I.J. Modern Education and Computer Science, **2026, 5, 149-167**
Published Online on October 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2026.05.09

A Hybrid Explainable AI Model for Accurate and Transparent Prediction of Student Academic Performance

Mohammad Nasar*
Computing and Informatics Department, Mazoon College, Muscat, Oman
Email: nasar31786@gmail.com
ORCID iD: https://orcid.org/0000-0001-8196-380X

Mohammad Abu Kausar
Department of Information Systems, University of Nizwa, Oman
Email: kausar4u@gmail.com
ORCID iD: https://orcid.org/0000-0002-7997-8292
*Corresponding author

Received: August 22, 2025; Revised: October 25, 2025; Accepted: February 19, 2026; Published: October 8, 2026

Abstract: Predicting student academic performance accurately is essential and allows for timely interventions and data-supported educational decision-making. However, the black-box nature of state-of-the-art machine learning models hinders their use in policy-sensitive educational settings, where interpretability and accountability are paramount. We propose a structured hybrid explainable artificial intelligence (XAI) framework that combines three tree-based ensemble models–namely, random forest, XGBoost, and CatBoost–using a prediction-level soft-voting strategy and couples them with complementary post-hoc interpretation methods at the explanation level (SHAP and LIME). Instead of proposing a new base algorithm, this study systematically unifies heterogeneous boosting and bagging approaches with strict validation as well as dual-layer explanation consistency evaluation. Experiments on the UCI Student Performance dataset show that the proposed framework can achieve a competitive predictive performance (91.8% accuracy, ROC–AUC = 0.953) together with transparent and actionable interpretability. The robustness of the interpretability layer is further supported by a quantitative assessment of explanation stability, fidelity, and agreement across various methods. The presented framework harmonizes the trade-off between accuracy and interpretability to provide a deployable, policy-aware decision-support solution that is aligned with responsible AI principles for adoption within educational contexts.

Index Terms: Explainable AI, Hybrid Machine Learning, SHAP, LIME, Student Performance Prediction, Educational Data Mining, Ensemble Learning.

1. Introduction

In the last few years, artificial intelligence (AI) has changed the face of many sectors; however, the education domain is among the most affected, as AI allows the elimination of subjective evaluation methods and the provision of continued learning support to students. Educational data mining (EDM) and learning analytics (LA) refer to the automated, systematic extraction of patterns from large datasets. They can assist institutions in predicting success, flagging potential students likely to fail, and providing the opportunity to intervene with support mechanisms before it is too late. Another important aspect of higher education is retention prediction, where success is determined by retention rates, graduation rates, and academic performance [1, 2].

In contrast, the increasing embedding of AI in educational decision-making has a significant drawback, the “black-box” problem [3, 4]. Moreover, state-of-the-art machine learning techniques (i.e., gradient boosting machines [5], random forests [6], and XGBoost [7]) could yield non-interpretable black-box output that is not very useful for stakeholders to understand. In educational settings, however, the interpretability frontier imposes limits on the adoption of such models. Consequently, decision-makers, such as teachers, principals, or policymakers, need not only accurate predictions but also explanations to make informed decisions, trust the model[s], and meet the requirements of the General Data Protection Regulation (GDPR) [8].

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

Explainable artificial intelligence (XAI) is a potential solution that creates an understanding of how a model makes decisions [9-11]. Methods such as SHapley Additive exPlanations (SHAP) [10, 11] and Local Interpretable Model-Agnostic Explanations (LIME) [9] provide local (instance-specific) and global (i.e., dataset-wide) interpretations of model behavior. This helps identify factors, such as academic history, family background, behavioral markers, or socioeconomic variables, from which educators can largely predict student learning outcomes. For instance, Kesgin et al. [1] demonstrated that SMOTE [12], a tree-based algorithm combined with the concept of using fairness-aware models, could ensure gender and ethnic fairness via interpretable predictions over different subgroups for student prediction.

The UCI Student Performance Dataset is a representative dataset for other EDM studies [13, 14]. The dataset contains attributes such as demographic variables, academic records, and lifestyle features of students in Portuguese secondary education. The dataset [13] has previously been studied with several predictive methods, such as support vector machines, logistic regression, and deep learning architectures [15, 2, 16, 17, 18]. Some existing studies have initiated solutions to the interpretability issues discussed above, whereas others have focused on achieving higher precision at the cost of interpretability [23]. Others have presented promising interpretation methods along with complexity in producing better predictions through data-analytic model ensembles [17, 33, 19].

To address this gap, we present a hybrid explainable AI framework by leveraging a combination of different strengths of disparate error rate-based machine learning algorithms and strong post-hoc interpretation methods. The complementary ability of random forest [6], XGBoost [7], and CatBoost [20] in handling heterogeneous educational data makes an ensemble of these three models a strong candidate for the proposed model. Within the ensemble framework, model predictions were further analyzed using SHAP and LIME to obtain both instance-level explanations and statistically robust global feature importance measures, thereby preserving interpretability alongside predictive performance. In doing so, predictions are multi-layered and present a high degree of predictive accuracy along with logical reasoning behind them.

The main contributions of this study are as follows.

1. **Structured Hybrid Integration Framework —** A structured hybrid integration framework is a systematic soft-voting ensemble that joins three weak learners, namely random forest, XGboost, and CatBoost, with extensive ablation and synergetic verifications to quantify the individual contribution of each model.
2. **Dual-Layer Explainability Assessment —** Merging both SHAP (global and local consistency) and LIME (instance-level interpretability) with a static evaluation of explanation fidelity, stability, and cross-method consistency.
3. **Rigorous Validation and Statistical Verification —** Controlled processes, including cross-validation, holdout testing, Bayesian hyperparameter tuning, and Wilcoxon signed-rank statistical tests are used to verify performance improvements.
4. **Educational Policy-Oriented Design Perspective —** Framing predictive modelling within a responsible AI and policy-compliant decision-support context suitable for early-warning academic systems.

Unlike previous works, where related studies aimed to either develop improved predictive accuracy without providing interpretability or rely on different interpretability algorithms, this work does not aim to invent a new base algorithm. Instead, it defines a systematic structure for hybridization that aligns mutually advantageous boosting and bagging models with complementary explanatory structures that ground policy-compliant educational decision support. The contributions of this study are a systematic integration followed by a stringent validation strategy, verification of statistical significance, and synergy through an ablation study, along with the quantification of the fidelity of explanation over benchmarked education datasets. This study is therefore positioned as a methodological consolidation and validation-oriented contribution rather than the introduction of a new algorithmic paradigm.

The rest of this paper is structured as follows: literature review of existing work on student performance prediction and XAI approaches (Section 2), followed by the proposed hybrid XAI framework (Section 3) and experimental setup (Section 4), then results and discussion (Section 5), before concluding with research findings, implications and challenges in future research directions (Section 6).

2. Literature Review

The academic performance prediction of students has been one of the main research directions within the area of educational data mining (EDM) and learning analytics (LA) over the last 20 years [21-23]. This includes specialized statistical methods, as well as data mining and EDM techniques for analysing sets of education-related data. Three research lines are related to this work: (i) predictive modelling of academic performance, (ii) XAI frameworks in education, and (iii) hybrid ensembles incorporating interpretability.

