Journal of Information Technology Education: Innovations in Practice

An Official Publication of the Informing Science Institute
InformingScience.org

JITEiip.org

Volume 25, 2026

LEVERAGING EXPLAINABLE AI TO ENHANCE STUDENT METACOGNITION THROUGH EARLY RISK DETECTION

Saadia Malik	Department of Information Systems, Faculty of Computing and Information Technology, Rabigh, King Abdul-Aziz University, Jeddah, Saudi Arabia	Smalik1@kau.edu.sa
Muhammad Hamid*	Department of Computer Science, Government College Women University, Sialkot, Pakistan	mhamid@gcwus.edu.pk
Muhammad Saleem	Department of Industrial Engineering, Faculty of Engineering, Rabigh, King Abdul-Aziz University, Jeddah, Saudi Arabia	msaleim1@kau.edu.sa

*Corresponding author

ABSTRACT

Aim/Purpose	Early identification of at-risk students is a burning issue in higher education. Although conventional Machine Learning (ML) models have high predictive accuracy, they tend to be opaque black boxes and offer no diagnostic information. This paper aims to fill this diagnostic gap by developing an eXplainable AI (XAI)-based framework to convert technical risk scores into actionable prompts for students' self-regulation and reflection.
Background	Conventional academic assistance is reactive, meaning it is provided after failure has occurred. Although ML enables proactive identification, its lack of transparency prevents educators from offering specific support. This paper introduces a framework to improve student success by integrating explainability and fairness beyond classification, aligning technical AI performance with pedagogical objectives.
Methodology	Three ML models, Logistic Regression, Random Forest, and XGBoost, were used to analyze a refined sample of 278 student records. The methodology entailed a preprocessing pipeline of data. SHAP (Shapley Additive Explanations) was incorporated to support both global and local interpretability, whereas a

Accepting Editor Stamatis Papadakis | Received: September 21, 2025 | Revised: January 26, February 6, February 15, 2026 | Accepted: February 28, 2026.
Cite as: Malik, S., Hamid, M., & Saleem, M. (2026). Leveraging explainable AI to enhance student metacognition through early risk detection. *Journal of Information Technology Education: Innovations in Practice, 25,* Article 19. https://doi.org/10.28945/5739

	formal Fairness Audit was performed to guarantee that risks were fairly detected across gender groups.
Contribution	The research introduced a novel human-focused framework that fills the gap between predictive analytics and pedagogical theory. It shows how XAI can help turn a technical risk score into a metacognitive prompt, encouraging data-driven conversations between educators and students and offering a clear roadmap for institutional interventions.
Findings	The analysis revealed that the Random Forest model achieved 92.9% accuracy and an AUC-ROC of 0.977. SHAP analysis found school absenteeism and mid-term grades as the most important risk predictors. Individualized diagnostics (waterfall plots) in the system provided the necessary evidence through student self-reflection, and the audit of fairness ensured that the model supports gender groups equally.
Recommendations for Practitioners	Educational institutions must implement risk prediction systems based on XAI to go beyond mere warning systems. Practitioners should employ individual-level diagnostics to tailor mentoring and motivate students to reflect on their learning behaviors through evidence-based self-reflection.
Recommendations for Researchers	Further studies are expected to include longitudinal pilot studies that quantify the actual behavioral effects of XAI-based prompts on student outcomes. Researchers are also advised to test the framework using larger, multi-institutional datasets to increase its generalizability.
Impact on Society	Transparent and fair ML systems can improve student retention and graduation rates, leading to better resource allocation and a more inclusive educational environment. By focusing on student agency, these systems foster a more successful, self-aware workforce that benefits society in the long term.
Future Research	To establish the global applicability and ethical soundness of the XAI framework, future studies must explore real-time application of XAI to learning management systems and test the cross-cultural validity of behavioral predictors.

Keywords	early risk detection, explainable artificial intelligence (XAI), fairness in AI, machine learning in education, metacognition, student retention

INTRODUCTION

The main objective of higher education is not merely to provide students with knowledge but also to ensure that they complete their academic journey. Retention of students and academic achievement have become performance indicators at universities across the world. Traditionally, education support was reactive; students were assisted after they failed an exam or dropped out. However, to enhance student performance, institutions have adopted proactive approaches that identify struggling students at an early stage, when there is still time for meaningful intervention (Algarni et al., 2023). This proactive approach requires a profound understanding of student behavior and the factors that influence success, including socio-economic status and prior academic achievements.

To achieve this early identification, the field of education has shifted towards data-driven technologies. Educational Data Mining (EDM) and ML are now commonly employed to analyze large masses of student data to predict academic risk (Feng et al., 2022; Hamid et al., 2026; Khine, 2024). XGBoost, Logistic Regression, and Random Forest algorithms have demonstrated high mathematical accuracy in identifying at-risk students (Chong et al., 2025). By training models on past student records, institutions can now produce risk scores before the final exams. Although a shift towards

automated ML prediction has enhanced the speed of identification, it has also presented new challenges in terms of the transparency and practical application of such models in a classroom environment.

The biggest issue with the existing ML solutions in education is that they act as Black Boxes. However, these models can effectively forecast that a student is at risk, but they are unable to tell why the prediction was made (Gunasekara & Saarela, 2025). Prediction without explanation is a major limitation in an educational context. For example, a teacher may be alerted that a student is likely to fail, but without knowing whether the risk is due to high absenteeism, poor mid-term grades, or study time absence, the teacher cannot design a suitable intervention. This results in a diagnostic gap in which the AI issues an alarm but provides no roadmap for support. Moreover, being labeled at-risk by an unknown algorithm may be stigmatizing to students, leading to a loss of motivation rather than behavior change.

This results in a significant research gap: the lack of integration between AI and Metacognition, the capacity of students to monitor and control their own learning process (Kara et al., 2024). Based on earlier theories by Flavell (1979) and Zimmerman (2000), students learn effectively when they can reflect on their behavior and modify it. This is not supported by traditional ML models since their logic is not visible to the student. This is where Explainable AI (XAI) is necessary. XAI is more transparent, unlike typical ML, as it displays the aspects or behaviors that contributed to a risk flag. The existing research has not fully investigated the use of XAI as a Metacognitive Prompt to assist students in comprehending their specific learning gaps. Without XAI, we cannot transform a technical risk score into a pedagogical tool that stimulates self-reflection for student development.

To bridge this gap, this paper proposes an XAI-based predictive framework that balances high accuracy with human-interpretable explanations. With a Random Forest model, we not only identify at-risk students but also provide a list of contributing factors that contribute to the risk. The transparency enables our model to serve as a Metacognitive Catalyst, turning the intervention process into a data-driven conversation between the educator and the student rather than a generic warning. By revealing the why behind a prediction, the system enables students to self-regulate and take informed measures, aligning technical AI performance with the main objectives of education.

The primary contributions of this study are:

- The study develops a high-fidelity predictive framework that uses ML algorithms to identify at-risk students early. Through benchmarking various classifiers, the paper finds the random forest as the best model for identifying academic risk with high precision, moving beyond a manual monitoring system to a scalable, data-driven identification system.
- An important contribution is the inclusion of an XAI layer that breaks down the black-box nature of traditional models. By applying SHAP analysis to provide global and local interpretability, the framework transforms the at-risk flag into a diagnostic tool. This enables the system to uncover the specific behavioral and academic drivers behind every prediction, providing the necessary 'why' for effective intervention.
- The research integrates a fairness audit as an essential ethical element in the methodology. Using Demographic Parity Difference to measure and reduce gender-based algorithmic bias, the study ensures that the predictive results are fair. This contribution will provide a guideline for responsibly applying AI in the education sector, ensuring every student group has equal and unbiased access to support.
- Ultimately, this research facilitates a shift from reactive institutional monitoring to proactive student agency. By turning technical explanations into metacognitive prompts, the framework empowers students to engage in self-reflection and self-regulation. The primary impact is the creation of a human-centered dialogue between educators and learners, where AI serves as a pedagogical catalyst for long-term metacognitive growth and academic success.

