August 2025

Multi-Modal XAI Framework for Trustworthy Higher Education Applications

Zaid M. Altukhi
University of Technology Sydney, zaid.m.altukhi@student.uts.edu.au

Sojen Pradhan
University of Technology Sydney, sojen.pradhan@uts.edu.au

Multi-Modal XAI Framework for Trustworthy Higher Education Applications

Emergent Research Forum (ERF) Paper

Zaid M. Altukhi
University of Technology Sydney
Zaid.m.altukhi@student.uts.edu.au

Sojen Pradhan
University of Technology Sydney
sojen.pradhan@uts.edu.au

Abstract

In this paper, we propose a framework that integrates XAI methods in higher education applications. The framework aims to provide clear, transparent, trustworthy AI-based systems. The framework we developed addresses different challenges that hinder the integration of XAI in education and utilising Design Science Research (DSR) to ensure the research rigour through the lens of affordance theory to examine the interaction between the AI-based application and the educators as end-users.

Keywords

XAI, Higher Education, Framework, Student at-risk.

Introduction

Artificial Intelligence (AI) is playing a significant role in higher education (HE), particularly in identifying at-risk students. The problem we are exploring to address is that despite the high predictive accuracy of using advanced machine learning (ML) models known as black-box models, their lack of explainability represents an obstacle in the adoption of AI-based systems by educators (Swamy et al., 2022). Non-technical stakeholders such as educators need transparent and interpretable explanations that enable them to understand, trust, and act on AI predictions (Maity & Deroy, 2024).

Using Explainable AI (XAI) methods to explain the black-box models is one of the effective solutions. Swamy et al. (2022) claim that employing XAI techniques enables educators to gain meaningful insights into the inner workings of black-box models, thereby improving trust and adoption in the education sector. However, there are some challenges in integrating XAI methods into educational systems.

Affordance Theory (AT) is used to reduce those challenges while formulating a novel framework that will be developed and carefully evaluated. The AT examines how educators perceive the explanations and the actionable possibilities provided by the XAI system. The perspectives from AT will be used to mitigate the challenges, including explainability, Human Computer Interaction (HCI), trustworthiness and policy and guidelines in the proposed framework.

Our proposed framework shows ML predictions and provides multi-modal explanations of AI-based application results for HE educators. Local and global explanations will be used in our framework. Local explanations concentrate on explaining individual predictions made by an ML model for specific data instances, whereas global explanations offer a comprehensive understanding of a model's behaviour across the entire dataset (Aechtner et al., 2022). By integrating XAI techniques: *local* (e.g., Shapley Additive exPlanations (SHAP), Local Interpretable Model Agnostic (LIME), and textual narratives) and *global* (e.g., SHAP and partial dependence plots), the framework allows educator to understand and engage with the model's decision-making process. By using the Design Science Research (DSR) methodology, we will validate the outputs from the system. The main purpose of the framework is to offer transparent and trustworthy explanations of the results of the AI-based applications in HE sector.

Background

Integrating XAI in HE is essential for ensuring that AI-based decisions can be transparently communicated to non-expert users. Prior studies have highlighted the importance of bridging the gap between sophisticated ML algorithms and end-user comprehension (Severes et al., 2023). Methods such as SHAP and LIME have been applied in various AI-based HE applications to help improve transparency and interpretability (Huang, 2023; Melgar-García et al., 2023). However, these methods have some negative impacts on different educational stakeholders (Maity & Deroy, 2024; Swamy et al., 2023), and there is a need to do more research to validate the effectiveness of these methods (Saarela & Podgorelec, 2024).

In HE, identifying students at risk refers to the process of using predictive analytics and AI techniques to early identify students who may be struggling academically or at risk of dropping out (Albreiki et al., 2022). It enables educators to provide timely support before a student fails or drops out (Al-Gahmi, 2024). Several studies utilised XAI techniques to help identify at-risk students by analysing educational data and academic history (Alwarthan et al., 2022; Ujkani et al., 2024). Although studies indicate potential benefits of using XAI for identifying at-risk students, they emphasise the importance of developing context-specific solutions (Al-Gahmi, 2024; Albreiki et al., 2022). Some researchers employed multiple ML models to improve prediction accuracy in order to find the best performance (Raji et al., 2024; Susnjak, 2024).