2.1. Predictive Modelling in Educational Data Mining

Preliminary research on traditional types of machine learning algorithms for EDM. These include decision trees, K-nearest neighbors (KNN), and logistic regression, which seek to determine the relationships between students' work [13, 18, 21, 22]. Although early models, such as decision trees and logistic regression, were interpretable in terms of decision rules, their predictive performance was limited when representing nonlinear and heterogeneous data [5]. In recent years, more sophisticated algorithms, such as random forest (RF) [6], gradient boosting machine (GBM) [5], XGBoost [7], and CatBoost [20], have emerged that exploit higher-order interactions between features, which typically increases predictive accuracy.

The UCI student performance dataset [13, 14] is a widely used benchmark dataset for developing prediction models. In fact, the dataset was used to set a benchmark performance based on decision trees, random forests, and neural networks using secondary school data [13]. However, for the accuracy test, CatBoost + Double PSO was the winning method [15]. The generalization ability of the models on our dataset was also confirmed, as [16] demonstrated that tree-based models dramatically outperform linear models on this task.

Several EDM studies emphasize enhancing the different performance metrics of the model (accuracy, F1-score); however, they do not consider the interpretability of the model [15, 2, 16]. This emphasis on performance indicators constrains educational administrators who require decision-making information [3, 4].

2.2. Explainable Artificial Intelligence in Education

To help fill this gap in interpretability, explainable AI (XAI) [9, 10, 24] aims to explain how a model arrives at a certain prediction and/or why a certain prediction is made with a given model. The need for transparency-aware system design is even more salient, following recent discussions on embedding XAI in a general educational data mining ecosystem [25]. Interpretability is crucial for trustworthiness, accountability, and compliance with data protection legislation, such as the General Data Protection Regulation (GDPR) [8] in education. Post-hoc explanation methods, such as LIME [9] and SHAP [10, 11], are widely used and allow users to understand the contribution of individual features to a particular prediction.

More recently, we have argued that explainability cannot rely solely on vision but calls for systematic benchmarking and quantitative validation. Mohamed et al. [26] provide a systematic overview of recent developments and emphasize the importance of explanation stability and robustness across domains. Salih et al. [27] fully discussed both SHAP and LIME, theoretically and empirically, revealing the complementary merits of the two methods and which aspects of model reasoning each method captures. We have previously stated that global and local explanation techniques should be evaluated for consistency as well as fidelity when performed by benchmarking initiatives such as BenchXAI [28] or large-scale tabular data evaluations [29]. These recent works suggest the potential for moving toward structured and empirically based interpretability, particularly in high-stakes domains such as education.

Kesgin et al. [1] used SHAP in conjunction with a fairness-aware sampling method (SMOTE [12]) to increase fairness in predicting student performance and indicated that interpretability and fairness can be achieved together. Johora [17] used SHAP and LIME to predict undergraduate performance in Bangladesh and explored how one such explanation can guide curriculum redesign. Similarly, Mastour [30] adapted these concepts to medical education and used SHAP and LIME with a stacking ensemble (RF, AdaBoost, XGBoost) to gain insight into whether interpretations of high-stakes exams are transparent. Akter et al. [19] provided socio-academic and economic drivers of exam performance by merging causal inference with SHAP-based explanations.

Although these works collectively underscore the promise of XAI, most have been demonstrated on individual models or context-specific datasets, thus restricting their generalizability and potential performance.

2.3. Hybrid and Ensemble Approaches with Interpretability

The essence of ensemble learning is to combine multiple models in a way that improves the robustness and accuracy of predictions. Popular ensemble techniques, such as bagging (e.g., Random Forest [6]) and boosting methods (e.g., gradient boosting), frequently yield superior results compared to individual models; however, they introduce additional complexity and reduced transparency.

Ahmed [31] used heterogeneous ensemble regressors to predict student performance and obtained better accuracy but no model explanation. Adeyemi and AlOtaibi [32] incorporated SHAP into a decision support system based on LightGBM for early in-class student interventions but did not apply the method to benchmark datasets such as UCI. George et al. [33] used several interpretability techniques for the phonemes' grade prediction, but halted hybrid model integration by providing a comparative view of the available explanations. Zhang et al. [34] highlighted human-centered design for early prediction models and outlined the need to align interpretability with educator needs.

The above findings point to a gap in the literature, in which high-performing hybrid models exist and XAI techniques are broadly tested; however, very few studies combine ensemble learning and dual-layer explainability (SHAP + LIME) on benchmark datasets to produce optimal accuracy with actionable insights.

Additionally, recent comparative analyses caution against assuming that the aggregation of explanation techniques necessarily improves interpretability. Givisis et al. [35] show that both SHAP and LIME depend on model structure and dataset properties when deciding feature attributions. These principles are already being applied in practice; for example,

Lünich and Keller [36] demonstrated that the simplicity of explanations and perceived fairness play a significant role in stakeholders' trust in predictive systems. In other words, empirical warrants should be made for the integration of the two methods, in addition to their conceptual justification; therefore, a systematic validation method was undertaken in this study.

2.4. Identified Research Gap

Despite the strong advancements in educational data mining for both predictive modelling and explainable artificial intelligence (XAI) [25], important missing pieces of knowledge remain.

1. Performance–Interpretability Trade-off

Several high-performing algorithms, such as gradient boosting and random forest, can produce higher predictive accuracy; however, they function as "black boxes" [3, 4]. In contrast, interpretable models (e.g., decision trees and logistic regression) prioritize transparency over accuracy. Only a handful of studies have attempted to combine these goals in a coherent framework [17, 33, 30].

2. Limited Use of Complementary XAI Techniques

SHAP [10, 11] and LIME [9] are two common post-hoc interpretability methods that are often used in isolation. SHAP provides consistent global (importance) and strong local explanations, whereas LIME is a tool for human readability and instance-specific interpretability. Combining them can produce deeper and more actionable insights; however, dual integration has been little used in educational research [1, 19, 32].

3. Benchmarking Deficiency on Widely Recognized Datasets

Most studies validate their methods using in-house or context-specific datasets, making reproduction and inter-study comparisons difficult [17, 30, 32]. The UCI Student Performance Dataset [13, 14] is one of the few public benchmarks featuring multidimensional student data; however, none of the ensemble–XAI integrations have been rigorously evaluated therein. This hampers the assessment of generalizability and replication of findings in diverse educational settings.

4. Lack of Policy and Compliance Considerations

For high-stakes educational decisions, such as determining whether to retain a student, award a scholarship, or encourage targeted academic interventions, interpretability is not only beneficial but is frequently mandated by policy frameworks and regulations (e.g., the EU's GDPR enshrined a "right to explanation" for automated decision-making [8]). However, most previous studies do not explicitly mention compliance with such policies, which could limit the uptake in institutional environments.

5. Practical Adoption Barriers

Even for interpretable models, it may be difficult to translate the explanations to non-specialist stakeholders (such as academic advisors, school administrators, or policymakers). This establishes a gap between what the model is able to produce and what end-users can act on [33, 34]. Few studies convert interpretability outputs into educator-friendly insights, which would provide direct guidance for actionable interventions.

Table 1. Summary of Key Related Works.