The rest of the paper is organized as follows. The next section presents a thematic literature review that synthesizes the background. The research methodology is then described, including the dataset characteristics, the ML algorithms, and the ethical fairness audit procedure. The results of the experiment, the comparison of models, and the analysis of feature importance are then described. The paper then concludes the study and gives directions for future research.

LITERATURE REVIEW

The rapid development of EDM and Learning Analytics (LA) has fundamentally altered higher education institutions' ability to track and predict student academic outcomes. In the past, institutions used reactive measures, but transitioning to data-driven decision-making has enabled proactive intervention. Studies (e.g., Algarni et al., 2023; Jiang et al., 2024) have shown that machine learning models, including Random Forests and Support Vector Machines, are capable of high classification accuracy in predicting student success. Although this advancement has been made, there is still a major limitation: these studies are largely based on mathematical performance measures, and in most cases, the underlying behavioral, psychological, and non-academic factors are not adequately considered in understanding student persistence.

The use of AI in instructional models is no longer a choice, as Aldowah et al. (2019), Khine (2024), and Kong et al. (2021) claim that this approach is essential in the modern educational process. The online and distance learning environment, which has now become the most popular form of instruction, needs more than just an efficient design; it needs a sense of student engagement. Chong et al. (2025) emphasize that individual engagement patterns and learning styles are more important than instructional design in driving student success in virtual environments. To address these complex needs, Moreno-Marcos et al. (2020) argue that EDM and LA are essential fields that enable educators to uncover patterns previously hidden in educational datasets. These patterns that would otherwise be hidden are obtained using advanced statistical and machine learning methods that enable a transformation of raw data into actionable pedagogical strategies.

But the quest for greater precision has led to the proliferation of complex automated systems. Aldowah et al. (2019) caution that with the growing sophistication of models towards deep learning architectures as used by Al-Shabandar et al. (2019), there is a risk that they lose their theoretical connection to the well-established learning sciences. Williamson and Eynon (2020) highlight this issue, noting that the social, ethical, and power dynamics of student data use must be addressed when developing long-term learning analytics. Moreover, although Waheed et al. (2020) and Latif et al. (2023) used large interaction logs to predict performance, their models tend to provide context-specific outcomes that are hard to extrapolate to other institutional cultures or demographic groups. Although predictive frameworks can be technically sound, they seldom undergo formal technical audits to address algorithmic bias, a major ethical issue related to fairness and fair support (Al-Shabandar et al., 2019).

One of the main criticisms of existing ML-based interventions is the black-box nature of these interventions. The existing systems, as reported in Al-Shabandar et al. (2019), Iranzad and Liu (2025), and Embarak and Hawarna (2024), provide risk scores but do not elaborate on the reasoning behind them, which limits the possibility of targeted mentoring. Educators cannot provide the roadmap that struggling students need when they cannot explain why an algorithmic prediction works the way it does. It is in this diagnostic gap that the field of Explainable AI (XAI) is needed. Authors (Fiok et al., 2022; Gunasekara & Saarela, 2025; Lukyanenko et al., 2021) argue that transparency is the cornerstone of trust in educational AI, yet Pitts et al. (2025) and Chong et al. (2025) point out that many current XAI applications stop at simply building trust rather than functioning as practical tools for student development.

The proposed framework (Flavell, 1979; Zimmerman, 2000) is based on the theory of metacognition. It was determined that metacognitive monitoring and self-regulation are critical to academic

development. Students can learn when they have the capacity to reflect on their behaviors and correct them. Nevertheless, modern AI systems often do not undergo the reflection phase, since the decision-making logic of an algorithm is never disclosed to the learner.

According to Kara et al. (2024), engagement is mainly driven by personality traits and self-regulation, but studies such as Embarak and Hawarna (2024), although novel, do not include a direct feedback loop to enhance student self-awareness. Recent studies by Yang and Xia (2023), Lee et al. (2025), and Latif et al. (2023) reaffirm the need for such a feedback loop, suggesting that AI should be used as a scaffold for metacognition calibration rather than a monitoring tool.

Additionally, the advent of Generative AI in a short period of time has brought new challenges to the educational domain. Although Baidoo-Anu and Ansah (2023), Wu (2023), and Vieriu and Petrea (2025) report on the difficulties of AI tools such as ChatGPT, they all focus on the idea that AI should be used to facilitate academic agency among students instead of automating the education process to the extent of involving students. The study by Rienties et al. (2023) fairly observes that tools should be socially valid in the classroom; otherwise, they may be too complex or not transparent enough, making them unavailable to non-expert users, such as students.

A critical analysis of this literature shows a two-fold gap. First, there is a technical gap: high-accuracy predictive models are not interpretable enough for pedagogical action. Second, predictive analytics are not theoretically connected to the underlying processes of metacognitive development. The majority of current systems are black boxes that generate alarms but offer no diagnostics, leaving educators without data-based improvement paths. The lack of model-agnostic XAI frameworks that bridge the gap between high-fidelity risk prediction and the student's internal regulatory mechanisms is evident. This research fills these gaps by proposing an XAI-driven predictive system that includes ML transparency and metacognitive support.

METHODOLOGY

This study adopts a quantitative predictive modeling method grounded in EDM to specifically address the interpretability and metacognitive gaps in the previous literature (Khine, 2024). The methodology is structured to go beyond the black-box character of traditional ML models, which are frequently highly accurate but lack the diagnostic transparency required for pedagogical action. The framework aims to fulfill the theoretical needs of self-regulated learning proposed by Flavell (1979) and Zimmerman (2000). The implementation of XAI as an essential part of its framework and the transformation of technical risk scores into practical metacognitive prompts for students.

The proposed methodology, as shown in Figure 1, proceeds in a logical order, starting with data acquisition, a preprocessing stage with one-hot encoding, and a stratified train-test split to address class imbalance. This pipeline is then switched to a multi-model benchmarking phase, where Logistic Regression, Random Forest, and XGBoost are compared. Importantly, the methodology has built in an evaluation and fairness audit layer to address the ethical issues noted in past studies, within an outcome that is focused on student-teacher communication and metacognitive development.

DATASET AND VARIABLES

The experimentation phase of the study is built on the Student Performance Dataset, a benchmark-quality dataset downloaded from the UCI ML Repository. This dataset was chosen to ensure data quality and to focus on the academic and behavioral aspects of early risk identification. The original data set contained 395 records. However, to ensure high data integrity in the predictive model, rigorous filters were applied. Data with missing values for important behavioral attributes and students with incomplete academic records were removed. Thus, a clean set of 278 student records was obtained and used for modeling and interpretability analysis.

