International Journal of Innovative Research and Scientific Studies, 9(3) 2026, pages: 84-105

IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

An explainable hybrid machine learning model for data-driven assessment and enhancement of program learning outcomes in higher education

Awad M. Awadelkarim[1*], Khalid Al-Otaibi[1], Hafed Albalawi[1], Mohammed Mustafa[1], Anas Bushnag[1]

[1]*College of Computing and Information Technology, University of Tabuk, Tabuk 71491, Saudi Arabia.*

Corresponding author: Awad M. Awadelkarim (*Email:* awad@ut.edu.sa)

Abstract

The purpose of this study is to develop and validate an explainable hybrid machine learning framework for accurately assessing and enhancing Program Learning Outcomes (PLOs) in higher education. The study aims to overcome the limitations of conventional manual or heuristic evaluation methods by leveraging data-driven predictive analytics to identify key factors influencing student achievement. A stacked ensemble learning architecture is proposed, combining multiple gradient boosting and tree-based algorithms: LightGBM, XGBoost, CatBoost, Gradient Boosting, and Decision Tree, under a multinomial Logistic Regression meta-learner. The model was trained and tested on real academic data collected from the University of Tabuk, Saudi Arabia, incorporating academic, behavioral, and demographic variables. Comprehensive preprocessing, stratified k fold cross validation, and grid search optimization were applied to enhance robustness and generalization. SHapley Additive exPlanations (SHAP) were used to interpret model outputs and determine the relative importance of predictors. The hybrid model achieved a micro average ROC AUC of 0.998, along with consistently high precision, recall, and F1 scores across all grade categories (A–F). SHAP analysis revealed that Total Score, Project Score, and Final Score were the strongest predictors of PLO attainment, offering a clear insight into the learning dimensions that contribute most to academic success. Results confirm that the proposed hybrid ensemble outperforms conventional single model and deep learning approaches in both predictive precision and interpretability. By combining accuracy with transparency, the model serves as a valid analytical tool for institutional quality assurance and outcome based education. This framework enables educators and program evaluators to make data driven, evidence based decisions for early identification of at risk students, curriculum refinement, and continuous improvement of teaching strategies. It also provides a replicable methodology for integrating explainable AI into academic performance assessment in higher education institutions.

Keywords: Hybrid model, Machine learning, Program learning outcomes (PLOs), Stacked ensemble, SHAP, Student performance prediction.

DOI: 10.53894/ijirss.v9i3.11393
Funding: This study received no specific financial support.
History: Received: 9 January 2026 / **Revised:** 20 February 2026 / **Accepted:** 24 February 2026 / **Published:** 24 March 2026
Copyright: © 2026 by the authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).
Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Acknowledgements: The authors would like to thank the University of Tabuk (UoT) for their encouragement, for providing the data, and for offering valuable technical resources for data analysis, which were crucial to the successful completion of this study.
Publisher: Innovative Research Publishing

1. Introduction

With the rapid proliferation of digitised content and personalised learning environments, data-driven technologies are now entering classrooms in a variety of ways, leading to the emerging field of Educational Data Mining (EDM) and Learning Analytics (LA) that use statistical analysis and machine learning techniques on educational datasets to inform educators' decision-making [1]. These have been used more recently for tracking learning behaviors, predicting student performance, evaluating academic risk, and informing data-driven decision-making in teaching and in curriculum design [2]. The predictive validity of student achievement is a key issue in the evaluation and assurance of the quality of institutional performance, particularly in higher education contexts where academic success is strongly correlated with resource allocation, retention rates, and accreditation characteristics [3]. Due to the increasing availability of digital learning environments and e-assessment platforms, vast amounts of student data are stored (e.g., academic scores, behavioral indicators, demographic information, engagement logs), which can be fed into computational models [4]. Using such data in sophisticated predictive models, educators can identify early who is at risk, tailor interventions, and improve learning for everyone [5]. However, the prediction performance and generalizability are still far from satisfactory because of the heterogeneity of data sources, the complexity of features, as well as drawbacks in traditional classifiers [6].

Recent studies have shown that being able to generate accurate predictions on student overall performance is possible using single machine learning algorithms such as Decision Trees, Support Vector Machines (SVM), and Gradient Boosting [7-9]. However, these models suffer from a high variance or bias depending on input features and dataset formation [10, 11]. In reality, academic datasets are small and imbalanced [12], with many dimensions, so that a single method is unlikely to be able to handle the nonlinear relationships between values found here, including scores, participation (attendance), and behavior. Thus, hybrid models and ensemble learning are promising options. Stacked ensembles can combine the powers of various algorithms as well as reduce their weaknesses using multiple base learners with a meta-model [13, 14]. In addition, using explainability methods such as SHapley Additive exPlanations (SHAP) provides transparency to educators so they can interpret decisions made by the model and know which features have the biggest impact on student performance [15, 16].

Though these methods have shown potential, there are relatively few works on hybrid methods that can simultaneously offer predictive performance, interpretability, and computational reduction for the task of student performance classification. Many studies use only one method and are solely performed on limited academic datasets, leading to overfitting and a lack of generalization. Moreover, most of the existing models focus on predictive accuracy and hardly emphasize interpretability, with educators facing unclear reasons for the x-variable relation to learning outcomes. Lastly, there is a notable absence of the comprehensiveness of academic attributes and behavior as a determining factor of learner success. Furthermore, some frameworks do not use stratified cross-validation and fail to let a separate feature relevance analysis deprive them of reproducibility and practical utility. For these reasons, there appears to be a demand for a hybrid, explainable machine learning model that achieves a high predictive score and sheds light on the key factors of student learning outcomes.

To bridge the aforementioned gaps, this study proposes a hybrid machine learning Model using a stacked ensemble consisting of multiple gradient boosting and tree-based classifiers with a logistic regression meta-model. As such, the main contributions of this study can be summarized as follows:

- This research suggested a stacked ensemble model that integrates multiple gradient boosting and tree-based learners (LightGBM, XGBoost, CatBoost, Gradient Boosting, and Decision Tree) under a Logistic Regression meta-learner. Unlike conventional ensemble approaches, the proposed framework synergizes model diversity and interpretability for multi-class grade prediction within an educational context.
- An innovative experimental design was used to predict academic grades with the help of a dataset of social media usage and emotional well-being. This integration of behavioral and academic metrics reaches further into the educational information mining tradition and evidences the ability of the model to process complex data of the learner.
- This study proposed a two-tier optimization process that is integrated by Grid Search with Stratified k-Fold Cross-Validation to both select balanced datasets and identify optimal hyperparameters, simultaneously. This technique can dramatically reduce overfitting and thus help the trained model generalize better for the imbalanced academic data.