In our prior review Altukhi and Pradhan (2024), we identified six key challenges when integrating XAI into HE applications: (1) Explainability that address the need to provide clear explanations (2) Ethical Considerations to ensure non-bias and fairness by revealing potential biases (3) Technical Constraints by addressing the integration of XAI methods into existing educational systems (4) HCI by developing user-friendly, web-based applications for non-technical users (5) Trustworthiness, which combines high-accuracy predictions with interpretable explanations to build user confidence (6) Policy and Guidelines to create frameworks that comply with educational policies, which can be adapted across many HE contexts.

Affordance Theory (AT)

AT will guide our development and evaluation of how educators interact with and perceive the explanations that would be provided by the system designed as per the proposed framework. In the realm of Information Systems, AT will provide a means to enhance the interactions between users and AI-based applications (technology), demonstrating how new features contribute to achieving better outcomes (Leonardi, 2013). AT guides the selection and design of explanation types within the framework by aligning them with the actionable possibilities educators perceive when interacting with AI-based systems. We have selected a combination of textual, visual, and rule-based explanation modalities to reflect the diverse interpretive needs and affordances of non-technical users. Textual explanations leverage the Large Language Models (LLMs) to align with educators' communication preferences. Visual methods (SHAP, LIME) support identifying pattern recognition through graphical affordances. Rule-based explanations (Anchors) enable logical reasoning behind the ML decision. This affordance-driven design ensures that explanations are not only interpretable but also practically useful for educators in real-world decision-making processes.

The framework will leverage UI components to enhance user interaction and comprehension. Features such as interactive widgets (e.g., buttons, and checkboxes), layout elements (e.g., columns and sidebars), and real-time responsiveness afford users the ability to engage with the system dynamically. These components support the principles of AT by providing clear, actionable guidance that guides users in exploring and interpreting model predictions effectively. This design approach ensures that the system's functionalities are perceivable and actionable, fostering a more intuitive and trustworthy user experience.

Proposed Framework

Our proposed framework comprises three major components and sub-components as shown in Figure 1: 1) Data Preparation: all the necessary work on data preprocessing, model training, and model selection. 2) XUI: predictions, risk assessments, graphical representations, and multiple explanations generation. 3) Educators Interaction: end-user role to perceive the affordance system's features, interaction with the system and feedback in order to improve the system.

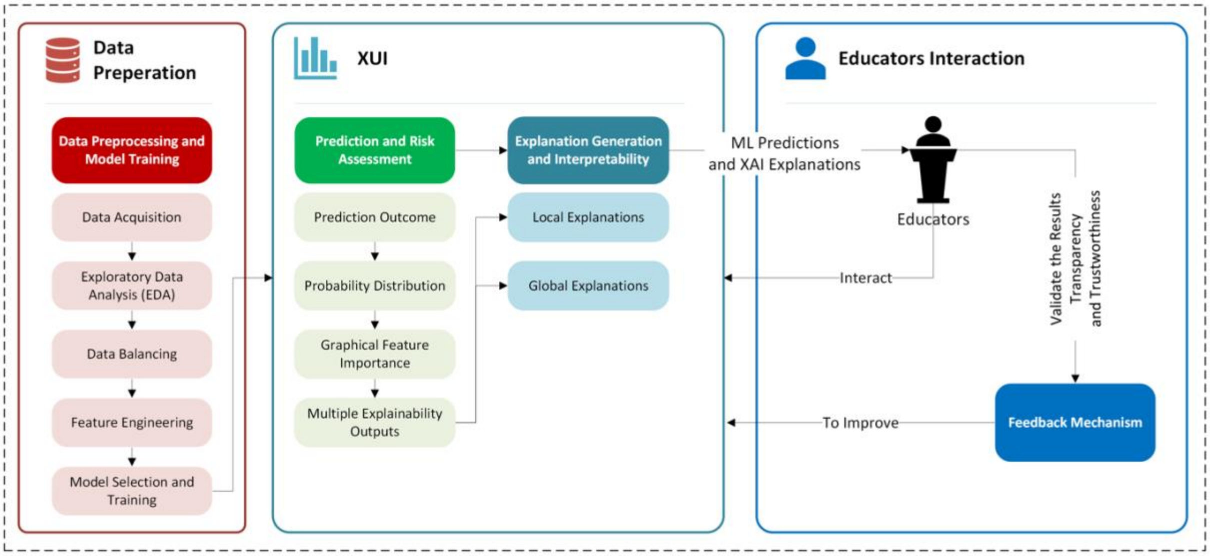

Figure 1. Proposed Trustworthy HE Framework Components.