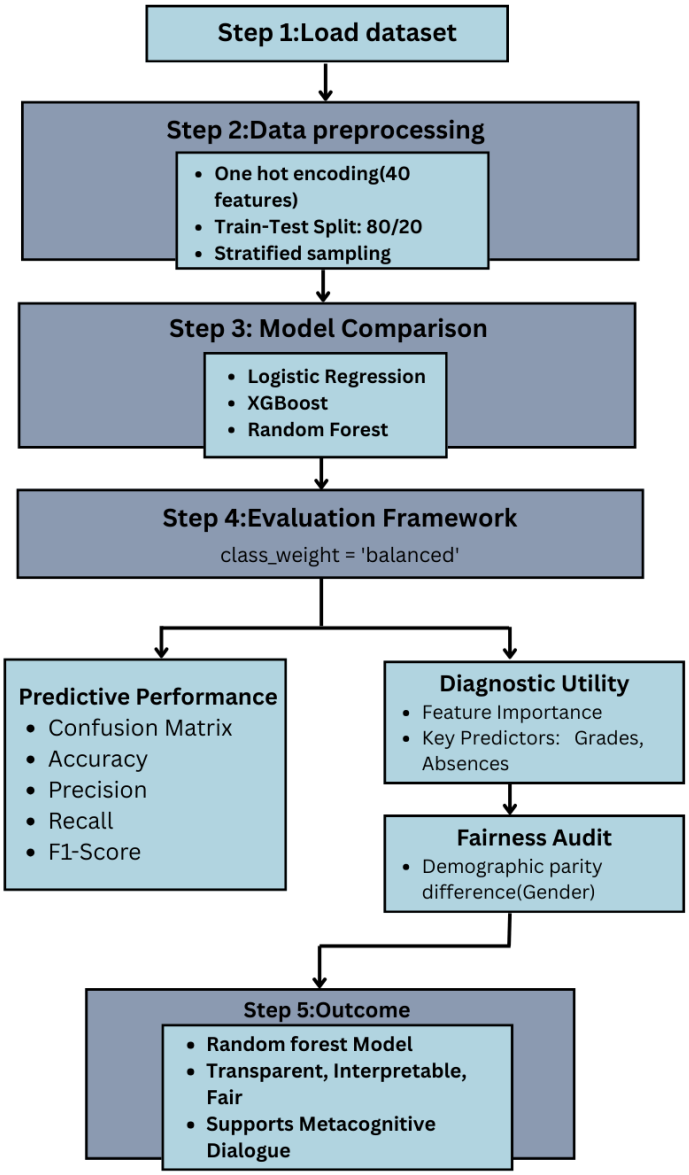


Figure 1. Proposed methodology for predicting at-risk students

The dataset will offer a comprehensive picture of students based on 33 attributes, which were grouped into four major categories to reflect the complexity of the learning scenario:

1. *Demographic Profiles:* Basic attributes such as Gender, Age, and Address_Type (Urban vs. Rural) that provide basic details of the student's background where he/she resides.
2. *Socio-economic and Family Background:* Strategic variables include Mothers_Education, Fathers_Education, and parents' occupations (Mothers_Job, Fathers_Job) and Family_Size.
3. *Academic Engagement:* The performance indicators include Weekly_Study_Time, Travel_Time_To_School, and access to Extra_Educational_Support.
4. *Behavioral and Lifestyle Indicators:* Behavioral indicators such as school absences, going out with Friends, and alcohol consumption, which are usually ignored in conventional performance models.

The target variable for the classification task is the binary variable At_Risk_Student, where 1 indicates At-Risk status, and 0 indicates non-At-Risk status. In accordance with ethical data-handling practices, the dataset used in this study is fully anonymized, with all personally identifiable information (PII), such as names and student IDs, removed at the source. It guarantees student privacy protection and the opportunity to analyze academic and behavioral patterns objectively.

One characteristic of this dataset is the imbalance in the classes, with about 32% of students at risk. To ensure reproducibility and reduce algorithmic bias, the methodology uses stratified sampling and the parameter class_weightbalanced during model training. To enable compatibility among ML models, all categorical string variables (e.g., parental jobs and school choice reasons) were encoded using one-hot encoding, increasing the feature space to 40 numerical predictors.

Dataset Characteristics and Exploratory Analysis

Before developing the AI model, an Exploratory Data Analysis (EDA) was conducted to understand the distribution and relationships among the data variables. This step is essential to ensure that the technical consequences of the AI model are appropriate for educational practice. The distribution of the target variable, At_Risk_Student, is shown in Figure 2. This indicates the class imbalance. In this data, 32 percent of students are at-risk (Class 1), and 68 percent are not (Class 0). This class imbalance aligns with the educational reality that a minority of students typically require intervention. To address this problem, stratified sampling and class weights were used to ensure the model remained sensitive to the at-risk minority class.

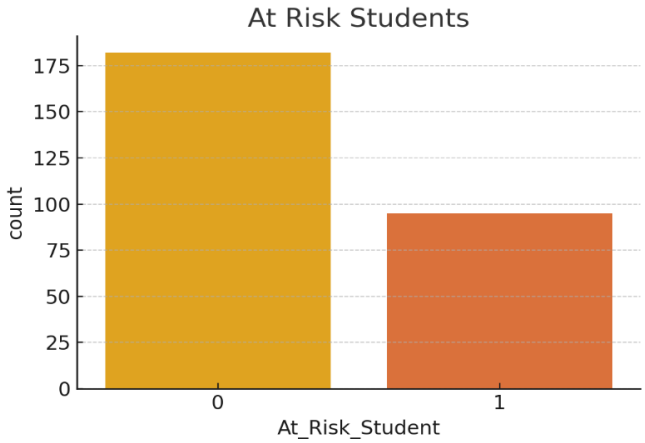


Figure 2. Distribution of the target variable (At_Risk_Student)

Figures 3(a), 3(b), and 3(c) give the demographic profile of the sample of students. The data shows a diverse sample with an age range of 15 to 22 years, predominantly residing in urban areas (U) with a fairly even split between genders. A preliminary analysis of behavioral variables indicates a strong correlation between social lifestyle and academic performance. For example, students with higher values for school absences and social activities are more frequently flagged as at risk. This is an important educational insight as it suggests that a decline in class engagement often precedes academic failure. Early detection of these patterns allows the framework to serve as a metacognitive prompt, prompting students to reflect on their time-management habits before their grades suffer.

Academic predictor analysis shows very strong and consistent patterns. Figure 4 shows a distinct relationship between first- and second-period grades, with at-risk students consistently falling in the lower-performance quadrant. Also, the visual heatmap (Figure 5) confirms the strong negative correlation between the final academic status and grade periods, and the strong positive correlation (0.8) between the two grade periods themselves. This means that past performance is a very reliable indicator of future risk. Non-academic factors in the heatmap, such as Past Failures and Mother's

Education, demonstrate that a student's risk is not entirely determined by their current marks but also depends on their socio-economic background. These correlations provide the necessary context for the XAI phase, ensuring that the model's eventual "explanations" are grounded in the actual behavioral patterns of the students.

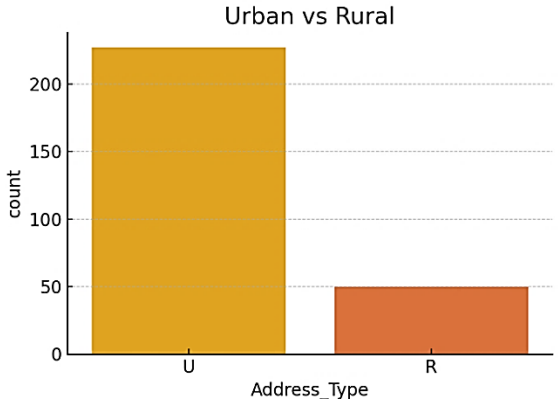


Figure 3(a). Demographic profile of the student sample showing distribution by address type

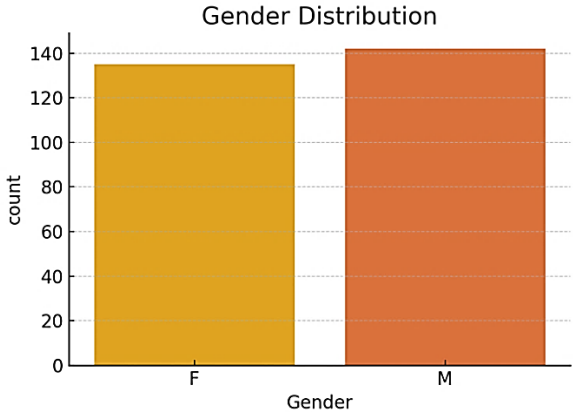


Figure 3(b). Demographic profile of the student sample showing distribution by gender

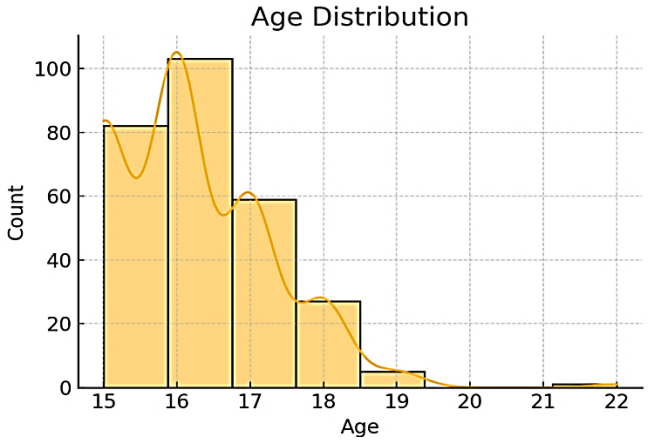


Figure 3(c). Demographic profile of the student sample showing distribution by age