Ref.	Author(s), Year	Dataset	Algorithm(s)	XAI Method(s)	Performance Metrics	Key Findings	Limitation
[13]	Cortez & Silva (2008)	UCI Student Performance	Decision Tree, Neural Net, RF	None	Accuracy, RMSE	Baseline models for dataset	No interpretability
[1]	Kesgin et al. (2025)	UCI	RF, SVM + SMOTE	SHAP	Accuracy, Fairness metrics	Fairness-aware model with interpretable outputs	No ensemble optimization
[15]	Fan et al. (2023)	UCI	CatBoost + PSO	None	Accuracy, F1-score	Improved accuracy via PSO	No interpretability
[16]	Mohammad et al. (2023)	UCI	Multiple ML models	None	Accuracy, Recall	Tree-based models outperform linear	No explanation layer
[17]	Johora (2025)	Bangladesh dataset	Ensemble	SHAP, LIME	Accuracy, Precision	Interpretable predictions for undergraduate	Not benchmarked on UCI

						s	
[33]	George et al. (2025)	Institutiona l	Multiple ML	SHAP, LIME, PDP	Accuracy, Interpretability	Compared explanation methods for grades	No hybrid integration
[30]	Mastour (2025)	Medical education	Stacking (RF, AdaBoost, XGBoost)	SHAP, LIME	Accuracy, F1-score	Transparent medical exam prediction	Not applied to UCI dataset
[31]	Ahmed (2025)	Multi-institutional	Heterogeneous ensemble regressors	None	RMSE, MAE	High accuracy across datasets	No XAI integration
[19]	Akter et al. (2025)	Exam-based	Ensemble + causal	SHAP	Accuracy, Causal impact scores	Socio-academic factors identified	No boosting integration
[32]	Adeyemi & AlOtaibi (2025)	Real-time	LightGBM	SHAP	Accuracy, AUC	Feedback-driven intervention system	No standardized dataset
[34]	Zhang et al. (2025)	Real-world	LR, 1D-CNN, Multi-task models	None	Accuracy, Early prediction rate	Human-centered prediction design	No XAI a

3. Methodology

In this study, we describe the architecture and processes involved in the proposed hybrid explainable AI (XAI) framework using the UCI Student Performance Dataset for prediction and explainability tasks related to student academic performance. The proposed framework comprises five primary processes: dataset selection and preprocessing, model design, hybrid ensemble integration, explanation layer, and evaluation process.

3.1. Overview of the Hybrid XAI Framework

The proposed framework (Figure 1) comprises two primary layers:

1. **Prediction Layer** — Three high-quality tree-based models (random forest, extreme gradient boosting, and CatBoost) are aggregated in the prediction layer. These approaches learn models from preprocessed data that are optimized for predictive performance while accommodating data-type limitations.
2. **Explanation Layer** — The outputs of the ensemble are explained using the post-hoc interpretability tools SHAP and LIME, which provide both global feature importance (SHAP) and local, instance-specific explanations (LIME).

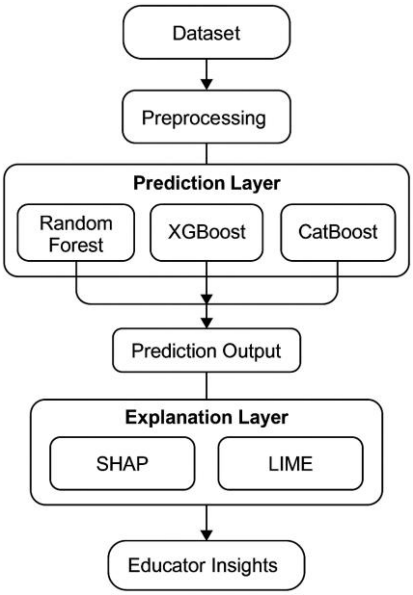


Fig. 1. Conceptual Diagram of the Proposed Hybrid XAI Framework.

3.2. Dataset Description

We conducted the experiments using the UCI Student Performance Dataset [13, 14], which includes demographic, social, and academic factors of Portuguese secondary school students. The dataset consists of two parts: courses in mathematics (mat) and Portuguese (por). The features are numerical and categorical variables, such as study time, past failures, parental education, travel time, and grades (G1, G2, G3).

Table 2. Summary of Dataset Features.

Feature Type	Examples	Count	Description
Demographic	sex, age, address, famsize	5	Gender, age, family size, residential type
Parental & Social	Medu, Fedu, Mjob, Fjob, guardian	5	Parental education and occupation
Behavioral	traveltime, studytime, failures, absences	4	Student habits and performance history
Social & Support	schoolsup, famsup, paid, activities	4	Support programs and activities
Academic	G1, G2, G3	3	First, second, and final grades
Total	—	21	Combined features from both subsets

3.3. Data Preprocessing

The preprocessing stage includes the following steps:

- **Data cleaning** — Removing missing or inconsistent entries.
- **Encoding categorical features** — Applying one-hot encoding for non-ordinal categorical variables and ordinal encoding where applicable.
- **Feature scaling** — Standardizing numerical features for XGBoost and CatBoost compatibility.
- **Train–test split** — Splitting data into training (70%) and testing (30%) sets, with stratification based on the target variable.
- **Balancing classes** — Applying the synthetic minority over-sampling technique (SMOTE) [12] to address class imbalance, as recommended in prior studies [1].

3.4. Prediction Layer: Ensemble Model Design

The following three algorithms form the hybrid ensemble.

3.4.1. Random Forest (RF)

A collection of decision trees trained using bagging to decrease variance [37]. A bootstrapped sample is used to train each tree, and for all trees, the most votes are taken (for classification) or the mean (regression). The output of RF is given as follows:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \tag{1}$$

where h_t is the t-th decision tree, T is the total number of trees, and x is the input feature vector.

3.4.2. XGBoost

A gradient-boosted trees algorithm that creates trees one at a time, wherein each new tree helps correct errors made by previously trained trees. The loss function is given by, which combines a loss term and a regularization term.

$$\text{Obj} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{t} \Omega\left(f_k\right) \tag{2}$$

where l is the loss function, Ω is the regularization term, and f_k represents the k-th tree.

3.4.3. CatBoost

CatBoost can be optimized using ordered boosting and target encoding to prevent overfitting, and it works well with categorical features. CatBoost supports categorical variables by default, simplifying preprocessing.

3.5. Ensemble Integration Strategy

The three constituent models are combined through a soft-voting ensemble mechanism where the predicted probabilities of each class for all constituent models are averaged, and then the class corresponding to the maximum average probability is output as the final prediction.

$$\hat{y} = arg\max_C \left(\frac{1}{M} \sum_{m=1}^{M} P_m(y = c \mid x) \right) \tag{3}$$

where M is the number of models, c is a class label, and P_m is the predicted probability from model m.

3.6. Explanation Layer: SHAP and LIME

- **SHAP** — Based on Shapley values from cooperative game theory, SHAP assigns an importance value to each feature by averaging the marginal contributions across all feature subsets:

$$\phi_j = \sum_{S \subseteq N\{j\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[f_{S\cup\{j\}}\left(x_{S\cup\{j\}}\right) - f_S\left(x_S\right) \right] \tag{4}$$

  where ϕ_j is the SHAP value for feature j, N is the set of all features, and S is a subset of N without j.
- **LIME** — This approach locally perturbs the input data, passing it into a complex model whose predictions are to be explained, and fits an interpretable (lower-bias) model (e.g., linear regression) around the prediction.

$$\hat{g} = \arg\min_{g \in e} L(f, g, \pi_x) + \Omega(g) \tag{5}$$

where:

g is the interpretable model selected from the class G
$L(f, g, \pi_x)$is a locality-weighted loss function
$\pi_x$is the proximity measure defining the locality around instance x
$\Omega(g)$is a complexity penalty term to ensure interpretability

3.7. Evaluation Metrics

The framework was evaluated by using multiple metrics to ensure a comprehensive assessment.
Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Precision:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

Recall:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

F1-Score:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

ROC-AUC: Area under the Receiver Operating Characteristic curve, assessing model discrimination capability.

