Education; Educational Data Mining; Student Performance Prediction; Learning Analytics; Machine Learning; SHAP; Trustworthy AI; Personalized Learning.

1. Introduction

The rise of Artificial Intelligence has fundamentally changed the way educational institutions acquire, analyze, and utilize information on teaching and learning. Higher-education institutions have adopted learning analytics platforms and machine learning models to support teaching and learning, academic administration, student services, and institutional research. Universities worldwide generate, collect, and store huge amounts of student-related data through learning-management systems, student-information systems, online examinations, library systems, attendance monitoring, and other digital platforms. Such data can allow educational data miners and machine learners to uncover patterns that may be used to improve teaching and learning, student services, and administration. One of the major applications of machine learning in higher education is student-performance prediction. Predictive models can be leveraged to identify students at risk of poor academic performance so that the appropriate interventions can be made in good time. It has been shown that machine learning, deep learning, and educational data mining have been widely applied to predict student performance. In particular, a recent study conducted a comprehensive literature review on student-performance prediction and revealed that tree-based algorithms and neural networks were the most commonly applied models, and that there is a strong emphasis on using explainable AI to make such models more interpretable and applicable in practice. Nonetheless, while complex models offer higher accuracy and performance, their application in education involves various challenges since most models are often “black-box” in nature, making it difficult for teachers, administrators, and students to understand the rationale behind the produced results.

This issue has prompted the need to conduct research on Explainable Artificial Intelligence (XAI) and its applications in higher education. XAI aims to enable educators, administrators, policymakers, and students to trust, monitor, validate, and effectively apply the results of machine learning and deep-learning models. Various research studies have demonstrated that SHAP, LIME, feature importance, rule extraction, and inherently-interpretable models can be used to explain the predictions of black-box models. Earlier studies conducted systematic literature reviews on explainable student-performance prediction and revealed research trends and challenges related to the use of explainable AI. More recently, research studies have been conducted on the application of XAI in different disciplines such as Computer Science and STEM education. Several studies have demonstrated the potential of applying XAI and post-hoc explanation methods for student-performance prediction and educational decision support. The application of XAI in education has, however, been limited due to various challenges associated with transparency, explainability, fairness, trust, privacy, and applicability.

This paper aims to review the state-of-the-art in XAI in higher education with a focus on student-performance prediction and educational decision support. In particular, this paper reviews the major application areas, predictive models, and explanation methods applied in XAI for education. This paper also reviews different types of explanations and discusses major technical, ethical, pedagogical, and human centric challenges in implementing XAI in higher education. Furthermore, this paper highlights the research gaps in the current literature on XAI in higher education. In addition, this paper proposes a conceptual framework for building trustworthy and actionable XAI-enabled systems in higher education.

1.1 Research Objectives

This paper seeks to achieve the following objectives:

1. Review the major applications of XAI in higher education
2. Identify the commonly-applied machine-learning models and XAI techniques
3. Examine how predictions are explained and utilized in student-performance prediction
4. Discuss the major technical, ethical, pedagogical, and human centric challenges
5. Highlight research gaps in the current literature on XAI in higher education
6. Propose a conceptual framework for trustworthy and actionable XAI-enabled higher-education analytics

1.2 Research Questions

This paper seeks to answer the following research questions:

RQ1: What are the major areas of application of XAI in higher education?

RQ2: What are the commonly-applied machine-learning models and XAI techniques?

RQ3: What are the major categories of explanations generated in XAI?

RQ4: What are the major technical, ethical, fairness, and human-centric challenges?

RQ5: To what extent can XAI be leveraged to generate educational interventions?

RQ6: What research directions can be pursued to improve the trustworthiness and practical usefulness of XAI in higher education?

2. Background

AI-based educational systems are increasingly being applied for prediction, recommendation, assessment, tutoring, and learning analytics. Predictive analytics can be used to identify students at risk of poor performance so that appropriate interventions can be implemented in good time.

The major challenge with predictive analytics lies in ensuring that the generated predictions are interpretable and actionable. This is particularly important in education since the decisions made based on predictive analytics have direct impacts on the lives of students.

2.1 Explainable Artificial Intelligence

Explainable AI encompasses various methods and techniques that have been developed to render AI models and their predictions interpretable and accessible to end-users.