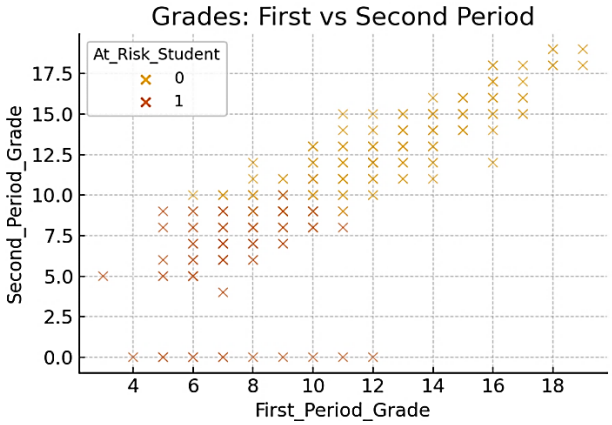


Figure 4. Correlation between first and second period grades

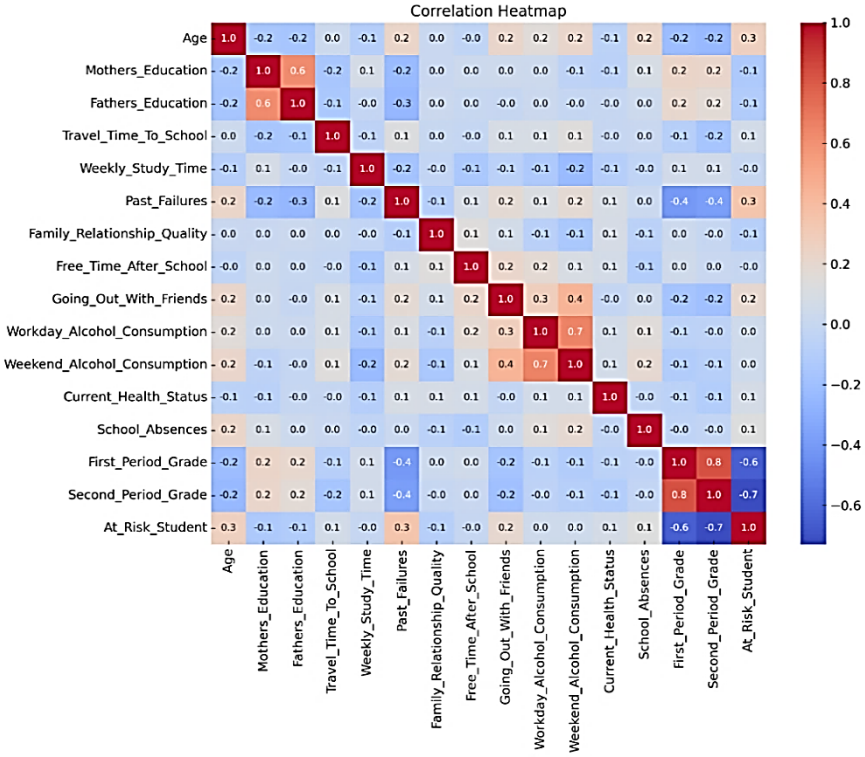


Figure 5. Correlation matrix of key numerical variables

Data Preprocessing

To ensure the integrity of the predictive model and the quality of the following diagnostic insights, a data preprocessing was developed. This process was designed to transform the raw values of students' academic and behavioral data into a format that can be used by ML algorithms. Our data comprised 33 attributes, and the majority of these attributes were categorical (e.g., parent's profession, reason for choosing the school, and urban or rural). We used one-hot encoding for all non-ordinal features to ensure these variables were represented in a form algorithms could use and interpret while preserving their meaning. As a result, a new feature set of 40 numerical variables was obtained. This is important because it enables the models to learn the impact of different levels of categorical variables, such as whether a student has "extra educational support", which is an important feature for early academic risk prediction. The data was split into two sets – training (80%) and testing (20%) –

to ensure a fair and rigorous evaluation of the predictive power of the models. One challenge in educational data is the underrepresentation of at-risk students (32% in this study). Stratified sampling was adopted to reduce the chances of sampling bias during the split. This guarantees that the ratio of students who are at-risk (Class 1) to those who are not at-risk (Class 0) is the same across the training and test subsets. This regular distribution compels the model to acquire the characteristics of the minority class well, eliminating the risk of the model becoming biased against the majority group, a frequent breakdown in conventional black box systems.

Predictive Modeling and Algorithm Selection

Three different classification models were benchmarked to find the best tool for early risk identification. This is a common practice in EDM, where model selection is justified by performance and interpretability. All algorithms were chosen to reflect a certain degree of architectural complexity:

1. *Logistic Regression:* Logistic regression is a linear model that estimates the probability of a binary outcome. It was trained with L2 regularization and the liblinear solver to ensure it converged stably on the 278-record cohort. The simplicity and direct interpretability of logistic regression make it extremely useful in educational tasks.
2. *Random Forest:* The random forest is an ensemble ML algorithm, which employs bootstrap aggregating (bagging) to create a number of decision trees. In this research, the model was optimized with 100 estimators and a maximum tree depth of 10 to avoid overfitting while maintaining accuracy. Its main advantage is its ability to compute feature importance, providing the diagnostic transparency required for a metacognitive framework. The Random Forest model goes beyond simple classification to offer actionable insights by discovering which academic or behavioral factors drive a prediction (Iranzad & Liu, 2025).
3. *Extreme Gradient Boosting (XGBoost):* XGBoost was adopted to provide a high-performance standard of predictive accuracy. It employs a sequential boosting method in which each successive tree is trained to correct the errors of the previous tree. The learning rate and maximum depth were adjusted to 0.1 and 6, respectively, to achieve the best performance on the dataset.

Following comparative analysis, the Random Forest Classifier was chosen as the best model in the proposed framework. It provided a better trade-off between high accuracy and human interpretability. To ensure the model reflects the educational priority of safeguarding vulnerable students, the class_weight parameter was set to 'balanced'. This technical change raises the penalty for misclassifying at-risk students and effectively reduces false negatives by ensuring that the system does not ignore struggling learners.

Evaluation of Models

A detailed evaluation was established to determine the predictive model's effectiveness. This assessment was conducted to ensure that the selected model is accurate, fair, and pedagogically effective. To measure the classification performance of the models on the test set, the following measures were used:

- *Accuracy:* It is the proportion of all classifications that were correct, whether positive or negative. It is defined as:

$$\text{Accuracy} = \frac{Number\ of\ correct\ classification\ instances}{Total\ number\ of\ instances}$$

- *Precision:* It is a measure of the model's reliability and is calculated as the ratio of true positives to all positive predictions.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

- *Recall (Sensitivity):* It measures how many of the actual positive cases were correctly identified by the model. It is important to note that missing a positive case (false negative) is more costly than false positives.

$$\text{Recall} = \frac{True\ Positive}{True\ positive + False\ Negative}$$

- *F1-Score:* It is the harmonic mean of precision and recall. It is useful when we need a balance between precision and recall, as it combines both into a single number. A high F1 score means the model performs well on both metrics. Its range is [0,1].

$$\text{F1} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

- To solve the black-box problem observed in the literature, the framework includes Feature Importance Analysis as a diagnostic utility measure. This is one of the key elements of XAI. By identifying the factors involved in a risk prediction, such as high absenteeism or a decline in mid-term grades, the model provides the why behind an alarm. This diagnostic transparency provides the objective, evidence-based feedback needed to foster student self-reflection and metacognitive monitoring.