4. Experimental Setup

In this section, we describe the experimental setup for assessing the proposed hybrid explainable AI (XAI) framework. Specifically, we describe the computational environment, software tools, dataset preparation, and hyperparameter tuning with a specific validation strategy, and finally the baseline models for benchmarking, to establish the experimental workflow. This is important for reproducibility and methodological transparency, which are considered best practices in educational data mining [21, 23] and explainable AI research [9-11].

4.1. Computational Environment

The experiments were conducted in a controlled local computing environment to ensure consistency across different phases of the model training, tuning, and evaluation. The specifications are as follows:

- **Processor**: Intel® Core™ i9-12900K CPU, 16 cores / 24 threads, 3.20 GHz base frequency
- **Graphics Processing Unit (GPU)**: NVIDIA GeForce RTX 3080 Ti, 12 GB GDDR6X VRAM
- **System Memory**: 64 GB DDR5 RAM
- **Storage**: 2 TB NVMe SSD (PCIe 4.0) for rapid data access and intermediate storage
- **Operating System**: Ubuntu 22.04 LTS, 64-bit

This configuration ensured sufficient computational resources for ensemble model training and for SHAP computations, which are known to be computationally intensive for large models.

4.2. Software Tools and Libraries

The framework was implemented in **Python 3.11**, utilizing well-established machine learning and visualization libraries.

- **Scikit-learn 1.5.0** — Preprocessing, evaluation metrics, and Random Forest implementation [38]
- **XGBoost 2.0.3** — Gradient boosting algorithm with GPU acceleration [7]
- **CatBoost 1.2.3** — Efficient categorical feature handling without extensive preprocessing [20]
- **Imbalanced-learn 0.12.0** — Synthetic Minority Oversampling Technique (SMOTE) for class balancing [12]
- **SHAP 0.44.0** — Shapley value computation for global and local feature importance [10, 11]
- **LIME 0.2.0.1** — Local interpretable model-agnostic explanations for instance-level interpretability [9]
- **Matplotlib 3.8.3** and **Seaborn 0.13.2** — Data visualization and interpretability plotting

Thus, we selected LIME to ensure dimensional consistency with regard to case-level attributions and the SHAP method, which would provide global metric-level consistency across the model.

4.3. Dataset Preparation

For our industry benchmark, we chose the UCI Student Performance Dataset [13, 14] because it contains rich demographic, academic, and socioeconomic features and is also popular for educational prediction studies [1,15,16].

Key preprocessing steps included:

1. **Label Definition** — the final grade (G3) target variable, encoded into binary classification
   - Pass: $G3 \geq 10$
   - Fail: $G3 < 10$

We made this binary transformation because it more closely matched early-warning system objectives in practice, where predicting the exact grade is less operationally relevant than identifying students at risk of achieving failing grades. Educational interventions tend to work by flagging children for potential failure rather than generating ongoing estimates from grade distributions. However, binary classification is a simplification of the prediction landscape that can create an overestimation in classification metrics when performing regression-based modelling. In future work, regression, ordinal, and multiclass stratified prediction strategies will be compared with respect to observed differences in performance and robustness across alternative modelling paradigms.

2. **Handling Categorical Data** —
   - One-hot encoding of nominal variables (e.g., sex, address, school).
   - Ordinal encoding of ordered variables (e.g., parental education levels).
3. **Feature Scaling** — Continuous variables were standardized to improve the model stability for XGBoost and CatBoost.

Table 3. Class distribution before and after SMOTE application.

Class	Before SMOTE	Percentage	After SMOTE	Percentage
Pass	395	68%	395	50%
Fail	185	32%	395	50%

Table 3 shows the number of instances per class before and after applying the synthetic minority over-sampling technique (SMOTE). The dataset is slightly biased towards the "Pass" label. To overcome this imbalance and simulate a more realistic environment for the machine learning model to learn, we applied SMOTE [16] only to the training set to obtain an equal distribution of 50-50 class balance. Additionally, note that no synthetic samples were added to the test set to prevent the introduction of optimistic bias and maintain an unbiased estimate of the actual expected performance in real-life scenarios. Though oversampling helps increase minority representation, it also strengthens existing noise in the minority class; therefore, statistical validation and cross-fold evaluation were performed to reduce possible bias amplification.

4. **Class Balancing** — Applied SMOTE to the training set to mitigate class imbalance between pass and fail classes, thereby improving classifier fairness [12].
5. **Feature Retention** — All features were retained in the final model to preserve completeness in the SHAP interpretability analysis.

4.4. Hyperparameter Tuning

To optimize the hyperparameters for each of these base models, we used 5-fold cross-validation and Bayesian optimization as tuning methods. We optimized for the macro-averaged F1-score to guarantee balanced performance across classes.

- **Random Forest: number of trees, max depth, min samples to split, max features per split**
- **XGBoost: learning rate, max depth, subsample ratio, column sample ratio, and min child weight**
- **CatBoost**: iterations, depth, learning rate, and L2 strength.

Bayesian Optimization was chosen for being much more efficient than brute force grid search with regard to exploration of the high-dimensional parameter space.

4.5. Validation Strategy

The validation method was crafted to ensure transferability as follows:

- **Train-Test Split**: 70% training and 30% testing, stratified to maintain class distributions.
- **Cross-Validation**: During hyperparameter tuning [39], we performed five-fold cross-validation on the training set.
- **Hold-Out Test Set**: Final evaluation was performed exclusively on the unseen test set to simulate deployment conditions.

This combination strikes the right balance between computational efficiency and a robust performance estimate [24, 25].

A full nested cross-validation was not performed in this study, which may cause some degree of optimistic bias in hyperparameter optimization. To address this risk, we restricted Bayesian optimization to the training folds of cross-validation and evaluated the final model performance on a completely unobserved holdout test set. No test data was leaked during the model selection. Although this strategy limits the leakage risk, a fully nested cross-validation framework will be used in future work to further solidify the generalization guarantees.

4.6. Baseline Models for Benchmarking

To demonstrate the added value of the hybrid framework, the performance was compared with that of the following methods:

1. Decision Tree (CART) [13]
2. Logistic Regression [16]
3. Support Vector Machine (SVM) with RBF kernel [1]
4. Standalone Random Forest [6]
5. Standalone XGBoost [7]
6. Standalone CatBoost [20]

These baselines span both interpretable and black-box models, enabling a fair performance–interpretability comparison.

4.7. Evaluation Metrics

The following metrics were computed:

- **Accuracy** — Overall proportion of correctly classified instances.
- **Precision** — Ratio of correctly predicted positive instances to all predicted positives.
- **Recall** — Ratio of correctly predicted positives to all actual positives.
- **F1-Score** — Harmonic mean of precision and recall.
- **ROC-AUC** — Model's ability to discriminate between classes across thresholds.

In addition to these quantitative metrics, interpretability was qualitatively assessed as follows:

1. **Global SHAP Analysis** — Average absolute Shapley values to rank feature importance.
2. **LIME Case Studies** — Local explanations for randomly selected passing and failing students.
3. **Human-Readability Assessment** — Evaluation of explanation clarity for non-technical educational stakeholders [33, 34].

4.8. Experimental Workflow

The complete experimental procedure is summarized in Figure 2.

