Journal scientific and applied research, vol. 29, 2025
International Journal

Original Contribution

ISSN 1314-6289 (Print)
ISSN 2815-4622 (Online)

EXPLAINING ARTIFICIAL INTELLIGENCE IN EDUCATION: APPROACHES TO INTEGRATION IN AUTOMATED INFORMATION SYSTEMS

Desislava N. Ivanova

DEPARTMENT OF COMMUNICATION AND COMPUTER ENGINEERING AND SECURITY TECHNOLOGIES, FACULTY OF TECHNICAL SCIENCES, KONSTANTIN PRESLAVSKY UNIVERSITY OF SHUMEN, SHUMEN 9712,115, UNIVERSITETSKA STR., E-MAIL: d.n.ivanova@shu.bg

ABSTRACT: *This paper provides a systematic review of approaches for integrating Explainable Artificial Intelligence (XAI) into educational Automated Information Systems (AIS). It categorizes XAI methods into post-hoc explanations, inherently interpretable models, hybrid approaches, and calibration/visualization techniques, analyzing their strengths, limitations, and applicability for tasks such as learner dropout prediction, early warning, and personalized mentoring. Practical examples illustrate the benefits and challenges of XAI adoption, including the trade-off between accuracy and interpretability, technical barriers, and privacy concerns. Future directions include role-adaptive explanations, visual and interactive interfaces, and standardized quality metrics for educational contexts.*

KEY WORDS: *Explainable artificial intelligence, Automated information systems, Machine learning, Education, Interpretability, Early warning, Learner dropout.*

1. Introduction

Automated Information Systems (AIS), especially those integrating Machine Learning (ML) algorithms, are emerging as an important tool in modern education [1]. They support processes such as predicting learners' dropout risk, providing early warning of potential learning difficulties, and delivering personalized tutoring to increase achievement and motivation.

Personalized tutoring is a pedagogical approach in which the content, methods and intensity of tutoring are adapted to the individual needs, abilities, interests and learning pace of the specific learner [2].

In the context of automated information systems and machine learning, this means:

- Needs diagnosis - the system analyses data on performance, learning habits, attendance, social activity to determine the learner's strengths and weaknesses.
- Personalized recommendations - based on the analysis, the system suggests specific learning resources, exercises or activities to help overcome specific difficulties.
- Adaptive scheduling and content - the system can change the difficulty, format, or volume of materials according to the learner's progress.
- Tracking and feedback - a tutor (teacher or automated module) receives progress data and can take specific actions in a timely manner.

Although machine learning models can achieve high "prediction" accuracy, they often function as "black boxes" - i.e., their algorithmic internal processes and decision logic remain hidden from the end user. The lack of transparency in explaining predictions makes it difficult for key stakeholders in the education system, including educators, administrators, and parents, to accept them. This raises critical issues related to the ethics of the decisions, the credibility of the system, and the accountability of the algorithms to the public [3].

In response to these challenges, the field of Explainable Artificial Intelligence offers methods and techniques that make predictions of complex patterns more understandable, both at the global level (generic to all predictions) and at the local level (case-specific) [4]. These include post-hoc techniques (methods that generate explanations after model training), models that are interpretable in nature (that are understandable in their structure), and calibration and visualization approaches (aimed at verifying and visually representing the plausibility of probabilistic estimates).

The aim of this paper is to propose a systematic review of existing approaches for the integration of XAI in educational automated information systems, analyzing their advantages, limitations and applicability in real educational environments.

2. Classification and critical analysis of existing approaches to explainability in machine learning

There is a wide range of methods and techniques for achieving explainability in machine learning, which can be classified into several main categories. In the context of XAI integration in educational automated information systems, it is essential to know their characteristics, advantages, limitations and disadvantages, as the choice of approach directly affects the effectiveness and credibility of the system.

2.1. Post-hoc explainability

Post-hoc methods are applied after model training and aim to provide an interpretation of already existing "black boxes" without changing their structure.