- The study employed SHAP (SHapley Additive exPlanations) for not only the hybrid model, but also base learners to explain feature importance separately. This interpretable dual viewpoint connects the worlds of predictive analytics and pedagogy – it renders tangible and actionable machine learning results to educators.
- The proposed framework was designed and validated using real academic program data from the University of Tabuk, Saudi Arabia.

The rest of this paper is divided into the following sections: Section 2 contains an extensive literature review, which provides an overview of previous literature on educational data mining, student performance prediction, and a hybrid machine learning approach, along with the methodological trends and existing research gaps. Section 3 presents the methodology, which describes the proposed hybrid model architecture, the dataset's description and preparation process, model training workflow, hyperparameter tuning, and experimental environment. Section 4 presents the results and findings, including EDA, experimental investigations, comparative performance analysis, and interpretation based on SHAP. Section 5 concludes with an overview of the main findings, their implications in terms of practical relevance for educational enhancement, and possible future research paths.

2. Literature Review

The predictive modeling of student performance has emerged as a core research topic in Educational Data Mining (EDM) and Learning Analytics due to the increased demand to improve learning outcomes and maximize the pedagogical interventions. In the last 10 years, there has been a growing range of research emphasizing the application of machine learning (ML) and artificial intelligence (AI) methods to predict, model, and interpret academic achievement in various ways, depending on different influences on academic outcomes that include educational, behavioral, and cognitive. Such studies have assumed the use of various computational systems, including classical classifiers and fuzzy inference systems, and modern ensemble and deep learning systems, to achieve better predictive performance and help make evidence-based decisions in education. This section is a critical review of the major contributions of the recent literature, their methodology, performance, limitations, and gaps in research that may constitute the basis of promotion of the proposed hybrid explainable model.

Kowalska, et al. [17] performed a machine learning analysis to categorize and determine the learning results in higher education based on the European Qualification Framework. The study was effective in detecting misclassified learning outcomes in the areas of knowledge, skills, and social responsibility, with a strong sensitivity and specificity of about 0.8 using TF-INF to vectorize the texts and various classification algorithms. Nevertheless, the study was limited because it used textual information on a limited group of 22 universities, and this might limit the ability of the model to be generalized to the larger academic setting.

Goyal, et al. [18] proposed a fuzzy inference-based method to assess the association between course attainment parameters (CAPs) and program attainment parameters (PAPs) in engineering education. The study created a systematic evaluation framework to deal with imprecise or vague correlations by including twelve PAP indicators and three correlated CAP measures. The presented system of fuzzy logic was a good representation of complex educational performance dynamics, though its validation was only allowed to be applied over a period of one year in one discipline, which limits long-term and cross-domain generalizability.

Zaki, et al. [19] designed an AI-based model utilizing Natural Language Processing (NLP) in the automation and verification of mapping Course Learning Outcomes (CLOs) to Program Learning Outcomes (PLOs) as a quality assurance measure in higher education. The system was tested based on datasets provided by two e.g. educational programs and has a high mapping accuracy of 83.1% and 88.1%, which is quite close to the ratings of the experts. Although it was effective, the main limitation of the study is its limited scope of data and the possibility of being further validated in various academic fields and linguistic contexts.

Pallathadka, et al. [20] used different machine learning classification models such as Naive Bayes, ID3, C4.5, and SVM to estimate the performance of students based on the UCI student performance dataset. It was shown that data mining and predictive modeling are useful in recognizing at-risk students and improving academic decision-making. The work, however, was constrained by the traditional algorithm and small datasets used, which did not incorporate ensemble or hybrid methods that could enhance predictive accuracy and generalization.

Adnan, et al. [21] developed a predictive model based on various machine learning (ML) and deep learning (DL) algorithms to predict at-risk students in online learning courses, such as MOOCs, and LMS platforms. It compared the model performance according to the percentage of course progress and determined that the highest accuracy (up to 0.91) and better precision, recall, and F1-scores were obtained using Random Forest. Although the study has good performance, its weakness is the fact that it uses behavioral and engagement data without considering the affective or contextual learning variables that can further improve the accuracy of the prediction.

Yağcı [22] presented a machine learning model to estimate the end-of-term test scores of undergraduate students based on midterm, department, and faculty performance of 1,854 undergraduate students taking a course in the Turkish Language. Several algorithms were compared, such as Random Forest, SVM, Logistic Regression, Naive Bayes, and KNN, with the general accuracy of the classification being 70-75%. The paper has successfully shown that early academic risk prediction is feasible, but it lacks the scope to validate the model due to using a limited range of features and only a single course dataset, limiting future practitioner use and scalability.

Nayak, et al. [23] examined the student performance prediction in a system of Outcome-Based Education (OBE) using two datasets: behavioral attributes (Kalboard 360) and an institutional dataset with no behavioral data. There were a number of machine learning models like Decision Tree (J48), Naive Bayes, random Forest, and an optimized Multilayer Perceptron

(Opt-MLP) that were used in addition to feature selection methods like information gain and correlation analysis. Opt-MLP demonstrated a maximum of 97.08% accuracy, and Random Forest demonstrated 100% when the features of student engagement are involved, which highlights the importance of data on student engagement.

Lamb, et al. [24] came up with a new method of predicting student performance based on neurocognitive data measured by a functional near-infrared spectroscopy (fNIRS) in a Synthetic Adaptive Learning Environment (SALEs). The study was able to predict the results of cognitive responses by studying how students were responding in varying modes of instruction no content, video, and virtual reality, with an average predictive accuracy of 85% and a low error rate of less than 15%. This shows that real-time neurocognitive analytics could be helpful in adaptive learning, but the study had a limited scope due to a small sample (n=40) and the absence of a combination with other traditional academic or behavioral data sources, which limits the general applicability.

Kabathova and Drlik [25] have discussed the issue of student dropout prediction in e-learning settings, which has remained a challenge over time, by highlighting the importance of data quality and feature selection in the predictive models. They compared various machine learning classifiers using educational data gathered in four academic years and produced an accuracy in the classifier between 77-93% to predict course completers and non-completers with limited feature sets. The authors emphasized that predicting dropouts with high confidence can be achieved using the minimum amount of data when properly evaluated metrics are considered, but the results of the study are limited by the small size of the data and the risk of overfitting because of the small number of features.