- To ensure the ethical nature of the predictive framework, we conducted a Fairness Audit by calculating the Demographic Parity Difference (DPD). Demographic parity dictates that the likelihood of a student being flagged as 'at-risk' should not depend on their protected characteristics (such as gender). This was done technically by measuring the absolute difference between the selection rates of male and female students:

$$DPD = \left| P\left(\hat{Y} = 1 \middle| Gender = Male\right) - P\left(\hat{Y} = 1 \middle| Gender = Female\right) \right|$$

If DPD is close to 0, the model provides strong support, meaning that a student is flagged based on their academic and behavioral factors (e.g., attendance, grades) rather than their demographic background. This analysis is important for the pedagogical integrity of the framework; it ensures the 'metacognitive prompts' (feedback to students) are fair, non-biased, and not over-targeting or under-targeting any group of students.

RESULTS

Empirical testing of the risk prediction was conducted on a test sample of 56 student records, representing 20% of the refined cohort (N=278). The study benchmarked Logistic Regression, XGBoost, and Random Forest to identify the best-performing classifier. The benchmarking results yielded accuracies of 91.1% for Logistic Regression, 87.5% for XGBoost, and 92.9% for Random Forest. As shown in Table 1, the Random Forest model achieved better overall performance, with an accuracy of 92.9% and an F1-Score of 0.895%. Although Logistic Regression demonstrated a marginally higher recall score, its lower precision indicates it is more likely to produce false alarms and is less viable in institutions with limited intervention resources. Random Forest was more robust to the dataset's characteristics and provided a balanced, reliable detection rate.

The high predictive fidelity is further justified by the Receiver Operating Characteristic (ROC) Curve (Figure 6). The three models had high discriminative power with an Area Under the Curve (AUC) of over 0.96. In particular, the Random Forest model demonstrated an AUC of 0.977, indicating that it can differentiate between at-risk and non-at-risk students with nearly perfect accuracy across different classification thresholds. The proximity of the curves indicates that all models are accurate; the choice of a Random Forest is supported by the best balance between high AUC and human-friendly interpretability.

Table 1. Comparative performance of optimized machine learning models on the test set (n=56)

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.911	0.818	0.947	0.878
Random Forest	0.929	0.895	0.895	0.895
XGBoost	0.875	0.800	0.842	0.821

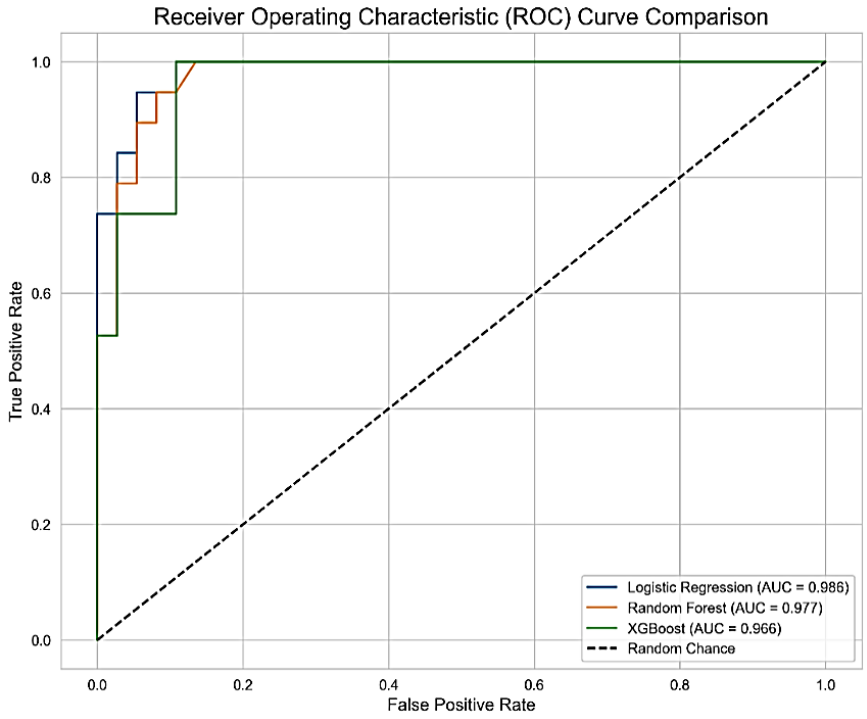


Figure 6. ROC curve for the Random Forest Model

The Confusion Matrix in Figure 7 provides a granular assessment of the reliability of the model. The system properly identified 17 of 19 at-risk students, with a miss rate (False Negatives) of only 10.5%. Pedagogically, the two missed cases are the only instances in which the system failed to invoke a required intervention. On the other hand, the model generated just two False Positives (5.4% error rate), which makes the institutional support targeted specifically to students with real needs. The framework is designed to reduce both missed opportunities and false alarms, providing a sound foundation for mentoring. Besides the model's predictive and discriminative performance, its ethical performance was evaluated using the predefined Fairness Audit.

The DPD was used to compare the rates of risk prediction for male and female students in the test set. The audit found that the DPD value of 0.03 is well below the acceptable threshold of algorithmic fairness (Figure 8). This suggests that the model predicts male and female students at similar rates, and demonstrates that the system is fair and offers equal support across genders. This is important from an educational perspective, as it means the "metacognitive prompts" are based solely on objective academic indicators rather than subjective ones (such as gender) and thus satisfy the fundamental ethical requirement of responsible AI in education.

After this fairness check, we analyzed the Feature Importance (Figure 9) to extract educational insights. This demonstrated that, while academic performance (Second_Period_Grade: 0.34, First_Period_Grade: 0.21) is the most important driver, behavioral features are crucial for early risk identification. The indicators School_Absences and Past_Failures are also seen to be important. The list provides a much-needed "Educational Roadmap" for educators; it suggests that academic performance issues could be a symptom of an engagement problem. For instance, excessive absenteeism can be seen as an early behavioral sign that guides practitioners before the student reaches a point of no return.