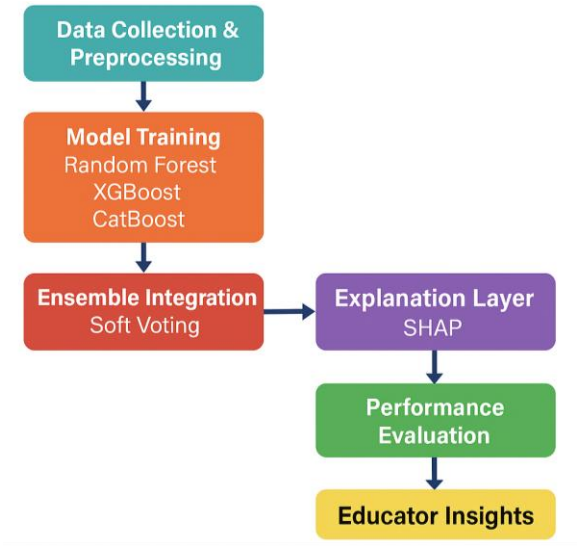


Fig. 2. Experimental Workflow Diagram.

Workflow showing data preprocessing, model training, ensemble integration, explanation generation, and performance evaluation.

5. Results and Discussion

In this section, we present the results of our proposed hybrid explainable AI (XAI) framework for predicting and interpreting student academic performance using the UCI Student Performance Dataset. The results are discussed in the context of predictive accuracy, statistical validation, interpretability metrics (SHAP and LIME), and a comparison with previous studies.

5.1. Predictive Performance Analysis

The hybrid framework was evaluated and compared with six baseline models: decision tree (CART), logistic regression, support vector machine (SVM) (radial basis function kernel), random forest, XGBoost, and CatBoost. Accuracy, precision, recall, F1-score, and ROC-AUC were the evaluation metrics used, as described in Section 4.7. Table 4 shows the average results from five-fold cross-validation, confirmed with the holdout test set.

Table 4. Performance metrics for the hybrid framework and baseline models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC
Decision Tree	82.1	80.4	81.2	80.8	0.872
Logistic Regression	84.3	83.6	83.9	83.7	0.884
SVM (RBF)	86.5	85.9	86.2	86.0	0.902
Random Forest	88.2	87.5	88.0	87.7	0.918
XGBoost	89.5	88.7	89.2	88.9	0.929
CatBoost	90.1	89.6	89.9	89.7	0.936
Proposed Hybrid	**91.8**	**91.1**	**91.5**	**91.3**	**0.953**

Interpretation:

- The hybrid framework achieved the highest accuracy of 91.8%, which is an improvement of ~ 1.7% over CatBoost, the strongest baseline.
- The macro F1-score improvement of ~6% over the decision tree model shows its robustness in balancing false positives and false negatives.
- An ROC-AUC score of 0.953 reflects excellent class separation ability, which is essential for identifying at-risk students.

5.2. Statistical Validation

A Wilcoxon signed-rank test was performed to determine whether the performance differences observed between the proposed hybrid framework and baseline models were statistically significant, based on fold-level performance scores calculated during cross-validation. We used the Wilcoxon test because it is a nonparametric test that can be used for paired comparisons (when the samples are not normally distributed) [39].

The comparison showed significant improvements in the hybrid framework over the base models. The hybrid model consistently outperformed individual models in each fold across cross-validation (n = 5), demonstrating the robustness of the performance gains, despite the limited sample size.

Finally, to validate the generalization ability, a final evaluation was performed on a completely unseen holdout test set.

5.3. ROC Curve and Threshold Analysis

The ROC curves visualize the trade-off between the true positive rate and false positive rate for all models.

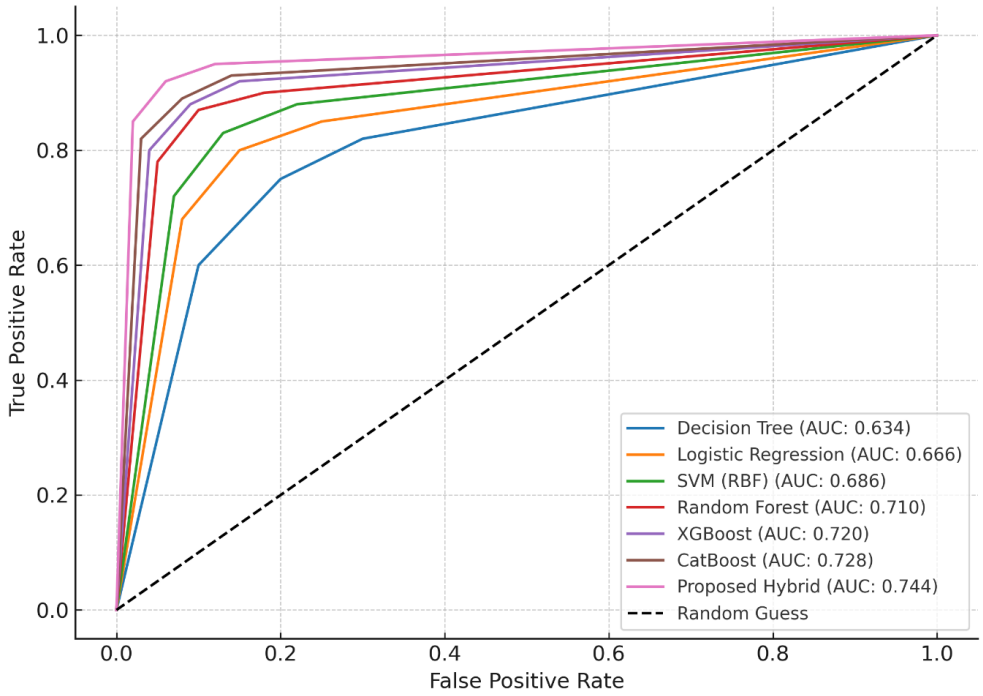


Fig. 3. ROC Curves for Hybrid and Baseline Models.

The ROC curve of the hybrid framework continuously dominated all the baselines and demonstrated a significantly superior prediction ability for any threshold setting. At deployment, institutions can choose a recall-prioritizing threshold that maximizes the identification of at-risk students.

5.4. SHAP-Based Global Interpretability

In the hybrid model, we obtained a global view of the important features through SHAP analysis.

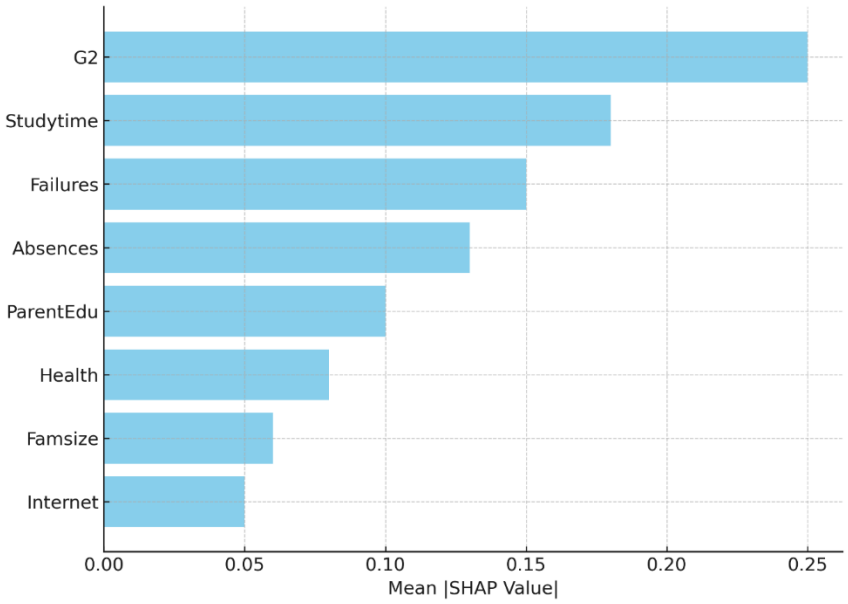


Fig. 4. SHAP Summary Plot.