Using more than 30,000 records of 74 teams, Giannakas, et al. [26] suggested a Deep Neural Network (DNN) system for early prediction of team performance in software engineering education. The model tested using several activation functions and optimizers showed a final accuracy of up to 82.39% and a learning performance of 86.57%, which proves the feasibility of deep learning in collaborative learning analytics. Moreover, SHAP analysis was used to understand the contribution of features, which makes the model more transparent. The study, however, is too close-ended on a single area (software engineering), and binary outcome prediction makes their results less generalizable in terms of application to education.

Hussain and Khan [27] investigated the application of supervised machine learning models to forecast the grades and marks of students based on past academic records of the Board of Intermediate and Secondary Education (BISE) Peshawar, which comprised seven regional divisions. Using regression and decision tree (DT) classification models, the research proved that the ML methods are effective in the accurate prediction of student performance and enhancement of educational planning. The study, however, had a geographical limitation of being confined only to one educational board, and no comparative experimental study with advanced models like ensemble or hybrid models.

Qiu, et al. [28] proposed the Behaviour Classification-based E-learning Performance (BCEP) prediction framework involving fusion and classification of behaviour features and process based on behaviour prediction to predict learning performance in real time. The proposed Process-Behaviour Classification (PBC) model, based on the Open University Learning Analytics Dataset (OULAD), showed a higher predictive accuracy than the traditional ones, emphasizing the importance of the modelling of behavioural interdependencies in online learning. Nevertheless, this research design had constrained its scope due to the reliance on behavioural log records and the absence of the need to combine them with cognitive or affective measures of learners, which can additionally deepen predictive validity.

Overall, the current literature illustrates that there has been considerable advancement in the application of machine learning and artificial intelligence to educational analytics, but it has a number of challenging issues. Most of the previous research has concentrated on the single-model methods or domain applications without paying much attention to the generalization and interpretation. Moreover, there are not many models that incorporate behavioral, academic, and contextual factors to be able to offer a comprehensive picture of the learning processes. Although the explainable AI-stances, including SHAP, are still emerging in more recent studies, they are not fully used as a part of the hybrid predictive framework. To solve these issues, a solid, explainable, and hybrid machine learning platform that can simultaneously fulfill both the high predictive accuracy and transparent interpretability goals is wanted, with the objective of which this study will be built. To provide an overview of current methodologies, findings, and the progress made towards developments in literature employing educational analytics for student performance prediction, Table 1 compares various literatures to demonstrate the range of model results.

Table 1 Summary of key studies related to machine learning approaches for student performance prediction and educational outcome analysis.

Table 1.
Summary of key studies related to machine learning approaches for student performance prediction and educational outcome analysis.

Reference	Model / Method Used	Key Findings	Limitations
Nayak, et al. [23]	Machine Learning classifiers with TF–IDF text preprocessing	Achieved ≈0.8 sensitivity and specificity in classifying learning outcomes; effectively identified formulation errors in outcome statements	Limited dataset (22 universities) and reliance on textual features restrict broader generalizability
Goyal, et al. [18]	Fuzzy Inference System (FIS) for mapping CAPs to PAPs	Demonstrated promising results for evaluating creative and collaborative skills	Applied only to one-year engineering data; lacks large-scale validation across disciplines
Zaki, et al. [19]	NLP-based automated CLO–PLO mapping system	Achieved 83.1% and 88.1% precision compared with expert mappings; demonstrated strong potential for automated quality assurance in education	Limited to two datasets; requires validation across varied academic and linguistic contexts
Pallathadka, et al. [20]	Naïve Bayes, ID3, C4.5, and SVM	Demonstrated effective student performance prediction using UCI dataset; helped identify at-risk students for academic support	Limited dataset and use of traditional algorithms; lacks ensemble or hybrid model comparison
Adnan, et al. [21]	Random Forest, ML & DL models for predictive analytics	Random Forest achieved up to 0.91 accuracy in identifying at-risk students across course phases; highlighted engagement and assessment as key factors	Focused mainly on behavioral data; excluded emotional and contextual learning features
Yağcı [22]	Random Forest, SVM, Logistic Regression, Naïve Bayes, KNN	Achieved 70–75% accuracy predicting final grades using midterm and institutional data; highlighted early detection of at-risk students	Limited to one course and small feature set; lacks scalability and generalization across disciplines
Nayak, et al. [23]	Decision Tree (J48), Naïve Bayes, Random Forest, Opt-MLP with feature selection	Opt-MLP achieved up to 97.08% accuracy; Random Forest reached 100% with behavioral features, showing behavior's strong role in learning outcomes	Possible overfitting; limited validation across institutions and datasets
Lamb, et al. [24]	fNIRS-based neurocognitive data with ML predictive modeling in SALEs	Achieved 85% predictive accuracy within 300 ms using real-time brain activity; demonstrated potential for adaptive AI tutoring systems	Small sample size (n=40); lacks integration with academic/behavioral features for comprehensive modeling
Kabathova and Drlik [25]	Multiple ML classifiers for dropout prediction	Achieved 77–93% accuracy using limited e-learning features across four academic years; emphasized importance of data quality and performance metric selection	Small, domain-specific datasets; possible overfitting and limited feature diversity
Giannakas, et al. [26]	Deep Neural Network (DNN) with SHAP interpretability	Achieved 82.39% accuracy and 86.57% learning performance for team performance prediction; enhanced explainability with SHAP	Focused on binary classification and a single domain; limited cross-context applicability
Hussain and Khan [27]	Regression and Decision Tree Classifier	Demonstrated accurate grade and marks prediction using 30 academic attributes from BISE Peshawar; validated ML's potential for educational forecasting	Limited to one regional dataset; lacked experimentation with ensemble or hybrid methods
Qiu, et al. [28]	Behaviour Classification-based E-learning Performance (BCEP) and Process-Behaviour Classification (PBC) models	Achieved superior prediction performance using OULAD dataset; emphasized feature fusion and behaviour interdependence in e-learning analytics	Dependent on behavioural data; lacks integration of cognitive and affective learner variables

3. Methodology

In order to offer a deep insight into the suggested hybrid framework, Figure 1 visualizes the overall methodological workflow. The figure shows the chronological nature of the research process, whereby data acquisition and preprocessing are followed by feature engineering, model development, and evaluation. The workflow illustrates how different machine learning algorithms were stacked as an ensemble architecture to use LightGBM, XGBoost, CatBoost, Gradient Boosting, and Decision Tree as their meta-learners. Each step of the process has a systematic contribution to the improvement of predictive performance, interpretability, and generalization in the model. The subsequent subsections describe in detail each of the mentioned components of this workflow, including the characteristics of the dataset, the methods of its preprocessing, the architecture of the hybrid model, the training and validation algorithms, and the analysis of its interpretability.