- SHAP (SHapley Additive exPlanations) - is based on cooperative game theory and calculates the contribution of each input variable to a specific forecast. SHAP offers both global (population-wide) and local (case-specific) explanations, making it particularly suitable for educational contexts where it is important to explain both the general logic of the model and specific solutions for individual learners [5].
- LIME (Local Interpretable Model-agnostic Explanations) - a local, model-independent method that approximates a complex model in a small region around a specific forecast by an easily interpretable linear model. This allows users to understand which factors had the greatest influence on a particular decision [5].
- Anchors - an approach that generates "anchor rules" describing the minimum set of conditions that, if satisfied, guarantee the same model prediction with high probability.
- Permutation Importance - a method in which the values of a variable are shuffled to assess the extent to which this degrades the performance of the model, which is an indicator of the importance of the variable.

2.2 Interpreted by the nature of the model

Interpretable patterns are by nature those whose structure and logic are simple enough to be understood without the need for additional explanation techniques.

- Logistic regression - a classical statistical model in which the influence of each variable is clearly expressed by coefficients, allowing direct interpretation.
- Decision Trees - represent a sequence of conditions (if-then rules) that lead to a decision, making them easy to follow and explain.
- Rule-based systems - expert-defined or automatically derived sets of rules that directly describe decision logic.

Despite their high explainability, these models often have limited predictive power with complex data.

2.3. Hybrid approaches

Hybrid approaches combine the advantages of powerful but opaque models with the explainability of simpler techniques [6].

An example is the use of a complex ML model for basic prediction, followed by a system with rules to validate or explain the results.

Another option is the application of surrogate models-simpler, interpretable models that are trained to mimic the behavior of the complex model, thus providing an understandable approximate version of its logic.

2.4. Approaches to calibration and visualization

The calibration aims to check how well the model's predicted probabilities match the actual observed event frequencies.

- Reliability diagrams - a graphical method of visualizing calibration showing the correspondence between predicted probability and actual frequency.
- Partial Dependence Plots (PDPs) - illustrate the relationship between one or more input variables and the model's predicted outcome, while fixing the remaining variables.
- Fairness dashboards - visual fairness dashboards that allow detection of systematic biases in predictions towards different subgroups (e.g. by gender, age or socio-economic status).

These visual techniques play a key role in building trust and provide a means for effective communication between technical experts and educational specialists.

Table 1. presents a comparative analysis of the main approaches to explainability in ML.

Table 1 Comparative analysis of the main approaches to explainability in ML machine learning

Approach Category	**Examples**	**Advantages**	**Disadvantages**	**Suitable Application Cases**
Post-hoc Explainability	SHAP, LIME, Anchors, Permutation Importance	- Can be applied to any model (model-agnostic) - Supports global and local explanations - Detailed assessment of each feature's impact	- Additional computational complexity - Risk of misinterpretation - Approximate, not exact explanations	When a complex 'black-box' model is already trained and its predictions need to be explained
Inherently Interpretable Models	Logistic regression, decision trees, rule-based systems	- High transparency - Easily understandable for non-technical users - Fast computation and analysis	- Often lower accuracy for complex tasks - Limited ability to model complex relationships	When transparency and trust are a priority, and accuracy is of secondary importance
Hybrid Approaches	Complex ML model + rules;	- Combine high accuracy and	- Increased system	When aiming for a balance

	surrogate models	explainability - Flexible adaptation to specific domains	complexity - Requires careful synchronization between components	between accuracy and explainability; in critical applications (education, healthcare)
Calibration and Visualization	Reliability diagrams, PDPs, fairness dashboards	- Improve understanding of model reliability and fairness - Visually accessible to a broad audience	- Do not directly explain the model's internal logic - Require skills for correct interpretation	When it is important to demonstrate reliability, fairness, and the effect of specific factors on the prediction

3. Examples from education

The implementation of explainable artificial intelligence in educational automated information systems is already finding real-world applications in a range of supporting processes, from dropout prevention to individualizing the learning process and increasing assessment transparency.