Explanations can take different forms depending on the type, nature, and complexity of the model. Two broad classes of explanation methods have been proposed:

1. Intrinsically interpretable models
2. Post-hoc explanation methods

Intrinsically-interpretable models are often referred to as inherently-explainable models. Such models do not require post-processing since they are relatively simple and operate on easily-interpretable logic. Commonly-applied intrinsically-interpretable models include Decision Trees, Rule-Based Systems, Linear and Logistic Regression, and Generalized Additive Models.

On the other hand, post-hoc explanation methods can be used to explain the predictions of complex "black-box" models. The major post-hoc explanation methods applicable in education include SHAP, LIME, Permutation Feature Importance, Partial Dependence Plots, Individual Conditional Expectation, and Counterfactual Explanations.

SHAP is arguably the most prominent XAI technique that has been applied in education for explaining model predictions.

3. Literature Review

The application of Artificial Intelligence (AI) and Machine Learning (ML) in higher education has been growing rapidly in areas such as student-performance prediction, dropout-risk identification, learning analytics, assessment, personalized learning, and academic decision support. Although predictive models can reveal complex patterns in student data, the use of complex ML algorithms has challenged efforts to render such models interpretable and applicable in practice. Explainable Artificial Intelligence (XAI) has therefore become an important research area that aims to promote transparency, accountability, trust, and human-level understanding of educational decisions made by ML models.

Alamri and Alharbi (2021) conducted a systematic review of explainable student-performance prediction models and emphasized the need to consider both predictive accuracy and interpretability in model design. In particular, the authors demonstrated that machine-learning models can be effectively applied to predict student performance while simultaneously providing explanations of the generated predictions. The study by Alamri and Alharbi provides an important foundation for subsequent studies on explainable student-performance prediction. More recent research studies have been focusing on combining conventional machine-learning algorithms with post-hoc explanation techniques. Tree-based algorithms such as Random Forest and XGBoost have been widely applied due to their relatively high predictive accuracy as well as their compatibility with post-hoc explanation techniques such as SHAP. Compared to tree-based and logistic-regression models, more complex algorithms such as neural networks and deep-learning models generally require more sophisticated post-hoc explanation methods (e.g., SHAP, LIME, Counterfactual Explanations).

Various research studies have investigated the application of XAI techniques in Computer Science and STEM education. In particular, Choi et al. (2025) conducted a systematic literature review on the application of XAI in Computer Science and STEM education. Their work identified the key areas of application, algorithms, and explanation methods that have been commonly applied to predict student performance in Computer Science and STEM education. Choi et al. (2025) revealed that SHAP was the most common explanation method, and that academic and behavioural features were the most common predictors of student performance. However, the review also identified the limited application of XAI for course improvement, visualization, and recommendation (Choi et al., 2025). This suggests that most studies emphasize the role of XAI in providing explanations without fully leveraging such insights to recommend educational interventions and improve learning analytics. In a related study, Boujmiraz et al. (2026) conducted a comprehensive literature review on machine-learning, deep-learning and explainable-AI approaches to student-performance prediction. Their study demonstrated that various researchers have been using predictive analytics to estimate student performance with a focus on tree-based models and neural networks. The study by Boujmiraz et al. (2026) indicates that there is a growing recognition of the role of AI in education, as well as the need to ensure the transparency of such systems in practice.

Recent research studies have sought to integrate educational domain knowledge into explainable machine-learning models. Qiang, Liu, and Zhang (2026) investigated the possibility of incorporating educational domain knowledge into an explainable AI approach for student-performance prediction. This work is important because it recognizes that the mere inclusion of easily-interpretable features such as GPA often fails to account for the complex relationships between various educational variables (Qiang, Liu, & Zhang, 2026).

3.1 XAI Techniques in Higher Education

Various explanation techniques have been applied in higher education to render complex models interpretable. One of the most prominent XAI techniques is SHapley Additive exPlanations (SHAP). SHAP is a model-agnostic explanation technique that assigns Shapley values to input features to indicate their contribution to a given prediction. In particular, SHAP can be used to provide both individual and global explanations for a given model. For example, SHAP can provide an explanation of why a particular student has received a certain grade. Local Interpretable Model-Agnostic Explanations (LIME) is another model-agnostic explanation technique that approximates the behaviour of a complex model locally to provide explanations for a particular input instance. One of the major limitations of LIME is its inconsistency in providing stable explanations. Feature-importance approaches such as Permutation Importance allow researchers to assess the relative importance of each input variable in a given model. However, while permutation importance can be used to provide global explanations, it is not always informative when it comes to individual predictions. Counterfactual Explanations (CE) is another approach that has been increasingly applied in education to improve model interpretability. Unlike other explanation approaches, CE allows researchers and practitioners to determine what changes to the input variables would have led to a different prediction. Such information can then be used to recommend specific actions that could have improved a given student’s performance or reduced the risk of failure or dropout.