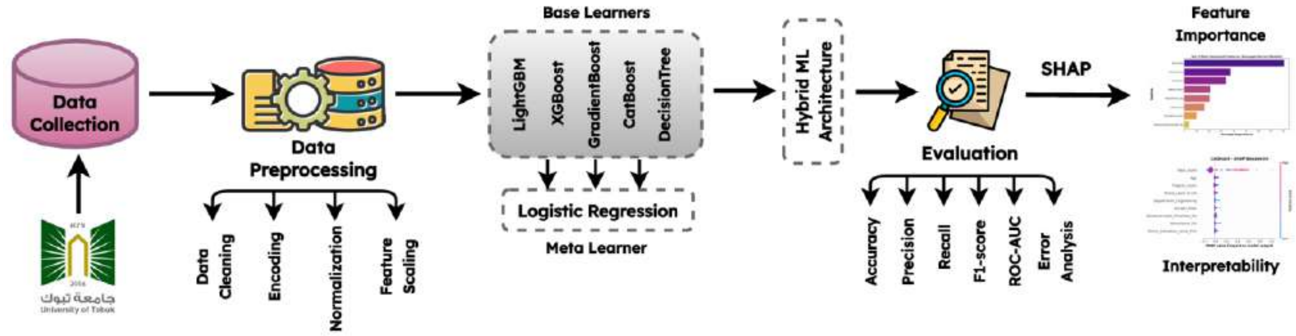

Figure 1.
Proposed methodology workflow.

3.1. Dataset Description

As mentioned earlier, the proposed framework was designed and validated using real academic program data from the University of Tabuk, Saudi Arabia. Accordingly, the study employed a dataset comprising the academic performance records of 1,646 University of Tabuk (UoT) students, including both male and female participants. Students' demographics, pre-university performance (MARK, TAHSEL_MARK, SCHOOL_AVG), and the most significant indicators of academic performance (course grades, average, GPA2, course attempts) are among the features. The entire feature set is described in Table 2.

Table 2.
Dataset Description.

Feature Name	Data Type	Description
STUDENT_ID	Integer	Unique identifier assigned by the university
MARK	Integer	Secondary school final score
TAHSEL_MARK	Integer	Achievement (Tahsili) test score
SCHOOL_AVG	Integer	Average school score
SCHOOL_CODE	Integer	School code identifier
JOIN_SEMESTER	Integer	Semester when the student joined
SCHOOL_TYPE	Integer	Type of school attended (1 = Private, 2 = Public)
GENDER2	Integer	Gender of student (1 = Male, 2 = Female)
COURSE_NO	Integer	First course code
COURSE_NO2	Integer	Second course code
COURSE_NO3	Integer	Third course code
COURSE_NO4	Integer	Fourth course code
Status, Status2, Status3, Status4	Object	Pass/Fail status for each course
N.of.Times to N.of.Times4	Object	Number of times the student repeated each course
Average	Float	Computed average from the selected course grades
GPA2	Float	GPA computed from the 4 course scores
GRADE2, Grade	Object	Categorical GPA grade (e.g., Excellent, Very Good, Good)
SCHOOL	Object	Type of school (Private or Public)
GENDER	Object	Gender (Male or Female)

3.2. Data Preprocessing

The preprocessing steps are done to make sure that raw educational data is cleaned, organized, and standardized, and then presented to the machine learning framework. The operations at this stage are important to reduce noise level, deal with missingness, and make sure that all features make an equal contribution to the predictive models. Each step of preprocessing is elaborated on in the succeeding subsections:

(i) Data Cleaning: Columns that are not informative Student_ID, First_Name, Last_Name, and Email, were dropped:

$$X = D\setminus\{Student\_ID, First\_Name, Last\_Name, Email\}$$

By taking these attributes out, any potential bias and flow of information due to the uniqueness of identifiers can be avoided, which do not have any predictive power.

(ii) One-hot Encoding Al-Shehari and Alsowail [29]: Categorical features have been automatically identified through their data type and encoded with one-hot encoding by setting the drop_first to 'True' in order to prevent collinearity.

If $x_j \in \{c_1, c_2, \dots, c_k\}$ is a categorical variable, then

$$x_{ij}^{(l)} = \begin{cases} 1, & x_{ij} = c_l \\ 0, & otherwise, \end{cases} \quad l = 1, \dots, k-1.$$

The final matrix is hence a combination of both numeric and binary variables to be ingested into the model.

(iii) Label Encoding Sailasya and Kumari [30]: LabelEncoder was used to encode the grades, with the textual classes being changed into integer labels, but maintaining the existing hierarchy of the performance:

$$y_i' = \phi(y_i) = \begin{cases} 0, & y_i = F, \\ 1, & y_i = D, \\ 2, & y_i = C, \\ 3, & y_i = B, \\ 4, & y_i = A. \end{cases}$$

The ML algorithms can decipher advancement in achievement levels numerically as a result of this transformation.

(iv) Noise Injection for Regularization Al-Gethami, et al. [31]: In order to improve model generalization and reduce overfitting, small random Gaussian noise was introduced to the whole feature matrix X:

$$X' = X + \epsilon, \ \ \epsilon \sim N(0, 0.015^2)$$

This step will be a simulation of slight variability in expected measurement in assessment scores so that models may be encouraged to learn more fixed characteristics instead of learning precise values of inputs.

(v) Train-Test Splitting: A stratified random partition was used to split the data into 80% training and 20% testing data sets to preserve the distribution of classes of grades:

$$(X_{train}, X_{test}, y_{train}, y_{test}) = Split(X', y', ratio = 0.8, stratify = y')$$

The stratification ensures equal representation at all grades within the splits and does not bias between frequent classes.

(vi) Feature Scaling Ahsan, et al. [32]: All the numeric variables were normalized through the z-score transformation by the use of StandardScaler:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_J}, \ \ \mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \ \ \sigma_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \mu_j\right)^2}$$

Scaling brings similarities between the magnitudes of features, which speeds the convergence process when using the particular gradient-based learners like LightGBM and XGBoost, and averts dominance of high-range variables.

3.3. Proposed Hybrid Model Architecture

The suggested hybrid architecture utilized a stacked ensemble learning architecture, where various heterogeneous base learners are joined together with one meta-model. This architecture is aimed at taking advantage of predictive properties of different complementary algorithms, where more robustness to classification and enhanced generalization are aimed at predicting Program Learning Outcome (PLO) achievement.