Data Preparation

The majority of learning analytics research relies on structured data (Ali, 2024; Prinsloo & Slade, 2018), and the framework is also built upon structured educational datasets. This phase consists of several key steps: Data Acquisition involves the collection of relevant HE datasets. Exploratory Data Analysis (EDA) familiarises the data, examines distributions, and identifies feature imbalances. Data Balancing mitigates these imbalances. Feature Engineering selects and transforms informative predictors. Model Selection and Training train multiple classifications of ML models and evaluate their performance. To select the most suitable classifier for integration into the proposed framework, we will evaluate multiple ML models using standard comparative metrics, including accuracy, precision, recall, and F1-score. The model will be selected based on its performance across the comparative metrics and its suitability for generating reliable and interpretable explanations for educators.

eXplainable User Interface (XUI)

XUI is an emerging area of research within the broader field of XAI, focusing on the design and implementation of UI that effectively communicate AI explanations to end-users (Ridley, 2022). It aims to help users better understand AI application results by creating intuitive and interactive interfaces that explain how the AI system makes decisions (Rahimi et al., 2025). Furthermore, XUI strives to simplify the complex algorithms behind AI, ensuring that these powerful technologies are more accessible and relatable to more people. Predictions and Risk Assessment employ ML-trained model in the previous component to predict student outcomes (e.g., Pass, Fail or Dropout) and compute probability scores of each class. The outcomes of the model will be deployed within a web-based application, allowing educators to input student data and receive prediction outcomes that show the classification of student's risk, a probability distribution for each class, feature importance visualising the rankings of influential features on the predictions, and multiple explainability outputs, illustrating different explanations through various XAI methods.

In the Explanation Generation and Interpretability, several local explanation methods utilised, including textual, visual, and rule-based methods. It empowers end-users to select their preferred LLM (from different options: GPT-4, LLaMA, or Mistral, for human-readable narratives that describe the contribution of each input feature to the model's prediction) to generate textual explanations. This flexibility ensures adaptability to different deployment contexts and supports both accessibility and reproducibility. The visualisation methods include LIME and SHAP. LIME provides a model-agnostic, feature-level explanation with bar charts showing positive and negative contributions to the prediction outcome. The SHAP is used to illustrate how each feature shifts the prediction. For the rule-based explanations, we perform Anchor

Explanations, which offer if-then decision rules to pinpoint the conditions leading to a specific classification. To get a holistic overview, we employed several global explanation methods: Feature Importance presents a bar chart of the most influential features across the entire dataset. SHAP provides a global overview of feature impacts across all predictions. The Partial Dependence Plots (PDP) visualises how varying a single feature influences the prediction outcome, including interaction effects.

Educators Interaction (and Feedback Mechanism)

The educator's interaction in our framework consists of receiving, interacting, and validating the ML, XAI results, and XAI-based system usability. To evaluate the effectiveness of the proposed framework, we will adopt a multi-dimensional assessment strategy that aligns with our research objectives. This strategy includes both qualitative and quantitative metrics under key evaluation dimensions: design and usability, engagement and interaction, explanation needs, explanation quality, system-level trust, and overall trust and understanding. These dimensions will guide the design of interview questions and structured feedback instruments to assess the system's interpretability, usability, and trustworthiness from the perspective of HE educators. To ensure the framework aligns with user needs, educator feedback will be integrated into iterative refinement cycles following the principles of the DSR methodology. Feedback obtained from interviews and structured evaluation tools will be analysed thematically to identify usability, trust, and explanation-related concerns. These insights will drive targeted improvements in both the user interface and the explanation mechanisms. Each iteration will be validated with educators to ensure the framework is practical, trustworthy, and enhances interpretability. Additionally, to support a more rigorous assessment of textual explanations, further validation efforts may incorporate quantitative metrics proposed by Ichmoukhamedov et al. (2024), which evaluate textual explanations across dimensions such as coherence, relevance, and interpretability from a user-centred perspective.

Conclusion and Future Work

This paper presents the progress of developing and implementing a multi-modal XAI framework for predicting student at-risk in HE. By combining advanced ML techniques with state-of-the-art XAI explainability methods, our approach addresses the dual challenge of achieving high predictive performance and offering transparent explanations. Future work will focus on conducting comprehensive user studies to evaluate the framework's impact on educators' decision-making and trust in AI systems. Future research can also expand the framework to support a wider range of educational stakeholders, data types, including unstructured data, and prediction tasks. Our work aims to bridge the gap between complex AI models and practical, trustworthy applications in educational settings, ultimately supporting students' success through informed intervention.