3.2 Explain ability and Student Performance Prediction

Student-performance prediction is one of the key application areas of AI in higher education. Educational institutions generate and store a wealth of students’ data that can be used to build predictive models that can help improve teaching and learning. Student-performance prediction involves using students’ academic records, attendance, and other relevant data to estimate a student’s likely performance, risk of failure, and likelihood of dropping out. Predictive models can then be used to recommend appropriate interventions for students at risk. However, while such models can be useful, their application is limited by various challenges. One of the major challenges of applying predictive models in student-performance prediction is the fact that a prediction is rarely actionable on its own. For example, a prediction about a student’s risk of failing a course requires additional information about why such risk occurred and what interventions can be implemented to mitigate the risk. SHAP, LIME, and other XAI techniques can be used to explain the rationale behind such a prediction. Literature review has shown that most researchers have been focusing on using XAI to provide explanations rather than developing approaches for actionable interventions. Another insight from the review is the fact that different stakeholders require different types of explanations. In particular, most students and teachers lack the technical expertise to understand explanations generated using SHAP values, and thus such explanations should be presented differently to students and teachers. In addition, most XAI studies emphasize model transparency while paying limited attention to issues such as fairness and privacy. In particular, a number of researchers have raised concerns about the potential risks associated with the use of predictive analytics in student-performance prediction.

3.3 Fairness, Privacy and Trust

Fairness, privacy, and trust are major ethical concerns that must be considered when applying predictive analytics and machine learning in higher education. Many studies have demonstrated that predictive models in education tend to be biased against female and minority students due to historical disparities in various aspects of education. As a result, it is essential to ensure that the use of predictive analytics and machine learning for student-performance prediction and educational decision support does not lead to the same level of discrimination. One of the major approaches for promoting fairness in machine learning is to incorporate fairness constraints in model design. Some researchers have suggested the need to consider different fairness notions such as demographic parity, equal opportunity, and equalized odds. Another major concern in the application of machine learning in higher education is student privacy. It is essential to ensure that students' data are collected, processed, and stored in a manner that guarantees confidentiality and security. In particular, educational institutions must adopt responsible data-governance policies and practices that minimize the risks of data breaches and unauthorized access to students' data. Finally, privacy concerns must also be considered when designing and applying XAI techniques. Some researchers have raised concerns about potential privacy risks associated with the application of XAI techniques such as SHAP, LIME, and CE. In particular, such techniques can be used to infer sensitive information about students' behaviours, attitudes, and demographics based on their educational data.

3.4 Research Gaps Identified from the Literature

Based on the review of literature on XAI in higher education, several research gaps can be identified. First, most researchers have been focusing on improving model performance and using XAI to provide explanations of model results without explicitly considering whether such explanations are actionable. Second, many studies emphasize the identification of important features while providing limited guidance on how such features can be leveraged to develop educational interventions that improve student performance. Third, most researchers have been applying post-hoc explanation methods rather than leveraging educational domain knowledge to render models more interpretable. Fourth, fairness is rarely considered alongside model accuracy and interpretability. Fifth, much of the research has been conducted using individual datasets and/or institutional data, thus limiting the generalizability of findings. Finally, little attention has been paid to the stability of various explanation approaches. In particular, most researchers have overlooked the fact that minor changes to input data can lead to very different explanations, thus casting doubt on the reliability of such explanations. One of the major contributions of this study is the identification of these research gaps, as well as the proposal of a conceptual framework for trustworthy and actionable XAI-enabled higher-education analytics.

3.5 Review Methodology

This study utilizes a systematic review methodology to identify and synthesize research on XAI in higher education with a focus on student-performance prediction and educational decision support.

The review follows the standard procedures for conducting literature searches, screening, eligibility assessment, data extraction, classification, and thematic analysis.