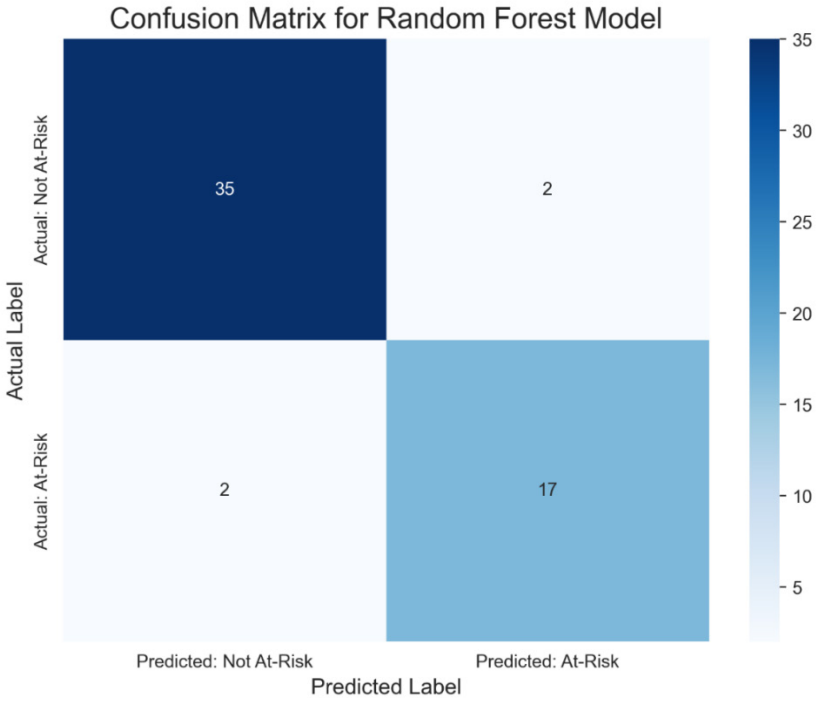


Figure 7. Confusion matrix for the selected random forest model on the test set

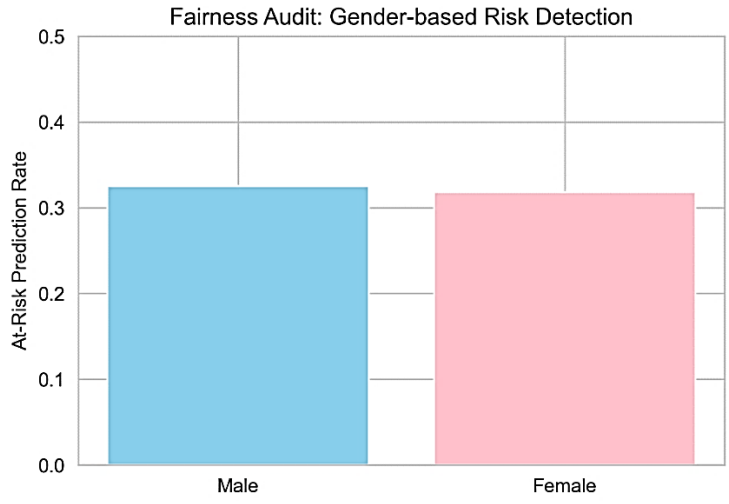


Figure 8. Fairness audit: gender based risk detection

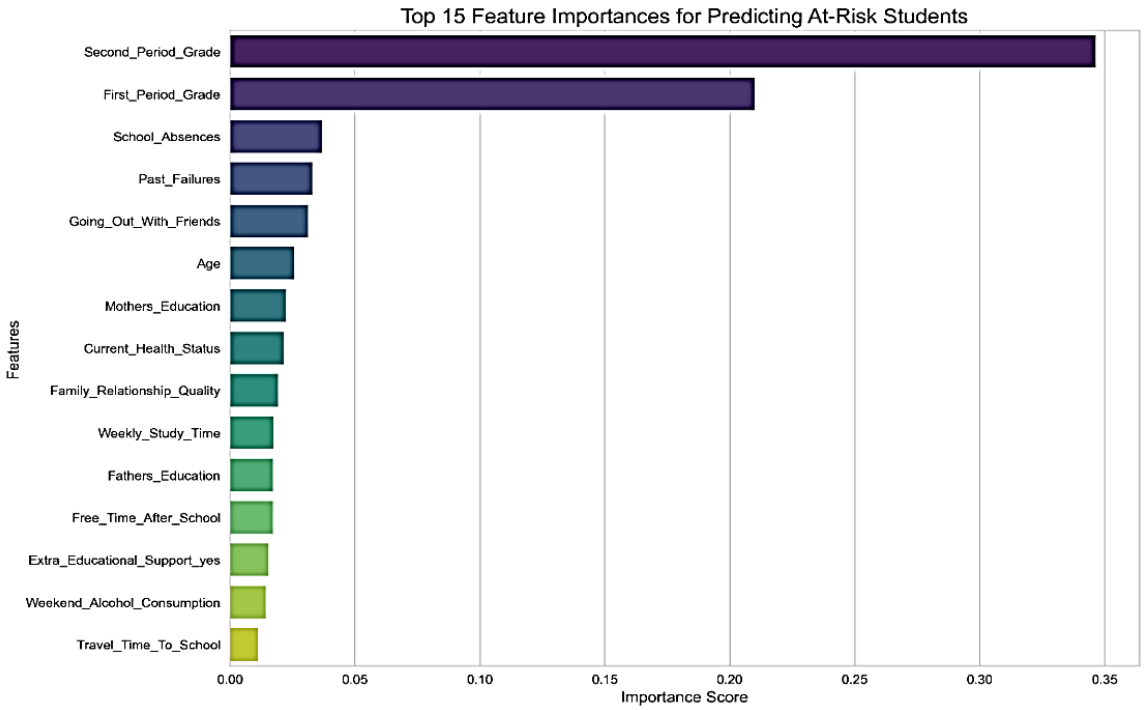


Figure 9. Top 15 feature importance for predicting at-risk status

XAI and Metacognitive Triggers

The SHAP summary plot in Figure 10 provides a global view of the impact of features on the risk prediction. The directionality of the impact is clear: low values (blue dots) of grades, and high values (pink dots) of absences and previous failures are pushing the model towards a prediction of at-risk. This global transparency ensures that the systems' reasoning is based on observable patterns in students' behavior to address the ethical issues posed by the black box of previous research (Pitts et al., 2025).

To facilitate local and individual intervention, the SHAP Waterfall Plot (Figure 11) provides a diagnostic report for the student. As illustrated in Student #5, this framework explains how particular academic skills (purple bars indicate the grades) were effectively leveraged to reduce the student's risk from a baseline of 0.509 to a final prediction of 0.02. The image is an effective Metacognitive Trigger. By sharing this evidence-based analysis with the student, the AI supports what Flavell (1979) described as cognitive monitoring. It facilitates the Self-Reflection phase of Zimmerman (2000) as the student sees how their particular behaviors affect their performance. This transparency turns a technical prediction into a pedagogical dialogue that enables students to make strategic decisions on their learning process.

Lastly, the XAI approach addresses the diagnostic gap in previous high-accuracy but black-box models. The approach supports the early detection of risks as actionable and fair by leveraging both high technical performance (92.9% accuracy and 0.977 AUC) and interpretability, along with a formal Fairness Audit. This human-centered approach aligns technical AI performance with the core values of modern pedagogy, fostering a supportive environment for student metacognitive growth.