3.1. Early Warning Systems (EWS)

Early Warning Systems (EWS) aim to identify learners at increased risk of dropping out well before critical events occur. An example of such an approach is the Early Warning System in the state of Indiana, USA, which integrates machine learning and explainable methods (e.g., SHAP) to provide school administrators with clear reasons for any prediction [7]. If the model predicts an 85% probability that a learner will drop out, SHAP can show that the root causes are 15% frequent absenteeism, 10% drop in math grades, and 8% low engagement in the e-learning platform. This allows the teaching team to take specific action in a timely manner.

3.2 Personalized learning

Personalized learning systems adapt the learning content and pace to the individual profile of the learner [17]. Knewton and DreamBox Learning are examples of platforms that use adaptive learning algorithms and explainable recommendation mechanisms to show the trainee and learner why a particular module is suggested [8]. If a learner is recommended an additional algebra course, the system can explain that the decision was based on a combination of low test scores in that area and better performance on visual materials. This increases learner confidence and motivation, and learners receive valuable information about individual needs.

3.3 Analysis of participation and performance

Tracking trainee participation and academic performance is a key component of maintaining a high level of success. In Finland, the MyData initiative provides learners and parents with access to personalized visual explanations of the factors influencing outcomes through interpretable models such as Generalized Additive Models (GAM) and decision rules [9]. Such visualizations show how frequency of participation in virtual discussions affects average success, and can be integrated into dashboards accessible to learners, educational leaders, and parents, improving transparency and collaborative work to support learners.

4. Benefits of XAI integration

Integrating explainable artificial intelligence into educational automated information systems provides significant benefits that extend beyond the purely technical aspects of modeling. These benefits translate into increased trust among stakeholders, improved decision support, and ensuring ethical accountability.

One of the main challenges in implementing machine learning-based systems is the "black box" barrier that reduces the propensity of trainers and administrators to use the results. SHAP visualizations, lead to a significantly higher acceptance rate of predictions in an educational context [10].

XAI facilitates the translation of complex models into concrete actions. For example, the MyData project in Northern Europe [11] demonstrated how visual explanations of predictions guide teachers to the right interventions, from extra help in a particular subject to social support to increase attendance. In the context of dropout risk management, this means that education teams can proactively address problems before they get worse.

The ethical aspect of AI in education requires mechanisms to monitor and control algorithmic decisions [12]. XAI allows to identify systematic biases, for example if the model disproportionately marks learners from certain social groups as "at risk". By analyzing feature importance by demographic segments, it can be determined whether the model is fair. The academic literature has documented cases where lack of explainability has led to students being inappropriately excluded, and XAI integration has prevented such scenarios from recurring [13]. Figure 1 provides an example of the role of XAI in educational AIS:

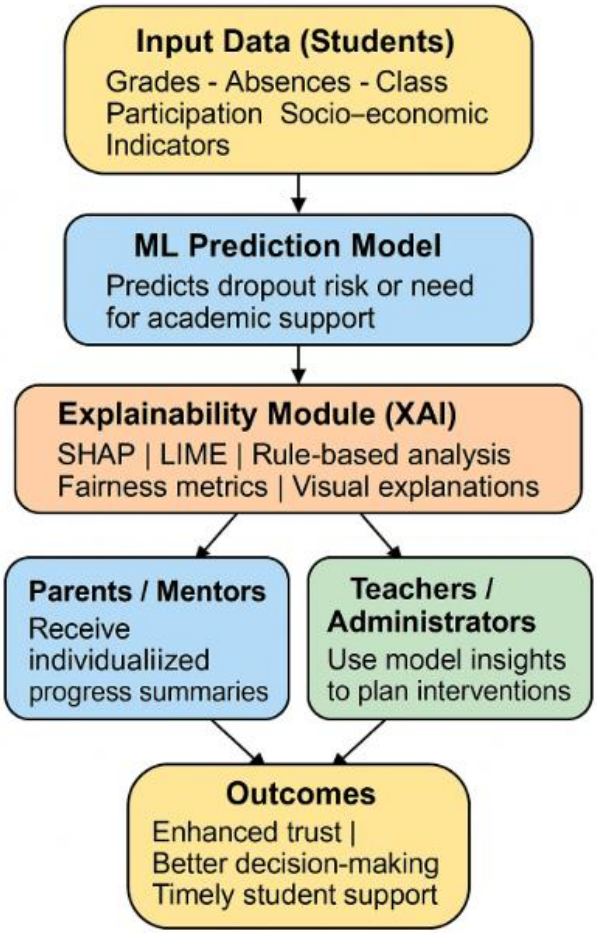
Input Data (Students)
Grades - Absences - Class Participation Socio–economic Indicators
ML Prediction Model
Predicts dropout risk or need for academic support
Explainability Module (XAI)
SHAP | LIME | Rule-based analysis Fairness metrics | Visual explanations
Parents / Mentors
Receive individualiized progress summaries
Teachers / Administrators
Use model insights to plan interventions
Outcomes
Enhanced trust | Better decision-making Timely student support

Fig. 1. Role of XAI in Educational Information Systems

5. Limitations and challenges

Despite the significant potential of explainable artificial intelligence in educational AIS, its integration faces a number of practical, technical and even ethical obstacles:

- risk of misinterpretation - even when using well-established methods such as Local Interpretable Model-agnostic Explanations (LIME), there is a danger that end-users - trainers or administrators - will misinterpret local explanations, especially if sufficient statistical and methodological training is lacking [14]. This can lead to inappropriate interventions or decision-making based on incomplete or misleading information.
- technical barriers - implementing XAI functionalities in real-time requires significant computational resources, architecture optimization, and a carefully designed user interface (UI) that presents explanations in an accessible and understandable manner [15]. A poorly designed UI can negate the technical advantages of XAI.
- Accuracy vs. interpretability - there is the classic trade-off: simpler models (e.g., linear regression, decision trees) are usually easier to explain, but often lose accuracy with complex and heterogeneous

educational data. On the other hand, deep neural networks and ensemble methods achieve better predictive power, but are "black boxes" and require post-hoc explanation techniques.

- Ethical considerations and privacy - explanations generated by XAI systems may reveal sensitive information - e.g., socioeconomic status, personal behavioral patterns, or specific difficulties of a learner. There is a need to ensure that such data is processed and visualized in a way that complies with applicable privacy regulations (e.g. GDPR) and the ethical use of AI in education.

6. Future development directions

Despite advances in the integration of explainable artificial intelligence into educational AIS, there are a number of promising directions for development that can enhance their effectiveness and stakeholder acceptance. Similar cognitive automation concepts have been explored in other intelligent system design contexts, including the development of web applications supported by cognitive machines [16].

- Automatically adapt explanations to the user's role. One of the key trends is the creation of dynamic XAI modules that customize the content, format and detail of explanations depending on the user profile.
- Integration with visual and interactive explanation interfaces.
- In the future, XAI components in educational systems should be able to use interactive dashboards, simulations, and animated visualizations that not only present results but also allow the user to experiment with the input data to understand how changes affect predictions.
- Standardized metrics for explanation quality in educational contexts. There is a need for clear and commonly accepted metrics for evaluating XAI in an educational setting that cover both the technical accuracy and robustness of explanations and their comprehensibility, usefulness, and impact on decisions. Such metrics can serve as a basis for certification and regulation of XAI solutions in education.

7. Conclusion

This paper highlights that the integration of explainable artificial intelligence into educational automated information systems has the potential to significantly change the way decisions are made, risks identified and interventions implemented in educational environments. XAI not only increases transparency and trust, but also facilitates communication between technical experts, educators, administrators, and parents.