3.5.1 Search Strategy

The review will leverage a combination of keywords and controlled vocabulary to identify relevant studies.

The search will utilize the following combination of terms:

("Explainable AI" OR "Explainable Artificial Intelligence" OR XAI OR "Explainable Machine Learning")
AND
("higher education" OR university OR college OR "tertiary education")
AND
("student performance" OR "academic performance" OR "learning analytics" OR dropout OR "student prediction")
This search can be executed in various databases, including Scopus, Web of Science, IEEE Xplore, ScienceDirect, SpringerLink, ACM Digital Library, and Google Scholar.
The final journal submission will provide details on the search strategy (i.e., database-specific search strings).

3.5.2 Inclusion Criteria

The following criteria will be used to select eligible studies for this review:

1. Address AI, ML, or XAI;
3. Concern higher education or educational settings directly relevant to higher education;
4. Discuss prediction, classification, learning analytics, assessment, recommendation, or educational decision-making;
5. Provide sufficient methodological detail;
6. Be published in peer-reviewed publications;
7. Be available in English;
8. Meet all other eligibility criteria.

3.5.3 Exclusion Criteria

The following criteria will be used to exclude studies:

1. Do not address AI, ML, or XAI
2. Are not concerned with education
3. Are focused on primary or secondary education (if applicable)
4. Are reviews, opinion pieces, or advertisements
5. Do not provide sufficient methodological detail
6. Are duplicates

4. Findings from the Existing Literature

The literature reveals several key areas of application.
Foremost among these is the prediction of student performance. Predictive models use academic, behavioural, demographic, engagement, and learning-management-system variables as inputs and grade, course outcome, or performance category as an output.
Explainable AI can detect students at risk of failure, withdrawal, or dropout. XAI can explain such a classification.
Explainable recommendations disclose information about why a specific learning resource, activity, or recommendation has been made.
XAI can explain patterns in engagement, assessment, attendance, and learning analytics.
AI is used for formative and summative assessments, and a recent systematic review identified 159 studies of AI-assisted educational assessment conducted in 2026. These utilised classification algorithms for student-performance prediction.

5. Machine Learning and XAI Techniques

Various predictive models have different degrees of intrinsic explainability.

For instance, logistic regression is often used for performance and risk prediction, because its coefficients are often easily interpretable. Decision trees can be used to classify students, because their structure is intuitive. Others include random forest, XGBoost, and support vector machine, all of which can be explained using SHAP and feature-importance values.

Neural networks and deep learning architectures can be explained using SHAP, LIME, and counterfactuals.

The literature suggests that tree-based models are favoured for their high performance-to-explainability ratios and applicability in SHAP-based explainers.

6. From Prediction to Explanation

An important finding from the literature is that prediction and explanation are not equivalent to intervention.

For instance, if Student A is likely to fail a course, a predictive model can identify this fact, but an educationally useful system would go further and attempt to answer these questions:

1. Why was the student classified as highly likely to fail?
2. Which factors contributed the most?
3. Which factors can be intervened?
4. Are the explanation aligned with educational theory?
5. Is the prediction fair?
7. What is the suggested academic intervention?
8. How certain is the model about the prediction?

This leads to the cycle below:

Data → Prediction → Explanation → Interpretation → Intervention → Outcome Evaluation

It is noteworthy that most studies focus on the first three steps, whereas future research should focus on developing a system that embodies all steps.

7. Major Challenges

7.1 Explainability versus Accuracy

Complexity often undermines explainability, but simpler models are sometimes outperformed by more complex ones in terms of predictive performance. Future research should focus on predictive accuracy, calibration, explanation fidelity, stability of explanations, human comprehension, and actionability.

7.2 Individual-Level Explanations

While global explanations can tell us which factors contribute to likelihood of dropout on average, individual explanations are necessary for intervening.

7.3 Fairness

Students' backgrounds are often represented in the data, and failure to account for this can result in XAI techniques perpetuating the biases embedded in the training data. Future research should focus on developing fairness-centred XAI, which accounts for potential discrimination.

7.4 Privacy

Education data can contain private information about students' lives. Future research should focus on privacy-preserving algorithms, data minimisation, and appropriate governance.