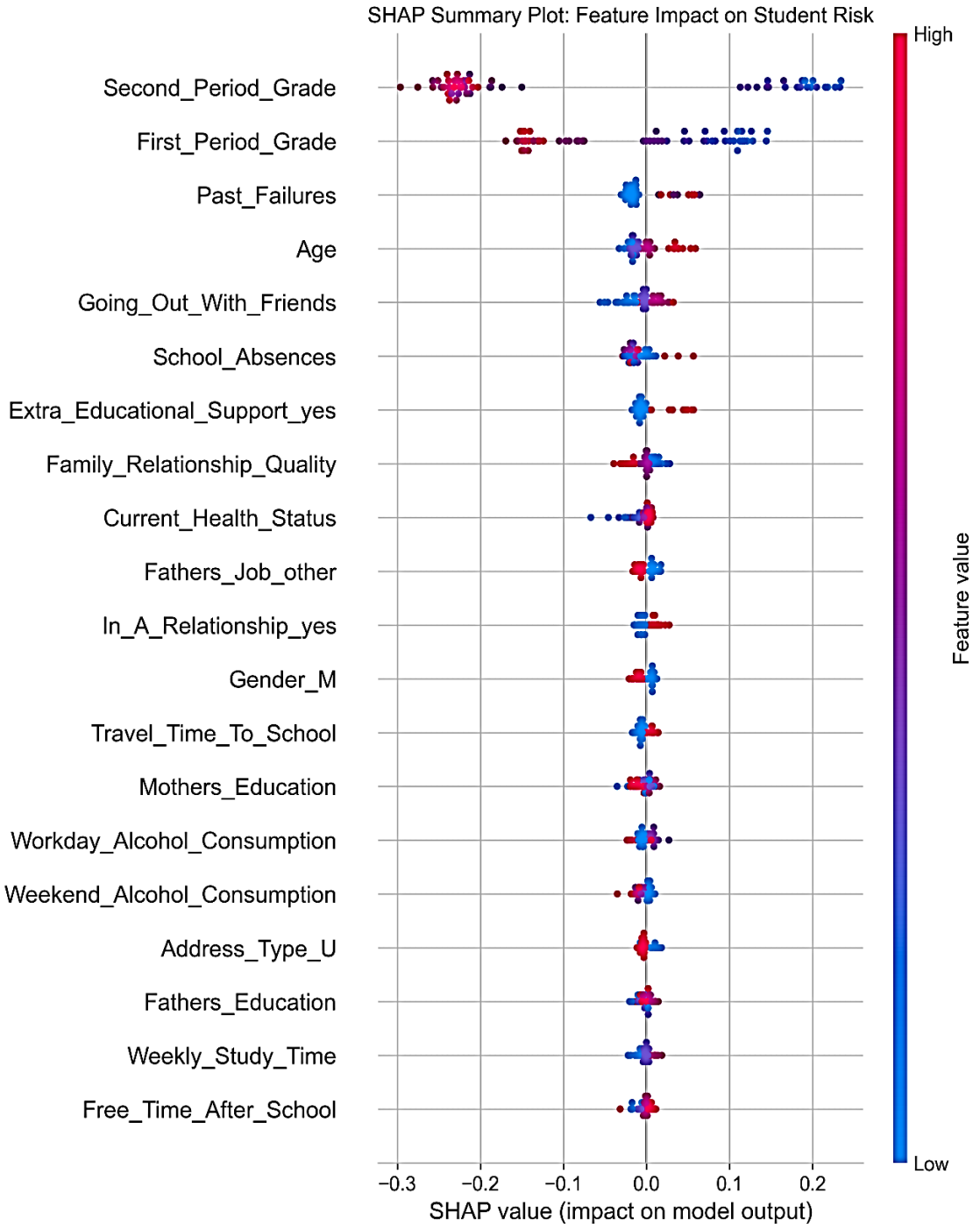


Figure 10. SHAP summary plot

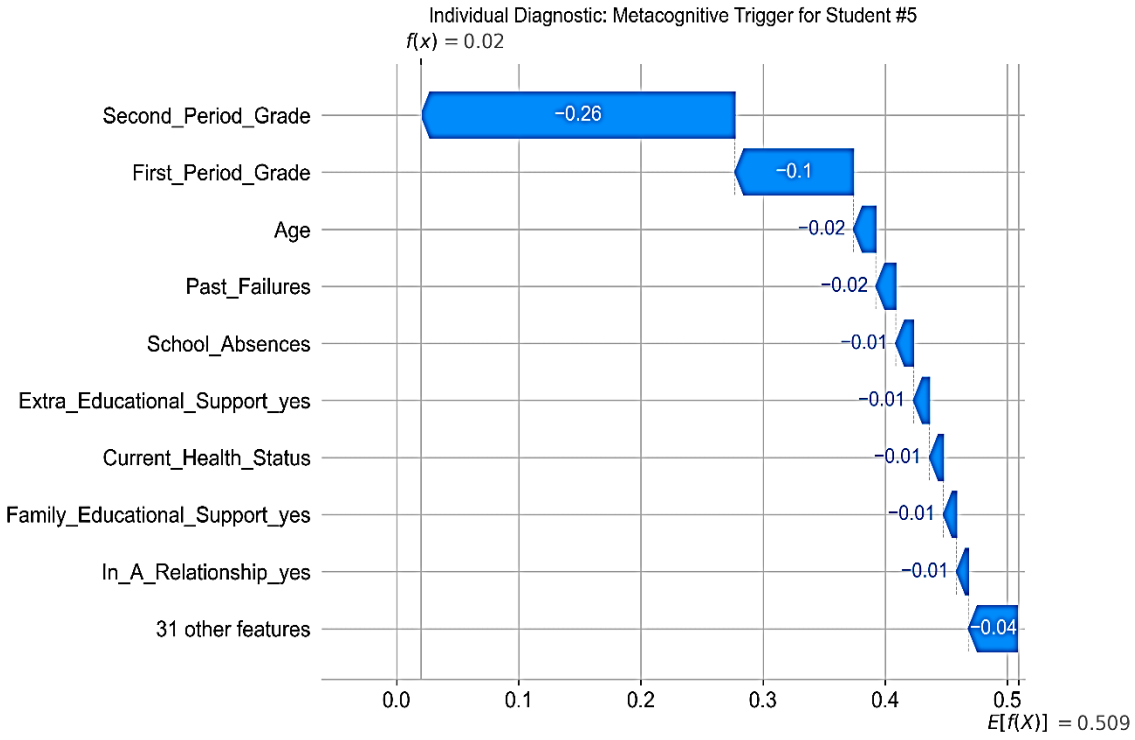


Figure 11. SHAP waterfall plot

Comparative Analysis with State-of-the-Art

To evaluate the significance of our findings, we conducted a systematic comparison of our proposed framework with existing studies in the field of EDM. While many researchers have achieved high predictive accuracy, our approach is unique in its integration of XAI and Fairness Audits, specifically designed to trigger metacognitive growth.

As shown in Table 3, our framework offers a better balance between pedagogical utility and technical performance. The study by Al-Shabandar et al. (2019) showed a slightly better accuracy with deep learning, but their model was a black box, making it impossible for a teacher to explain risk factors to a student. In contrast to the works of Algarni et al. (2023) and Latif et al. (2023), who focus on prediction, our framework implements a Fairness Audit to guarantee ethical implementation.

However, the most notable difference is the metacognitive link. Using SHAP waterfall plots, our framework converts technical information into what we consider a Metacognitive Trigger. This directly addresses the research gap on student agency; it enables students to shift from passive subjects of prediction to active participants in their own learning, meeting the criteria of the Self-Regulated Learning (SRL) cycle. On a deeper level, the 0.89 Recall score of our model has significant educational implications. It means that the framework is extremely sensitive to student struggle, so that 89 percent of the genuinely at-risk students are detected before it is too late. For an educator, this means the system acts as a reliable safety net. Moreover, the outcomes of the Feature Importance (e.g., the absenteeism as a leading predictor) change the educational dialogue from inherent ability to manageable behavior. Such interpretation is essential in the metacognition process. Once a student understands that their risk is caused by absences, not by a deficiency in intelligence, they will be in a psychological position to maintain their attendance and improve their results. This interpretive synthesis confirms that our framework is not just a technical success but a practical pedagogical tool.

Table 3. Comparison of the proposed framework with state-of-the-art studies

Study	Methodology	Accuracy	Interpretability (XAI)	Ethical fairness	Meta-cognitive link
(Algarni et al., 2023)	RF, SVM, LR	91.2%	Low (black box)	Not addressed	None
(Latif et al., 2023)	Multi-classifiers	88.5%	None	Not addressed	None
(Al-Shabandar et al., 2019)	Deep learning	93.5%	Very low	Not addressed	None
(Jiang et al., 2024)	Stacking	89.7%	Moderate	Not addressed	None
(Waheed et al., 2020)	DT, NN, RF	82.0%	None	Not addressed	None
Proposed framework	RF + SHAP (XAI)	92.9%	High (SHAP)	Verified (DPD)	Direct (SRL Trigger)

CONCLUSION AND FUTURE WORK

This study closes the diagnostic gap in educational analytics by developing and validating an XAI-based framework for early student risk detection. Conventional ML models in education have traditionally been black boxes that emphasize mathematical precision rather than pedagogical disclosure. With a predictive accuracy of 92.9% and a very high AUC-ROC of 0.977, when an optimized Random Forest model is used, the framework serves as a highly reliable early-warning system that reduces both false alarms and missed interventions. One of the main contributions of this work is the integration of XAI via SHAP analysis, which can convert a technical risk score into a personalized diagnostic roadmap by revealing the why behind a prediction, pointing to behavioral drivers such as absenteeism, past failures, and midterm grades. The system is not just a classification but an action. Importantly, a formal Fairness Audit is a crucial ethical element of the framework. Using the Demographic Parity Difference measure, the study found that the model is not biased and provides fair risk detection across gender groups. This guarantees that the academic assistance is grounded in objective behavioral patterns, with critical attention to algorithmic bias in automated educational tools. In the end, this research provides a viable, scalable solution for higher education institutions to transition from reactive to proactive student support. The suggested framework will ensure that AI is a collaborative pedagogical partner that empowers educators and learners with clear, evidence-based feedback. The system creates a setting in which students can self-reflect and act intelligently in their academic path by identifying specific behavioral obstacles. Future studies will focus on longitudinal classroom studies to empirically determine how these transparent diagnostic prompts can influence student retention and behavior change across a wide variety of institutional settings.