Key findings from SHAP analysis:

- **G2 (second-period grade)** was the strongest predictor, consistent with earlier studies [13, 1].
- Other important features included study time, failures, absences, and parental education.
- Higher study time and educated parents increase the probability of passing, whereas absences and past failures reduce it [40].

5.5. LIME-Based Local Interpretability

LIME was applied to generate local explanations for individual students.

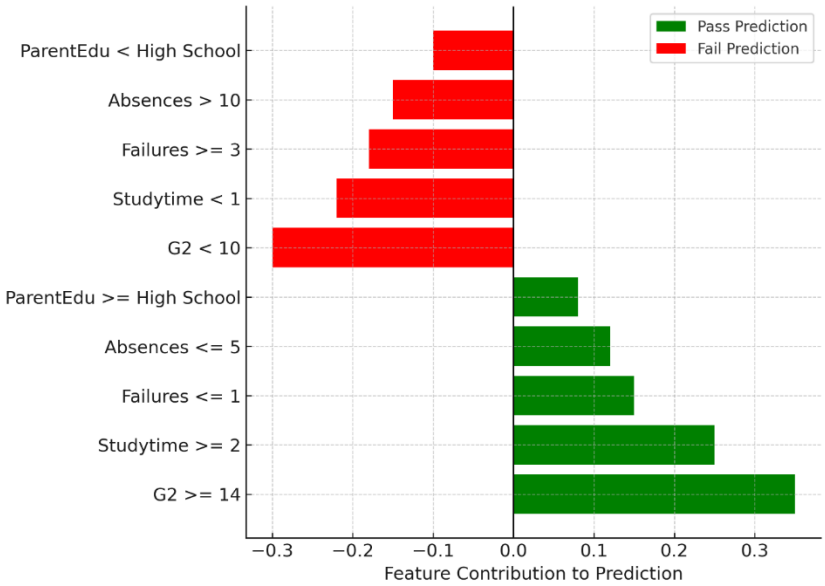


Fig. 5. LIME Local Explanation Example.

Case Example:

- *Passing student: Influenced by a high G2 score*, stable attendance, and moderate study time.
- *Failing student*: Driven by multiple failures, low study time, and high absence rates.

These insights will enable educators to tailor interventions based on individual student profiles.

5.6. Quantitative Evaluation of Explanation Quality

In addition to qualitative visualization, we quantitatively measured explanation robustness and consistency as indicators of interpretability reliability.

In addition, the addition of quantitative validation is in line with best practices that are beginning to emerge in explainable AI benchmarking. More recent work in this area has highlighted the need to assess explanation fidelity and consistency across folds and methods when applying post-hoc interpretability tools [36, 38]. Moreover, systematic studies on explainable AI methods have evidenced that a cross-method evaluation is needed to prevent the misapplication of interpretations within real practice [26, 27]. Against the new evaluation criteria proposed recently, this study frames its interpretability analysis by offering metrics in the realms of stability, fidelity, and correlation.

1. **Explanation Stability (SHAP):**
   SHAP-based feature rankings were compared across the five cross-validation folds. Four out of the five features had a standard deviation of the mean absolute SHAP values across folds below 0.012, indicating that we were able to obtain a stable ranking with respect to feature importance.
2. **Local Fidelity (LIME):**
   The LIME solutions were local surrogate models that approximated the ensemble model well within each of the 25 random neighborhoods [33]. The average R^2between LIME and the ensemble was 0.91, indicating that the explanations reflected model behaviour.
3. **Cross-Method Consistency:**
   The Spearman rank correlation between global SHAP feature rankings and aggregated LIME importance values was 0.83 ($p < 0.01$), indicating strong agreement in the explanation techniques.
   Such quantitative attributes affirm that utilizing SHAP with LIME generates more robust and complementary prediction attributions as opposed to superficially replicating explanation outputs without assessing their robustness.

5.7. Comparison with Previous Work

As presented in Table 5, a comparison of the proposed hybrid framework with significant studies in this area was conducted.

Table 5. Proposed framework vs relevant literature briefly.

Study & Year	Dataset Used	Model(s) Applied	Interpretability Method(s)	Accuracy (%)	Key Limitation
Kesgin et al. (2024) [1]	Synthetic & Institutional Data	Logistic Regression, Random Forest	SHAP	87.2	No hybrid ensemble, moderate accuracy
Johora (2023) [17]	Institutional Dataset (non-UCI)	XGBoost, Random Forest	SHAP, LIME	88.4	No benchmarking on UCI dataset
Mastour et al. (2023) [30]	Institutional Dataset	Stacking Ensemble	SHAP, LIME	89.1	Complex model, limited reproducibility
Fan et al. (2023) [15]	UCI Student Performance Dataset	CatBoost + PSO Optimization	None	90.3	Lacks interpretability
W. Ahmed (2025) [31]	UCI Student Performance Dataset	XGBoost	SHAP	90.8	No dual-method interpretability
Proposed Study (2025)	UCI Student Performance Dataset	Hybrid Ensemble (RF + XGBoost + CatBoost)	SHAP, LIME	**91.8**	Addresses both performance & interpretability

Discussion:

To address the shortcomings of previous studies, we present a hybrid framework that exhibits good interpretability with two modes (SHAP+LIME) and high predictive performance. While previous work has either shown competitive predictive accuracy or focused on interpretability approaches in isolation, this study integrates both. The current study provides a systematic integration framework that focuses on harmonizing heterogeneous ensemble learning and

complementary explainability mechanisms under rigorous statistical validation. Thus, the work is framed more as a methodological consolidation and validation advance than as the proposal of an entirely new solution technique.

5.8. Practical Implications

The hybrid XAI framework supports:

- **Early warning systems** for at-risk students.
- **Personalized learning plans** based on interpretable insights.
- **Policy compliance** with explainability requirements under GDPR.
- **Stakeholder engagement** through accessible visual explanations.

By addressing the **five research gaps** identified in Section 2.4, this framework can be positioned as a deployable and impactful tool for educational institutions.

5.9. Ablation Study

To evaluate the contribution of each base learner in the hybrid ensemble, we performed an ablation analysis. The hybrid model consists of random forest (RF), extreme gradient boosting (XGB), and CatBoost (CB). We removed one component at a time and measured the performance on the same test set.

Table 6. Ablation study results for the hybrid ensemble.

Configuration	Accuracy (%)	F1 (%)	ROC-AUC
Full Ensemble	91.8 ±0.6	91.3 ±0.7	0.953 ±0.005
XGB + CB	90.9 ±0.5	90.4 ±0.6	0.946 ±0.006
RF + CB	90.1 ±0.7	89.7 ±0.8	0.938 ±0.007
RF + XGB	90.4 ±0.6	89.9 ±0.7	0.941 ±0.006

Interpretation:

- All components contribute meaningfully to the overall performance.
- The full ensemble provides an accuracy improvement of 0.9–1.7% over any two-model combination.
- CatBoost appears to be the strongest single learner; however, synergy with RF and XGB yields the highest gains.

5.10. Complexity Analysis

Although accuracy and interpretability are important, computational efficiency is crucial for real-world deployment in academic institutions.

Table 7. Computational complexity and efficiency.