Formally, given a training dataset

$$D = \{(x_i, y_i)\}_{i=1}^{N},$$

where $x_i = [x_{i1}, x_{i2}, \dots, x_{id} \in R^d$ represents a feature vector for the *i-th* student and $y_i \in \{1,2,3,4,5\}$ denotes the corresponding class label (grades A-F), the ensemble framework constructs M base learners $\{h_1, h_2, \dots, h_M\}$, each trained independently on the same dataset.

3.4. Base Learners

The base layer is composed of five different machine learning algorithms, each with its own inductive bias and regularization. The outputs of these base learners constitute the meta-feature matrix for the following learning stage.

1. Light Gradient Boosting Machine (LightGBM) Li, et al. [33]: LightGBM is a fast and efficient gradient boosting framework. It constructs the additive decision trees through recursive steps, during which the differentiable loss function $L(y, \hat{y})$ is minimized by the process of gradient descent.

Optimization at iteration *t* can be given as:

$$f_t(x) = \arg\min_f \sum_{i=1}^{N}\left[g_i f(x_i) + \frac{1}{2} h_i f(x_i)^2\right] + \Omega(f)$$

where g_i and h_i represent the first and second derivatives of the loss with respect to the prediction $\hat{y}$, d $\Omega(f)$ represents the regularization expression that regulates the complexity of the model using parameters including num_leaves, learning rate and feature fraction.

2. Extreme Gradient Boosting (XGBoost)[10]: XGBoost is also a boosting algorithm that employs both L1 and L2 regularization to prevent overfitting:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 + \alpha \sum_{j}^{T=1} |w_j| \, ,$$

where T is the number of leaves, 2 will be the weights of the leaves, and will be hyperparameters of regularization (reg_alpha, reg_lambda). This makes sure that strong generalization is achieved and there is minimized model variance.

3. Categorical Boosting (CatBoost) Joshi, et al. [34]: CatBoost also adds variety to an ensemble specifically by supporting categorical variables using an ordered target statistics method. It decreases bias associated with predictions by generating symmetric oblivious trees by having all the nodes at the same level to share the same splitting criterion. The model is an optimization of a similar purpose gradient-based goal, but with an inbuilt ordered boosting and shrinkage to reduce overfitting.
4. Gradient Boosting Classifier (GBM) Mashagba, et al. [35]: The classical GradientBoostingClassifier from scikit-learn applies forward stagewise additive modeling. A weak learner $h_t(x)$ is introduced at each iteration to reduce the residual error:

$$F_t(x) = F_{t-1}(x) + \eta h_t(x)$$

where $\eta \in [0,1]$ represents a learning rate where the contribution of each tree can be controlled. This method of incremental approaches minimizes bias and variance, and offers features important to its interpretation.

5. Decision Tree Classifier (DT) Muraina, et al. [36]: The DT classifier is an easy-to-understand, simple baseline in the ensemble. It recursively breaks the space of features by minimizing the Gini impurity:

$$G = 1 - \sum_{k=1}^{K} p_k^2,$$

where p_k is the proportion of samples belonging to class k within a node. Though very simple, the decision tree identifies simple non-linear trends that intricate models may fail to identify.

3.5. Meta-Model: Logistic Regression

The meta-model combines the probability estimates of all base-learners to form a decision on its own Wang, et al. [37]. Let each base learner $h_m(x)$ produce a probability vector over C classes:

$$h_m(x) = [p_{m1}, p_{m2}, \dots, p_{mC}], \quad \sum_{c=1}^{C} p_{mc} = 1.$$

The joint probabilities of all M base learners give rise to the meta-feature vector:

$$z = [h_1(x), h_2(x), \dots, h_M(x) \in R^{M \times C}.$$

The meta-model, a regularized multinomial logistic regression, estimates the posterior probability of class c as:

$$P(y = c|z) = \frac{\exp(w_c z + b_c)}{\sum_{k=1}^{C} \exp(w_k z + b_k)},$$

where w_c and b_c are class-specific weight vectors and biases, respectively.

The L2 regularization (Ridge penalty) is applied to prevent overfitting; thus, the optimal optimization objective is given by:

$$\min_{w,b} - \sum_{i=1}^{N} \sum_{c=1}^{C} y_{ic} \log P(y_i, c|z_i) + \lambda \|w\|_2^2$$

where $\lambda = \frac{1}{C}$ is the inverse regularization.

The prediction of a new instance $x*$, after all is done, can be obtained as:

$$\hat{y} = \arg\max_{c \in C} P(y = c|z\,*) \, ,$$

where $z *= [h_1(x), h_2(x), \dots, h_M(x)]$.

3.6. Model Training and Validation

The training and validation process was oriented to make the model robust and avoid overfitting in order to enhance A-F student grades. The proposed hybrid framework is composed of a two-tier stacking ensemble where each base model is independently trained and contributes with its predicted probabilistic output to the meta-model, acting as input features.

All base learners, LightGBM, XGBoost, CatBoost, Gradient Boosting, and Decision Tree were trained on the same train dataset, scaled, and loaded with 80% of the data. All features were preprocessed using z-score normalization, and categorical values were encoded before model training.

For each student instance x, every base model $h_m(\cdot)$ generates a probability distribution over the 5 grade classes (A, B, C, D, F), as:

$$h_m(x_i) = [p_{i1}^{(m)}, p_{i2}^{(m)}, p_{i3}^{(m)}, p_{i4}^{(m)}, p_{i5}^{(m)}],$$

where $p_{ic}^{(m)}$ is the probability assigned by the *m-th* base learner for class c. These probabilities were concentrated to form a new meta-feature matrix $Z \in R^{N \times (M \times C)}$, which served as the input to the meta-model.

The base models were trained with hyperparameters listed in Table 3, tuned by Grid Search considering the weighted F1-score as the main metric.

Table 3.
Hyperparameter Settings for Base Models.

Model	Key Hyperparameters	Description
LightGBM	num_leaves = 31, learning_rate = 0.05, feature_fraction = 0.8, bagging_fraction = 0.8, min_data_in_leaf = 20, n_estimators = 300	Optimized for speed and generalization; controls complexity through shallow trees and subsampling.
XGBoost	max_depth = 6, learning_rate = 0.05, n_estimators = 400, subsample = 0.8, colsample_bytree = 0.8, reg_alpha = 0.1, reg_lambda = 1.0	Balances bias-variance trade-off with strong regularization.
CatBoost	depth = 8, iterations = 500, learning_rate = 0.03, l2_leaf_reg = 3, bagging_temperature = 0.2, border_count = 128	Handles categorical variables and reduces overfitting using ordered boosting.
Gradient Boosting	n_estimators = 300, learning_rate = 0.1, max_depth = 3, min_samples_split = 2, min_samples_leaf = 1	Uses incremental additive learning to minimize residual errors.
DT	criterion = ‘gini’, max_depth = 5, min_samples_leaf = 5, min_samples_split = 10	Provides interpretable, low-variance weak learner output for stacking diversity.