DATA AVAILABILITY

The dataset analyzed during the current study is publicly available in the UCI Machine Learning Repository: https://archive.ics.uci.edu/dataset/320/student+performance

REFERENCES

Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics, 37*, 13–49. https://doi.org/10.1016/j.tele.2019.01.007

Algarni, A., Abdullah, M., Allahiq, H., & Qahmash, A. (2023). Predicting at-risk students in higher education. *International Journal of Intelligent Systems and Applications in Engineering, 11*(3), 1229–1239. https://ijisae.org/index.php/IJISAE/article/view/3382

Al-Shabandar, R., Hussain, A. J., Liatsis, P., & Keight, R. (2019). Detecting at-risk students with early interventions using machine learning techniques. *IEEE Access, 7*, 149464–149478. https://doi.org/10.1109/ACCESS.2019.2943351

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI, 7*(1), 52–62. https://doi.org/10.61969/jai.1337500

Chong, K. E., Ibrahim, N., Huspi, S. H., Wan Kadir, W. M. N., & Isa, M. (2025). A systematic review of machine learning techniques for predicting student engagement in higher education online learning. *Journal of Information Technology Education: Research, 24*, 005. https://doi.org/10.28945/5456

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/10.1016/j.procs.2023.12.187

Feng, G., Fan, M., & Chen, Y. (2022). Analysis and prediction of students' academic performance based on educational data mining. *IEEE Access*, *10*, 19558-19571. 10.1109/ACCESS.2022.3151652

Fiok, K., Farahani, F. V., Karwowski, W., & Ahram, T. (2022). Explainable artificial intelligence for education and training. *The Journal of Defense Modeling and Simulation, 19*(2), 133–144. https://doi.org/10.1177/15485129211028651

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Gunasekara, S., & Saarela, M. (2025). Systematic literature review on explainable learning analytics and educational data mining. *IEEE Access, 13*, 214387–214408. 10.1109/ACCESS.2025.3643645

Hamid, M., Malik, S., & Saleem, M., Zahary, A. T., & Jaghdam, I. H. (2026). Enhancing educational assessment through automated question classification using a RoBERTa-based ensemble model. *Scientific Reports*, *16*, Article 13754. https://doi.org/10.1038/s41598-026-45486-1

Iranzad, R., & Liu, X. (2025). A review of random forest-based feature selection methods for data science education and applications. *International Journal of Data Science and Analytics, 20*(2), 197–211. https://doi.org/10.1007/s41060-024-00509-w

Jiang, X., Du, Y., & Zheng, Y. (2024). Evaluation of physical education teaching effect using Random Forest model under artificial intelligence. *Heliyon, 10*(1), e23576. https://doi.org/10.1016/j.heliyon.2023.e23576

Kara, A., Ergulec, F., & Eren, E. (2024). The mediating role of self-regulated online learning behaviors: Exploring the impact of personality traits on student engagement. *Education and Information Technologies, 29*(17), 23517–23546. https://doi.org/10.1007/s10639-024-12755-3

Khine, M. S. (2024). Educational data mining and learning analytics. *Artificial intelligence in education: A machine-generated literature overview* (pp. 1–159). Springer. https://doi.org/10.1007/978-981-97-9350-1_1

Kong, S. C., Ogata, H., & Shih, J. L. (2021, November). The role of artificial intelligence in STEM education. In M. M. T. Rodrigo, S. Iyer, & A. Mitrovic (Eds.), *International Conference on Computers in Education*. Asia-Pacific Society for Computers in Education. https://library.apsce.net/index.php/ICCE/article/view/4336

Latif, G., Abdelhamid, S. E., Fawagreh, K. S., Brahim, G. B., & Alghazo, R. (2023). Machine learning in higher education: students' performance assessment considering online activity logs. *IEEE Access, 11*, 69586–69600. https://doi.org/10.1109/ACCESS.2023.3287972

Lee, H., Stinar, F., Zong, R., Valdiviejas, H., Wang, D., & Bosch, N. (2025, April). Learning behaviors mediate the effect of AI-powered support for metacognitive calibration on learning outcomes. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1–18). Association for Computing Machinery. https://doi.org/10.1145/3706598.3713960

Lukyanenko, R., Castellanos, A., Samuel, B., Tremblay, M., & Maass, W. (2021). Research agenda for basic explainable AI. *AMCIS Proceedings*. https://aisel.aisnet.org/amcis2021/art_intel_sem_tech_intelligent_systems/art_intel_sem_tech_intelligent_systems/7

Moreno-Marcos, P. M., Pong, T.-C., Muñoz-Merino, P. J., & Delgado Kloos, C. (2020). Analysis of the factors influencing learners' performance prediction with learning analytics. *IEEE Access, 8*, 5264–5282. https://doi.org/10.1109/ACCESS.2019.2963503

Pitts, G., Marcus, V., & Motamedi, S. (2025). *Student perspectives on the benefits and risks of AI in education.* PsyArXiv. https://doi.org/10.48550/arXiv.2505.02198

Rienties, B., Divjak, B., Eichhorn, M., Iniesto, F., Saunders-Smits, G., Svetec, B., Tillmann, A., & Zizak, M. (2023). Online professional development across institutions and borders. *International Journal of Educational Technology in Higher Education, 20*, Article 30. https://doi.org/10.1186/s41239-023-00399-1

Vieriu, A. M., & Petrea, G. (2025). The impact of artificial intelligence (AI) on students' academic development. *Education Sciences, 15*(3), 343. https://doi.org/10.3390/educsci15030343

Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior, 104*, Article 106189. https://doi.org/10.1016/j.chb.2019.106189

Williamson, B., & Eynon, R. (2020). Historical threads, missing links, and future directions in AI in education. *Learning, Media and Technology, 45*(4), 321–333. https://doi.org/10.1080/17439884.2020.1798995

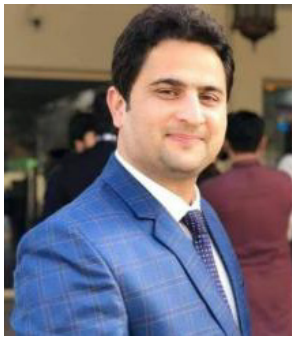
Wu, Y. (2023). Integrating generative AI in education: how ChatGPT brings challenges for future learning and teaching. *Journal of Advanced Research in Education, 2*(4), 6–10. https://doi.org/10.56397/JARE.2023.07.02

Yang, Y., & Xia, N. (2023). Enhancing students' metacognition via AI-driven educational support systems. *International Journal of Emerging Technologies in Learning, 18*(24), 133–148. https://doi.org/10.3991/ijet.v18i24.45647

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–39). Academic Press. https://doi.org/10.1016/B978-012109890-2/50031-7

AUTHORS

Saadia Malik holds an MSc degree in Computer Science and a PhD in Information Systems. Her PhD dissertation focuses on structured information retrieval. One aspect is using artificial intelligence to analyze user interaction with a structured information retrieval-based system. Dr. Malik has approximately 13 years of experience in administration, research, and teaching at the University of Duisburg-Essen, Germany, and King Abdulaziz University, Kingdom of Saudi Arabia. Her research interests include artificial intelligence, data mining, information retrieval, and software engineering.



Muhammad Hamid holds an M.Sc in Information Technology, an MPhil, and a PhD in Computer Science. His doctoral research focused on enhancing Pakistan's software export potential by addressing recurring industry challenges through the application of Artificial Intelligence (AI). During his PhD studies, he conducted collaborative research with the Department of Computer Science and Operations Research at the University of Montreal, Canada. He brings over 14 years of combined experience in teaching, research, and academic administration. He is currently serving as an Assistant Professor at Government College Women University, Sialkot, Pakistan. His research interests include artificial intelligence, software engineering, and the integration of AI in educational systems. He actively contributes to the academic community as a reviewer, technical committee member, and editorial board member for several reputable national and international conferences and journals.

Muhammad Saleem holds a Master's degree in Computer Science and Communications Engineering from the University of Duisburg-Essen, Germany, and a PhD in Engineering from the University of Federal Armed Forces, Munich, Germany. Dr. Saleem has more than 15 years of teaching, research, and administrative experience in the Department of Industrial Engineering at the University of Duisburg-Essen, Germany, and King Abdulaziz University, Kingdom of Saudi Arabia. Dr. Saleem also worked as a project manager in Siemens Power Generation's energy sector in Munich, Germany, for 3 years. His work is focused mainly on Industrial Quality Control, Artificial Intelligence, and Renewable Energy. He is actively involved in curriculum development and accreditation processes of engineering programs.