REFERENCES

Aechtner, J., Cabrera, L. M., Katwal, D., Onghena, P., Valenzuela, D. P., & Wilbik, A. (2022). Comparing User Perception of Explanations Developed with XAI Methods. *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1-7.

Al-Gahmi, A. (2024). On the Portability and Robustness of Early Student Performance Predictions. *2024 ASEE Annual Conference & Exposition Proceedings*.

Albreiki, B., Habuza, T., & Zaki, N. (2022). Framework for automatically suggesting remedial actions to help students at risk based on explainable ML and rule-based models. *International Journal of Educational Technology in Higher Education*, *19*(1), 49.

Ali, E. S. G. (2024). The Effect of Educational Guidelines on Late Adolescent Female's Knowledge and Attitude Regarding Human papilloma Virus. *Helwan International Journal for Nursing Research and Practice*.

Altukhi, Z. M., & Pradhan, S. (2024). Systematic Literature Review: Explainable AI Definitions and Challenges in Education.

Alwarthan, S., Aslam, N., & Khan, I. U. (2022). An explainable model for identifying at-risk student at higher education. *IEEE Access*, *10*, 107649-107668.

Embarak, O. H., & Hawarna, S. (2024). Automated AI-driven System for Early Detection of At-risk Students. *Procedia Computer Science*, *231*, 151-160. https://doi.org/https://doi.org/10.1016/j.procs.2023.12.187

Huang, K. (2023). Automatic Analysis of Peer Feedback using Machine Learning and Explainable Artificial Intelligence. In.

Ichmoukhamedov, T., Hinns, J., & Martens, D. (2024). How good is my story? Towards quantitative metrics for evaluating LLM-generated XAI narratives.

Leonardi, P. M. (2013). When does technology use enable network change in organizations? A comparative study of feature use and shared affordances. *MIS quarterly*, 749-775.

Maity, S., & Deroy, A. (2024). Human-Centric eXplainable AI in Education. *arXiv preprint arXiv:2410.19822*.

Melgar-García, L., Troncoso-García, Á., Gutiérrez-Avilés, D., Torres, J. F., & Troncoso, A. (2023). Explainable Artificial Intelligence for Education: A Real Case of a University Subject Switched to Python. Computational Intelligence in Security for Information Systems Conference,

Prinsloo, P., & Slade, S. (2018). ' s repository of research publications and other research outputs Building the Learning Analytics Curriculum : Should we Teach (a Code of) Ethics ? Conference or Workshop Item.

Rahimi, M., Yaghoobzadeh, Y., Department, M. R. D. B. E., Engineering, S. o. E., Science, I. U. o., Technology, Tehran, Iran, Electrical, Department, C. E., Tehran, U. o., Sciences, S. o. M., Sciences, I. f. R. i. F., Studies, T. I. f. A., & University, K. (2025). Explanations of Deep Language Models Explain Language Representations in the Brain.

Raji, N., Kumar, R. M. S., & Biji, C. (2024). Explainable Machine Learning Prediction for the Academic Performance of Deaf Scholars. *IEEE Access*.

Ridley, M. (2022). Explainable Artificial Intelligence (XAI). *Information Technology and Libraries*, *41*(2). https://doi.org/10.6017/ital.v41i2.14683

Saarela, M., & Podgorelec, V. (2024). Recent Applications of Explainable AI (XAI): A Systematic Literature Review. *Applied Sciences*.

Severes, B., Carreira, C., Vieira, A. B., Gomes, E., Aparício, J. T., & Pereira, I. (2023, 2023). The Human Side of XAI: Bridging the Gap between AI and Non-Expert Audiences.*SIGDOC '23*

Susnjak, T. (2024). Beyond predictive learning analytics modelling and onto explainable artificial intelligence with prescriptive analytics and ChatGPT. *International Journal of Artificial Intelligence in Education*, *34*(2), 452-482.

Swamy, V., Du, S., Marras, M., & Kaser, T. (2023). Trusting the explainers: teacher validation of explainable artificial intelligence for course design. LAK23: 13th International Learning Analytics and Knowledge Conference,

Swamy, V., Radmehr, B., Krco, N., Marras, M., & Käser, T. (2022). Evaluating the Explainers: Black-Box Explainable Machine Learning for Student Success Prediction in MOOCs. *ArXiv*, *abs/2207.00551*.

Ujkani, B., Minkovska, D., & Hinov, N. (2024). Course Success Prediction and Early Identification of At-Risk Students Using Explainable Artificial Intelligence. *Electronics*, *13*(21), 4157. https://www.mdpi.com/2079-9292/13/21/4157