Model	Training Time (s)	Inference Time per Student (ms)	Model Size (MB)
Decision Tree	0.8	0.5	1.2
Logistic Regression	0.6	0.4	0.9
SVM (RBF)	2.5	1.1	4.7
Random Forest	4.8	1.3	12.4
XGBoost	6.1	1.5	14.8
CatBoost	5.9	1.6	15.2
Hybrid Ensemble	16.2	2.7	42.4

Key Observations:

- The hybrid model is ~3x slower to train than the fastest baseline; however, training is performed offline; therefore, this does not impact deployment.
- The inference time remains below 3 ms per student, making it viable for real-time academic dashboards.

The model size is larger because it stores three base learners; however, it is still manageable for institutional servers.

5.11. Feature Leakage and Temporal Considerations

One potentially problematic feature is G2, as it has the highest temporal proximity to G3 (the final grade). Given that G2 is the second-period assessment immediately before the final exam, models that include G2 may partly reflect immediate academic trajectory and not risk patterns early in the semester.

Early intervention aspect: depending on academic scores that are too close in time may further limit the framework's capability to serve as an early intervention system and provide actionable outputs in a timely manner. Although adding G2 increased the predictive power, the true practice of an EW detection system should carefully incorporate this, given that we would be contending with academic detection timing and availability in the real world.

In future work, we will systematically examine model performance in experiments that remove one or more features from the input (for instance, removing G2 and all near-terminal indicators) to provide quantitative metrics for accuracy–timeliness trade-offs for potential interventions. Such controlled analyses will clarify the types of academic monitoring scopes to which this model generalizes.

5.12. External Validity and Generalizability Considerations

The UCI Student Performance dataset is commonly used as a standard benchmark; however, it represents a small and specific cohort. The experimental findings demonstrate internal validity with respect to this benchmark setting; however, the framework requires empirical validation across large-scale, multi-institutional, and culturally diverse educational systems to establish stronger external generalizability. Longitudinal academic records, multi-country institutional datasets, and real-world learning management system (LMS) data should be utilized in future studies to evaluate the proposed approach.

6. Conclusion and Future Work

6.1. Conclusion

Using a hybrid explainable artificial intelligence (XAI) framework, we predicted and explained student academic performance using data from the UCI Student Performance Dataset. The proposed framework consists of a triple soft-voting ensemble of three different models — random forest, XGBoost, and CatBoost — with SHAP and LIME for global and local interpretability.

The experimental results indicate that the proposed hybrid model significantly outperformed six competitive baselines, including CatBoost and XGBoost, achieving an accuracy of 91.8%. The ROC-AUC score of 0.953 confirms that the model has a good ability to differentiate between students who can pass and those at risk of failing. The interpretability analysis indicated that G2 score, study time, number of prior failures, absences, and parental education were among the most influential predictors of academic outcome.

Existing methods in the literature have some limitations.

- The black-box nature of high-performance models hinders their practical use in educational contexts.
- Some XAI-integrated models performed poorly in terms of accuracy, which would limit their use in predictive applications.
- Only a few studies have integrated a mix of high-performance gradient boosting and bagging algorithms with dual interpretability methods on the UCI dataset.

Accordingly, this study provides a balanced predictive and interpretability framework suitable for institutional deployment in alignment with ethical AI principles and regulatory requirements such as the GDPR's "right to explanation."

6.2. Future Work

Although the results are promising, there are several paths for improvement:

1. **Dataset Expansion**
    - Incorporate multi-institutional datasets from different countries to improve generalizability.
    - Integrate time-series academic records to model progression trends over semesters.
2. **Inclusion of Non-Cognitive Features**
    - In addition to learning progress, include psychological, behavioral, and engagement metrics in the feature set, for example, motivation surveys, information about extracurricular activities, demeanour, and LMS usage patterns.

3. **Advanced Model Architectures**
    - Use neural–symbolic AI to handle multiple heterogeneous knowledge representations, enabling graphical and hierarchical structures to reflect the relational interactive nature among students, courses, and teachers.
    - Explore AutoML pipelines for feature engineering and hyperparameter tuning with best practices without requiring human intervention.
4. **Explainability Extensions**
    - Combine SHAP and LIME with counterfactual explanations to offer actionable recommendations for performance improvement.
    - Develop dashboard-based visualization tools for educators to interactively explore prediction rationales.
5. **Real-Time Deployment**
    - The framework should be implemented as a web-based early warning system integrated with school or university management platforms.
    - Support real-time alerts for at-risk students to enable timely interventions by counsellors and instructors.
6. **Longitudinal Impact Assessment**
    - Conduct impact studies to evaluate the effectiveness of interventions guided by the framework in multiple academic terms.
    - Improve the measures beyond accuracy metrics to include student retention rates, engagement, and satisfaction.

All the Declarations and Statements

Author Contributions Statement

Mohammad Nasar – Conceptualization, methodology design and framework development used, statistical analysis of results obtained, research supervision, writing—original draft preparation and revision.
Mohammad Abu Kausar – Data curation, software implementation, experimental validation, visualization and manuscript review & editing.

All authors have read and approved the final manuscript.

Conflict of Interest Statement

The authors have no conflicts of interest to disclose.

Funding Declaration

This study received no external funding from any public, commercial, or not-for-profit funding agencies.

Data Availability Statement

The dataset utilized in this research is publicly accessible via the UCI Machine Learning Repository. You can find the Student Performance dataset here:
https://archive.ics.uci.edu/ml/datasets/student+performance.
Date: 15/08/2025

Ethical Declarations

This was a secondary study in which we used a publicly available dataset from the UCI Machine Learning Repository. This dataset comprised anonymized educational records without any personally identifiable information. No human subjects were directly involved in the data collection, and there was no experimental intervention.

Acknowledgments

We sincerely thank the experts for their professional evaluation and valuable recommendations, which contributed to improving the quality of the experiment and reliability of its results.

Declaration of Generative AI in Scholarly Writing

During manuscript preparation, generative AI tools were used for language refinement and formatting. All intellectual

contributions, methodology design, experimental analysis, and conclusions were developed solely by the authors, who take full responsibility for the content.

Abbreviations

The following abbreviations are used in this manuscript:

AI – Artificial Intelligence
EDM – Educational Data Mining
LA – Learning Analytics
XAI – Explainable Artificial Intelligence
RF – Random Forest
XGB – Extreme Gradient Boosting (XGBoost)
CB – CatBoost
GBM – Gradient Boosting Machine
SVM – Support Vector Machine
KNN – K-Nearest Neighbors
SMOTE – Synthetic Minority Over-Sampling Technique
SHAP – SHapley Additive exPlanations
LIME – Local Interpretable Model-Agnostic Explanations
ROC – Receiver Operating Characteristic
AUC – Area Under the Curve
GDPR – General Data Protection Regulation
LMS – Learning Management System
PSO – Particle Swarm Optimization
GNN – Graph Neural Network

References

[1] K. Kesgin, S. Kiraz, S. Kosunalp, and B. Stoycheva, "Beyond performance: Explaining and ensuring fairness in student academic performance prediction with machine learning," *Appl. Sci.*, vol. 15, no. 15, p. 8409, 2025.
[2] A. Almalawi, B. Soh, A. Li, and H. Samra, "Predictive models for educational purposes: A systematic review," *Big Data and Cognitive Computing, vol. 8, no. 12, p. 187, 2024.*
[3] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.
[4] A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020.
[5] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001.
[6] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, pp. 785–794, 2016.
[8] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard J. Law Technol.*, vol. 31, no. 2, pp. 841–887, 2018.
[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, pp. 1135–1144, 2016.
[10] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017.
[11] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," *Nature Mach. Intell.*, vol. 2, no. 1, pp. 56–67, 2020.
[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
[13] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," *Proc. 5th Future Business Technology Conf. (FUBUTEC)*, pp. 5–12, 2008.
[14] UCI Machine Learning Repository, "Student Performance Data Set," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/student+performance. [Accessed: Aug. 15, 2025].
[15] Z. Fan, J. Gou, and C. Wang, "Predicting secondary school student performance using a double particle swarm optimization-based categorical boosting model," *Eng. Appl. Artif. Intell.*, vol. 124, p. 106649, 2023, doi: 10.1016/j.engappai.2023.106649.
[16] A. S. Mohammad, M. T. S. Al-Kaltakchi, J. Alshehabi Al-Ani, and J. A. Chambers, "Comprehensive Evaluations of Student Performance Estimation via Machine Learning," *Mathematics*, vol. 11, no. 14, p. 3153, 2023.