The meta-model, a multinomial logistic regression classifier, was trained on the probabilistic predictions of all 5 base learners. The regularization is performed by minimizing the negative log-likelihood. The meta-model was coded with LogisticRegression(multi_class = ‘multinomial’, solver = ‘lbfgs’, C = 1.0, penalty = ‘l2’, max_iter = 1000) to have convergence and stability.

The hybrid method was further validated with k=5 using Stratified K-Fold cross-validation in order to avoid model overfitting. This approach ensures a proportional representation of all classes during cross-validation, thus avoiding the bias introduced by class imbalance.

3.7. Model Explainability and SHAP Analysis

The SHapley Additive explanations (SHAP) framework was used to interpret the contributions of each feature to the prediction of an outcome in order to increase the interpretability of the model and offer transparent information about the feature contribution to the base models across the hybrid ensemble model [38] using their respective tree-based explainers (shap.TreeExplainer). SHAP provides an integrated, conceptually based set of methods to describe the output of any machine learning model by giving each feature its value in the end prediction. This method is founded on the cooperative game theory, where each feature is a so-called player in a coalition, and the SHAP value is calculated as the average of the marginal contribution of the feature in all the combinations of features.

For a given model $f(x)$ and an input instance $x = [x_1, x_2, \dots, x_3]$, the prediction can be decomposed as:

$$f(x) = \phi_0 + \sum_{i=1}^{d} \phi_i,$$

where ϕ_0 is the expected model output over the training data (the baseline value), and ϕ_i is the SHAP value representing the contribution of the feature i to the deviation of the prediction from the baseline.

The SHAP value for a given feature *i* is computed as:

$$\phi_i = \sum_{S \subseteq N\{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right],$$

where N is the collection of all the features, and S is the collection of features without i. The acronym $f_S(\cdot)$ is used to denote the model output of a model that is trained or tested on features in a subset S.

This formulation is such that attributing features in a fair and consistent fashion is achieved, satisfying the major properties of efficiency, symmetry, dummy, and additivity, and is consequently one of the strongest interpretability methods that can be applied to complex models.

3.8. Experimental Setup

All the experiments were performed on a high-performance workstation that was set up in a way that would provide good model training and assessment. The system introduced an Intel Core i5-12450H processor with 14 cores and 20 threads with a base clock of 2.4GHz, which is capable of performing adequate computing power to execute a parallel operation. As the workstation had 16GB of DDR5 RAM, it could easily manage large datasets and use memory-intensive ensemble computation. To support accelerated learning and visualization tasks, a system was installed with an NVIDIA GeForce GTX 1650 Laptop, an NVIDIA graphics card with the using NVIDIA GDDR6 VRAM 4 GB, and this graphics card can be used to parallelize algorithms, including LightGBM, XGBoost, and CatBoost. The experiments have been carried out in a Windows 11 Pro (64-bit) operating system, which is compatible with the Python-based machine learning libraries, and it offers a stable and optimized environment platform on which to run the hybrid framework effectively.

In order to control the reproducibility and performance consistency of their experiments, all the experimental procedures, model development, and training, as well as evaluation, were done in a controlled computational environment. Table 4 displays the main software components and version of the library employed in the experimental setup.

Table 4.
Computational software environment and library details

Component	Version/Description
Programming Language	Python 3.10
Distribution Environment	Anaconda Distribution (stable scientific environment)
Machine Learning Libraries	scikit-learn (v1.4) for model training and evaluation
Gradient Boosting Frameworks	LightGBM (v4.0), XGBoost (v2.0), CatBoost (v1.2) for boosting implementations
Data Processing Libraries	NumPy (v1.26) and Pandas (v2.2) for numerical computation and data manipulation
Visualization and Interpretability Tools	Matplotlib (v3.8), Seaborn (v0.13), and SHAP (v0.45) for visualization and model explainability

4. Results

The following section shows the findings of the study, which includes two key analysis elements: The Exploratory Data Analysis (EDA) and Experimental Outcomes. The EDA (Section 4.1) allows seeing the full picture of the dataset in terms of its statistical properties, distributions, and relationships between the features, so that one can fully comprehend the patterns, correlations, and potential biases before training the model. This phase, through the insights provided by visualization, indicates the underlying structure of student performance data, providing interpretative background to the predictive modeling phase.

Thereafter, Section 4.2 compares the results of the suggested hybrid ensemble model, as well as the results of the separate base learners. Reducing different quantitative indicators, such as accuracy, precision, recall, F1-score, and AUC, is evaluated to confirm the classification efficiency of the model. Further, interpretability studies with SHAP (SHapley Additive explanations) are provided to reveal the role of each feature in the model predictions. These analyses combined allow it to determine the strength, stability, and transparency of the proposed hybrid learning framework.

4.1. Exploratory Data Analysis (EDA)

4.1.1. Correlation Analysis of Student Performance Indicators

The correlation heatmap was generated in order to analyze the relationships between academic, behavioral, and demographic variables in the data. Since most of the correlations have a weak coefficient between -0.3 and +0.1, as shown in Figure 2, the features do not appear to have a strongly linearly related response. There were negligible positive relationships between Total_Score, Projects_Score, and Final_Score, implying that students who excel in projects and final tests achieve better overall grades. On the other hand, low negative correlations were observed between Grade and departmental features like Department_Engineering and Department_Mathematics, which suggests that grades are slightly different between disciplines. Characteristics like Attendance (%), Study_Hours_per_Week, and Sleep_Hours_per_Night, Study Hours per Week, and Sleep Hours per Night showed minimal effects on the academics. All in all, the heatmap indicates that the multicollinearity is low, which means that the features make a contribution to the predictive modeling process independently.