7.5 Human Understanding

A human's understanding of an explanation is a critical factor in determining its efficacy. Post hoc explanation methods should be validated using human subjects.

7.6 Lack of Actionability

Most studies focus on explaining the factors associated with a prediction. Future research should focus on actionable explanations, whereby students, faculty, and administrators receive individual-level intervention recommendations based on the findings.

8. Research Gaps

Gap 1: The literature lacks human-centric studies, which explore whether post hoc explanation methods actually help teachers and students.

Gap 2: The literature lacks studies that go beyond simply identifying features associated with a prediction and utilise these insights to recommend individual-level academic interventions.

Gap 3: There is limited research on how well the results of machine learning align with educational knowledge and theory.

Gap 4: There is limited research on the fairness of educational XAI and how well it accounts for potential discrimination.

Gap 5: Most predictive studies are conducted on a single educational institution, which limits the generalizability of findings.

Gap 6: Most literature focuses on local explanation methods, which are prone to instability, especially when the input data contains noise.

Gap 7: An important question that remains unexplored is whether the interventions recommended by an educational XAI will lead to positive outcomes.

9. Proposed Conceptual Framework

Based on the research gaps, a conceptual framework was proposed. It consists of seven layers:

Layer 1: Educational Data: grades, attendance, assessments, assignment performance, learning-management-system analytics, course analytics, and academic history.

Layer 2: Data Governance: data quality, missing data, data privacy, data security, data bias, and data minimisation.

Layer 3: Predictive Analytics: Logistic Regression, Decision Tree, Random Forest, XGBoost, Support Vector Machine, and Neural Networks.

Layer 4: Explainability: SHAP, LIME, feature importance, counterfactuals, and partial-dependence plots.

Layer 5: Educational Knowledge: alignment of the findings from Layer 4 with educational theory and practice.

Layer 6: Human-Centred Interpretation: Different interpretations for teachers, students, and administrators.

Layer 7: Actionable Intervention: additional learning resources, academic tutoring, attendance reminders, learning analytics dashboards, and faculty intervention.

The findings from Layer 7 are utilised to inform further analytics and form a feedback loop.

10. Proposed Evaluation Framework

When implementing the framework, five categories should be considered:
Predictive Performance: Accuracy, Precision, Recall, F1-score, ROC-AUC, and PR-AUC.
Explainability: Fidelity, stability, consistency, complexity, and human interpretability.
Fairness: Demographic parity, equal opportunity, and equalised odds.
Actionability: Whether an explanation leads to an intervention.
Educational Effectiveness: Course completion, performance, attendance, and retention.

11. Discussion

The literature shows that XAI is a growing field, which seeks to make educational-AI models more transparent and trustworthy. While previous studies have demonstrated the ability of XAI to facilitate human-level understanding of machine learning, some of the critical questions remain unexplored. This study highlights the distinction between a technical explanation and an educational one. While the former focuses on the technical aspects, such as how a specific factor affected the outcome, the latter is concerned with whether the explanation facilitates learning.

For instance, a post hoc explainer could report that, for Student A, the decrease in attendance by 18% was associated with an increased likelihood of failure by approximately 0.18 . An educational explanation, on the other hand, could indicate that, since Student A’s class attendance decreased by 18% and assignment performance dropped by 20%, their likelihood of failure increased, which would require a faculty outreach.

Therefore, future post hoc explainers should go beyond merely providing technical explanations and utilise educational theory to facilitate teaching and learning.

In this regard, studies conducted by Qiang, Liu, and Zhang in 2026 are edifying. They demonstrate that an explainer, which incorporates educational knowledge, can be used to improve the predictive performance of a machine learning model. Specifically, Qiang, Liu, and Zhang found that aligning the learning process of a model with existing knowledge can enhance its performance. Future studies should adopt this knowledge-informed approach to highlight how specific factors, which contribute to an outcome, could be leveraged to improve educational effectiveness.

12. Future Research Directions

Future research should focus on the following areas:
1. Domain-knowledge-aware XAI for educational prediction.
2. Fairness-aware educational AI for prediction and recommendation.
3. Personalised educational explanations.
4. Educational counterfactuals.
5. Educational analytics using multimodal data.
6. Human-centred educational explainers.
7. Cross-institutional validation.
9. Longitudinal studies.
10. Educational analytics with privacy-preserving AI.
11. Explanation usefulness in education.