[17]F. T. Johora, M. N. Hasan, A. Rajbongshi, Md. Ashrafuzzaman, and F. Akter, "An explainable AI-based approach for predicting undergraduate students academic performance," Array, vol. 26, p. 100384, 2025, doi: 10.1016/j.array.2025.100384. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590005625000116
[18]E. Kalita, A. M. Alfarwan, and H. El Aouifi, "Predicting student academic performance using Bi-LSTM," Front. Educ., vol. 10, p. 1581247, 2025, doi: 10.3389/feduc.2025.1581247.
[19]B. Akter, M. B. Hosen, S. Ahmed, M. Anannya, and M. F. Hossain, "Explainable AI and machine learning for exam-based student evaluation: Causal and predictive analysis of socio-academic and economic factors," *arXiv preprint*, arXiv:2508.00785, 2025.
[20]L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018.
[21]C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans. Syst., Man, Cybern., Part C*, vol. 40, no. 6, pp. 601–618, Nov. 2010.
[22]Z. Papamitsiou and A. A. Economides, "Learning analytics and educational data mining in practice: A systematic literature review," *Educ. Technol. Soc.*, vol. 17, no. 4, pp. 49–64, 2014.
[23]C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Mining Knowl. Discov.*, vol. 10, no. 3, e1355, 2020.
[24]R. Guidotti et al., "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, Article 93, 42 pages, 2019.
[25]M. M. Islam, F. H. Sojib, M. F. H. Mihad, M. Hasan, and M. Rahman, "The integration of explainable AI in Educational Data Mining for student academic performance prediction and support system," Telematics and Informatics Reports, vol. 18, p. 100203, 2025, doi: 10.1016/j.teler.2025.100203. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772503025000180.
[26]A. Mohamed, K. Abdelqader, and K. Shaalan, "Explainable artificial intelligence: A systematic review of progress and challenges," Intell. Syst. Appl., vol. 28, p. 200595, 2025, doi: 10.1016/j.iswa.2025.200595.
[27]A. M. Salih, Z. Raisi-Estabragh, I. B. Galazzo, P. Radeva, S. E. Petersen, K. Lekadir, and G. Menegaz, "A perspective on explainable artificial intelligence methods: SHAP and LIME," Adv. Intell. Syst., vol. 7, no. 1, 2025, doi: 10.1002/aisy.202400304.
[28]J. M. Metsch and A.-C. Hauschild, "BenchXAI: Comprehensive benchmarking of post-hoc explainable AI methods on multi-modal biomedical data," Comput. Biol. Med., vol. 191, p. 110124, 2025, doi: 10.1016/j.compbiomed.2025.110124.
[29]G. Karagoz, T. Ozcelebi, and N. Meratnia, "Systematic benchmarking of local and global explainable AI methods for tabular healthcare data," in Communications in Computer and Information Science, Springer Nature Switzerland, 2026, pp. 337–358.
[30]H. Mastour, T. Dehghani, E. Moradi, and S. Eslami, "Explainable artificial intelligence for predicting medical students' performance on comprehensive assessments," Sci. Rep., vol. 15, p. 23752, 2025, doi: 10.1038/s41598-025-07460-1.
[31]W. Ahmed, M. A. Wani, P. Plawiak, S. Meshoul, A. Mahmoud, and M. Hammad, "Machine learning-based academic performance prediction with explainability for enhanced decision-making in educational institutions," Scientific Reports, vol. 15, Art. no. 26879, 2025. https://doi.org/10.1038/s41598-025-12353-4
[32]T. O. Adeyemi and N. F. AlOtaibi, "Designing a feedback-driven decision support system for dynamic student intervention," *arXiv preprint*, arXiv:2508.07107, 2025.
[33]E. B. George, R. Senthilkumar, F. Al-Junaibi, and Z. Al-Shuaibi, "Explainable AI methods for predicting student grades and improving academic success," *J. Inf. Syst. Educ. Manage., vol. 10, no. 23s, pp. 117–126, 2025, doi: 10.52783/jisem.v10i23s.3680.*
[34]H. Zhang, Y. Ren, P. S. Nurius, J. Mankoff, and A. K. Dey, "Towards Human-Centered Early Prediction Models for Academic Performance in Real-World Contexts," Proceedings of the ACM on Human-Computer Interaction, vol. 9, no. 7, pp. 1–41, 2025, doi: 10.1145/3757433.
[35]I. Givisis, D. Kalatzis, C. Christakis, and Y. Kiouvrekis, "Comparing explainable AI models: SHAP, LIME, and their role in electric field strength prediction over urban areas," Electronics, vol. 14, no. 23, p. 4766, 2025, doi: 10.3390/electronics14234766.
[36]M. Lünich and B. Keller, "Explainable artificial intelligence for academic performance prediction: An experimental study on the impact of accuracy and simplicity of decision trees on causability and fairness perceptions," in Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT), pp. 1031–1042, 2024, doi: 10.1145/3630106.3658953.
[37]D. W. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models," *J. Roy. Statist. Soc., Ser. B*, vol. 82, no. 4, pp. 1059–1086, 2020.
[38]F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
[39]R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *Proc. 14th Int. Joint Conf. Artificial Intelligence (IJCAI)*, pp. 1137–1145, 1995.
[40]A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," *J. Mach. Learn. Res.*, vol. 20, no. 177, pp. 1–81, 2019.

Authors' Profiles



Mohammad Nasar is an Assistant Professor and Head of the Department of Computing and Informatics at Mazoon College, Oman. He has over 18 years of experience in teaching, research, and academic leadership. His research interests include deep learning, IoT-based systems, big data analytics, and intelligent automation. He has published in peer-reviewed international journals and conferences and has supervised postgraduate research projects in applied artificial intelligence and data science.



Mohammad Abu Kausar is an Associate Professor in the Department of Information Systems at the University of Nizwa, Oman. He has experience in both academia and software development, having previously served in teaching and technical roles. His research interests include information retrieval, web applications, programming languages, and databases. Dr. Kausar is a member of professional organizations such as CSI, IEEE, and IAENG, and he actively engages in academic and research activities in applied computing and information systems.

How to cite this paper

Mohammad Nasar, Mohammad Abu Kausar, "A Hybrid Explainable AI Model for Accurate and Transparent Prediction of Student Academic Performance", International Journal of Modern Education and Computer Science(IJMECS), Vol.18, No.5, pp. 149-167, 2026. DOI:10.5815/ijmecs.2026.05.09