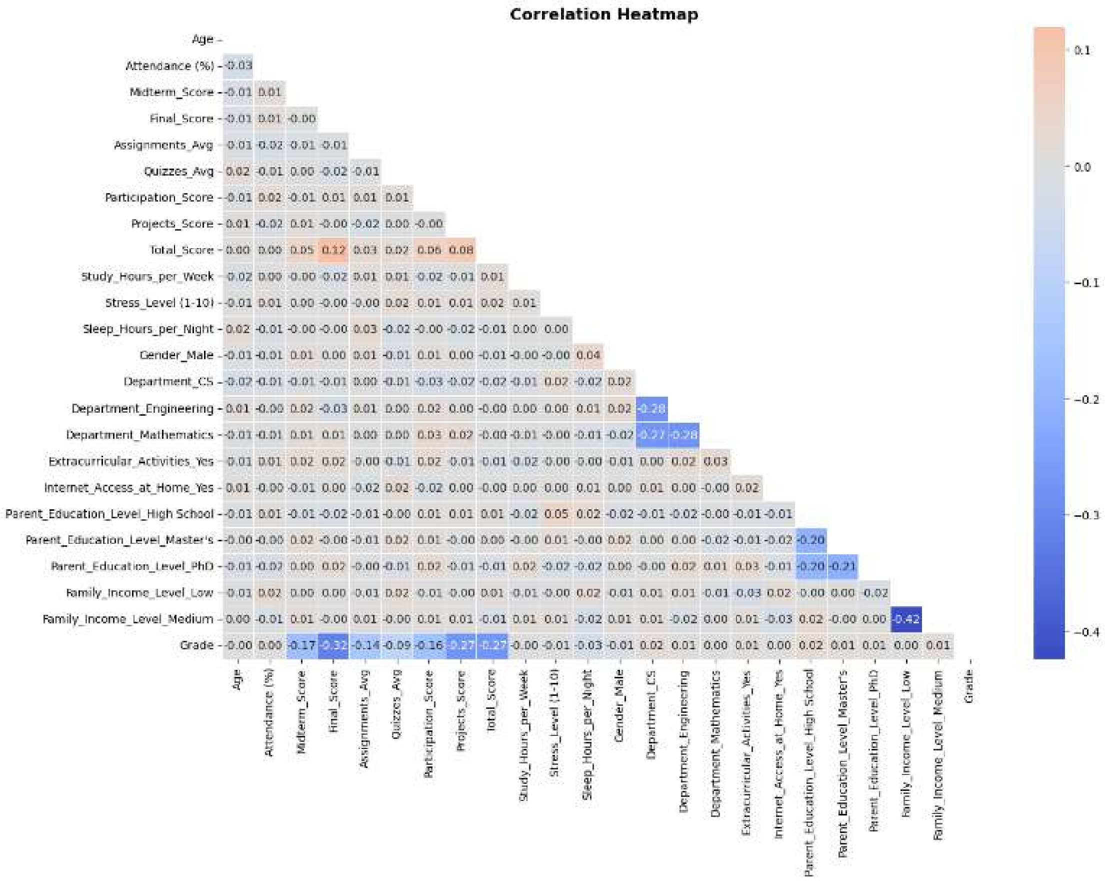

Figure 2.
Correlation heatmap depicting pairwise relationships among academic, behavioral, and demographic features, highlighting weak inter-feature associations and minimal multicollinearity.

4.1.2. Distribution and Variability of Input Features

Visualization of the distribution, spread, and possible outliers in all input features applied in the dataset was done using a boxplot display. As illustrated in Figure 3, most academic indices (Attendance (%), Midterm_Score, Final_Score, Assignments_Avg, Quizzes_Avg, and Projects_Score) show a distribution that is fairly symmetric around the median, combined with narrow to moderate interquartile ranges, which suggests uniform student performance across assessment dimensions. Participation_Score, Total_Score, etc have spread the highest, i.e. level of engagement and total scores vary a lot among students. The distribution of behavioral variables such as Study_Hours_per_Week, Stress_Level (1–10), and Sleep_Hours_per_Night is more unbalanced to reflect the diversity in study habits and general well-being. Binary and categorical features (e.g., Gender_Male, Department_CS, Extracurricular_Activities_Yes) present dense distributions as the values they can take are limited. A small number of mild outliers, especially in the scores for stress and participation, were kept as part of preserve the integrity of the data. In summary, the data boxplot does provide useful information on feature dispersion and existing heterogeneity among students.

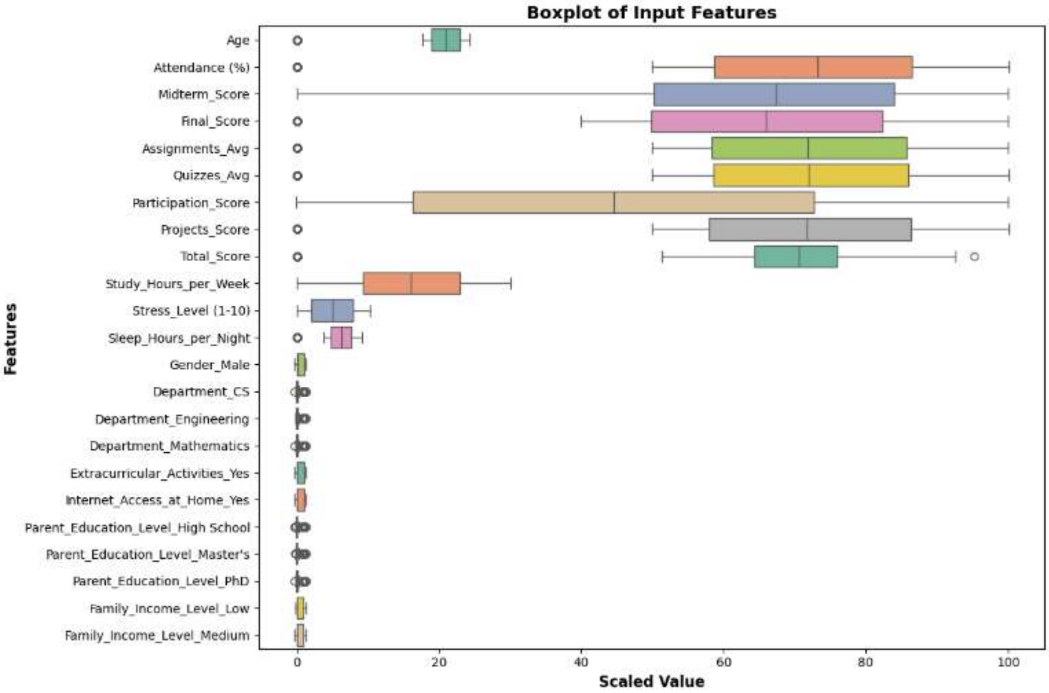

Figure 3.
Boxplot of input features.