13. Conclusion

Explainable Artificial Intelligence has the potential to transform higher-education analytics by making them more transparent and trustworthy. The literature reviewed in this paper shows that XAI can facilitate post hoc explanation of complex machine-learning algorithms, which is critical for adopting these algorithms in sensitive domains, such as higher education. However, the reviewed studies mostly focused on algorithmic transparency at the model level.

The next frontier of research should be to transform these explanations into educational knowledge and actionable interventions. This paper reveals several research gaps, which include limited consideration of individual-level explainers, limited focus on education-specific considerations when developing post hoc explainers, limited studies on fairness and validity of educational XAI, limited human-level evaluation of post hoc explainers, and limited experimental evidence on whether XAI-assisted recommendations lead to positive educational outcomes.

To advance the frontier of knowledge, this paper proposes a conceptual framework, which incorporates educational data and analytics, data governance, predictive modelling, post hoc explainers, educational knowledge, human factors, and interventions. The future of XAI in higher education will be defined by the shift from "Why did the model make this prediction?" to "How can an understandable, fair, and trustworthy AI explanation help educators and students make better decisions?"

References

1. Alamri, R., & Alharbi, B. (2021). Explainable Student Performance Prediction Models: A Systematic Review. IEEE Access, 9, 33132–33143. https://doi.org/10.1109/ACCESS.2021.3061368
2. Very important for your paper because it directly reviews explainable student-performance prediction.
3. Alamri, R., & Alharbi, B. (2022). Explainable Artificial Intelligence in Education. Computers and Education: Artificial Intelligence, 3, 100074. https://doi.org/10.1016/j.caeai.2022.10007
4. Useful for your sections on XAI, stakeholders, human-centred explanations, and educational applications.
5. Choi, W. C., Lam, C. T., Pang, P. C. I., & Mendes, A. J. (2025). A Systematic Literature Review of Explainable Artificial Intelligence (XAI) for Interpreting Student Performance Prediction in Computer Science and STEM Education. Proceedings of the 30th ACM Conference on Innovation and Technology in Computer Science Education, 221–227. https://doi.org/10.1145/3724363.3729027
6. Particularly useful for your discussion of SHAP, behavioural/academic features, individual explanations and personalized recommendations.
7. Qiang, M., Liu, Z., & Zhang, R. (2026). Explainable AI in education: Integrating educational domain knowledge into the deep learning model for improved student performance prediction. Scientific Reports, 16, 9515. https://doi.org/10.1038/s41598-026-40538-y
8. Strong reference for your proposed Educational Knowledge Alignment layer.
9. An explainable AI-based approach for predicting undergraduate students academic performance. (2025). Array, 26, 100384. https://doi.org/10.1016/j.array.2025.100384
10. Useful for your section comparing SHAP and LIME for undergraduate performance prediction.
11. The integration of explainable AI in Educational Data Mining for student academic performance prediction and support system. (2025).

12. Particularly relevant to your "Prediction → Explanation → Intervention" argument because the study combines ML, SHAP/LIME and a student support/recommendation system.
13. Jeenno, J., & Thongkam, O. (2025). A Systematic Review of State-of-the-Art Explainable AI in Education. Applied Machine Learning and Analytics. https://doi.org/10.64220/amla.v1i1.005
14. Useful as a broader review of XAI applications and challenges in education.
15. Prentzas, J., & Binopoulou, A. (2025). Explainable Artificial Intelligence Approaches in Primary Education: A Review. Electronics, 14(11), 2279. https://doi.org/10.3390/electronics14112279
16. Although focused on primary education, it can be useful for discussing the broader development of XAI in education and distinguishing your higher-education scope.
17. Explainable Machine Learning for Student Performance Prediction. (2026). AI Education, 2(2), 17. https://doi.org/10.3390/aieduc2020017
18. Very recent work using complementary XAI approaches, including Gradient SHAP and counterfactual explanations, for identifying at-risk students.
19. Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30.
20. 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, 1135–1144. https://doi.org/10.1145/2939672.2939778
21. Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-Precision Model-Agnostic Explanations. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1).
22. Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.
23. Useful for your theoretical discussion of global/local explanations, feature importance, SHAP, LIME and partial dependence.
24. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys, 51(5), 1–42. https://doi.org/10.1145/3236009
25. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012