Key findings from the analysis indicate that successful XAI implementation requires a careful balance between accuracy and interpretability, adaptation to the user context, and strict adherence to ethical and regulatory requirements.

In the long term, XAI should be seen not just as a technical add-on, but as an integral part of sustainable, ethical and student-centred educational AIS. This will not only ensure greater efficiency, but also help build trust and social responsibility in the processes associated with educational technology.

Acknowledgments

This article is supported by internal research project RD-08-109/05.02.2025 “Design and build a scalable research platform for Edge AI/HPS analytics”, Konstantin Preslavsky University of Shumen, Faculty of Technical Sciences, Department of Communication and Computer Engineering and Security Technologies.

References:

[1] Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. EDUCAUSE Review, 46(5), 30–40. (https://er.educause.edu/articles/2011/9/penetrating-the-fog-analytics-in-learning-and-education) (visited on 13.10.2025).

[2] Pane, J. F., Steiner, E. D., Baird, M. D., & Hamilton, L. S. (2017). Informing progress: Insights on personalized learning implementation and effects. RAND Corporation. (https://www.rand.org/pubs/research_reports/RR2042.html) (visited on 13.10.2025).

[3] Rudin, C. (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://arxiv.org/abs/1811.10154.

[4] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160. https://arxiv.org/abs/1802.01933.

[5] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4765–4774). https://arxiv.org/abs/1705.07874.

[6] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 93. https://arxiv.org/abs/1802.01933.

[7] Fancsali, S. E., Li, H., & Ritter, S. (2018). Towards an Early Warning System for At-risk Students Using Explainable Machine Learning. Proceedings of the 11th International Conference on Educational Data Mining, 1–10. International Educational Data Mining Society. https://doi.org/10.5281/zenodo.3554740.

[8] Pane, J. F., Steiner, E. D., Baird, M. D., Hamilton, L. S., & Pane, J. D. (2017). Informing Progress: Insights on Personalized Learning Implementation and Effects. RAND Corporation. https://doi.org/10.7249/RR2042.

[9] Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning.(http://fairmlbook.org) (visited on 13.10.2025).

[10] Bauer, K. (2023). Expl(AI)ned: The impact of explainable AI on trust in educational systems. INFORMS Journal on Education Technology. (https://madoc.bib.uni-mannheim.de/65911/1/bauer-et-al-2023-expl%28ai%29ned-the-impact-of-explainable-artificial-intelligence-on-users-information-processing.pdf) (visited on 13.10.2025)

[11] Poikola, A., Kuikkaniemi, K., & Honko, H. (2020). MyData – An introduction to human-centered personal data management. MyData Global. (https://mydata.org/wp-content/uploads/2020/08/mydata-white-paper-english-2020.pdf) (visited on 13.10.2025).

[12] Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable. (https://christophm.github.io/interpretable-ml-book/) (visited on 13.10.2025).

[13] Williamson, B., & Eynon, R. (2020). Historical threads, missing links, and future directions in AI in education. Learning, Media and Technology, 45(3), 223–235. https://doi.org/10.1080/17439884.2020.1798995.

[14] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (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.

[15] Gunasekara, S., & Saarela, M. (2025). Explainable AI in Education : Techniques and Qualitative Assessment. Applied Sciences, 15(3), Article 1239. https://doi.org/10.3390/app15031239.

[16] Atanasov, V. (2024). An approach to support web application development using cognitive machine. Yearbook of Shumen University "Bishop Konstantin Preslavsky", Vol. XIV F. Bishop Konstantin Preslavsky Publishing House. ISSN: 1314-8818.

[17] Atanasov, V, T., Transposition issues in digital learning process, Conference proceedings, vol. 1, Konstantin Preslavsky University Press, 2020, pp. 117 - 124, ISSN: 1314-3921.