4.1.3. Bivariate Relationships among Key Academic Indicators

Pairwise dependencies and joint distributions between major academic features were then investigated through a sequence of 2D Kernel Density Estimation (2D KDE) plots. Figure 4 shows these contour plots that illustrate the distribution of points as a function of the pair of features so we can understand how different types of assessment components correlate, or anti-correlate, with each other. A high positive relationship was particularly found between Total_Score and both Projects_Score and Final_Score, where denser contour trends along the diagonal indicate that higher project scores and exam scores lead to better overall academic performance. Likewise, Assignments_Avg and Quizzes_Avg exhibit a symmetric joint distribution, indicating that student performance on continuous assessments is consistent on average.

Conversely, relationships of Participation_Score and Midterm_Score seem to be more scattered, which suggests that the classroom participation and midterm results may not be directly proportional to other performance measurements. The density contours in each plot are rather similar and symmetric, indicating a small amount of heteroscedasticity as well as that scores are uniformly distributed, which also supports good data quality. Taken together, 2D KDEs offer a fine-grained view of the inner dynamics between features, capturing nonlinear relationships and emphasizing the strongest relations with academic performance.

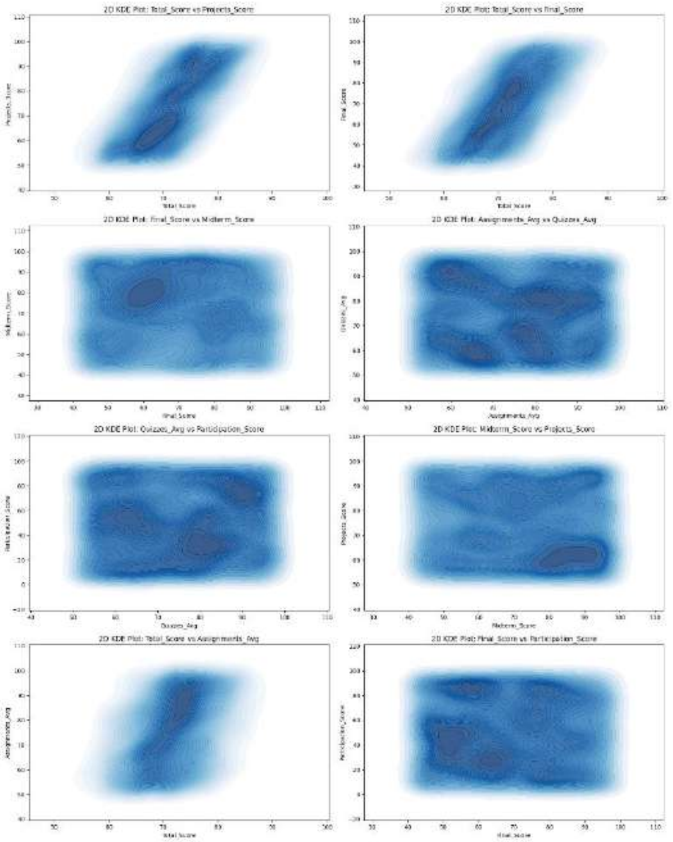

Figure 4.
2D KDE plots illustrating joint density distributions between key academic performance variables.

4.1.4. Distribution of Students by Department

To gain insight into the demographics of participants in the dataset, a department distribution analysis was performed. As depicted in Figure 5, the database includes a collection of 5000 student records that are evenly divided into four academic departments: Engineering, Business, Computer Science (CS), and Math. Engineering is the highest in terms of percentage, as it covers 25.5% (1274 students, closely followed by Business at 25.3%, CS at 24.8% and Mathematics at 24.5%.

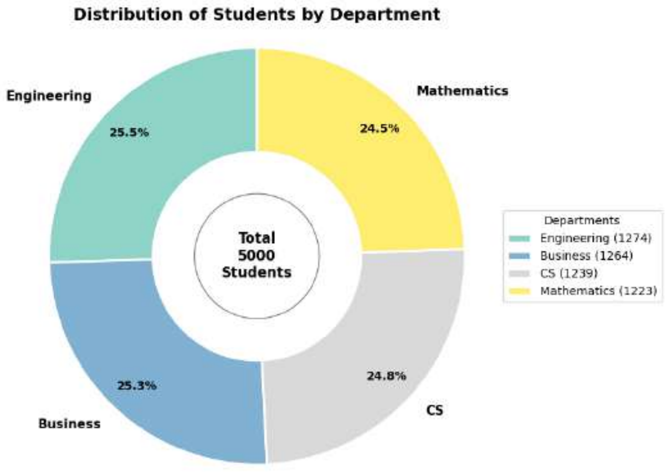

Figure 5.
Distribution of students across academic departments.

This emphasizes balanced departmental coverage as well, contributing to the generalizability of the resulting model across a wide range of academic disciplines. The even spread also enables alleviating the possible problem of class imbalance when training the models, so that a fair conclusion can be drawn with respect to whether PLO attainment is comparably evaluated by hybrid/ensemble model structures in varied fields.

4.2. Experimental Outcomes

4.2.1. Classification Performance Across Grade Categories

The performance of the hybrid ensemble model was satisfactory in all grade ranges (A–F). as shown in Table 5. The model attained high performance values for both accuracy, precision, recall, and F1-score measures, showing good discriminative capacity as well as balanced classification behavior.

Classes D and C showed the highest level of accuracy, 0.966 and 0.986, respectively, with a high F1-score (0.97, 0.96). This implies the model is able to identify well the underlying trends pertinent to mid-range academic performance. Perfect precision (1.00, perfect specificity) was attained for Class B, whereas recall (0.88) for this class was slightly lower, indicating a conservative but very precise classification habit. Likewise, class F achieved high precision (0.99) and recall (0.91), which noted the model's aptitude to correctly detect failing students. Class A had low to moderate but consistent scores (0.83 each) on all the metrics, suggesting a reliable capacity of predicting top-performing users, although they were less represented in the dataset.

Table 5.
Per-class classification performance metrics for the hybrid ensemble model.

Class	Accuracy	Precision	Recall	F1-Score
A	0.833	0.83	0.83	0.83
B	0.879	1.00	0.88	0.94
C	0.986	0.94	0.99	0.96
D	0.966	0.97	0.97	0.97
F	0.908	0.99	0.91	0.95

Overall, the radar visualization in Figure 6 displays clusters of closely packed polygons across all three metrics, indicating that the hybrid stacking model generalizes well through different performance groups and does not have a serious bias towards any specific grade group.

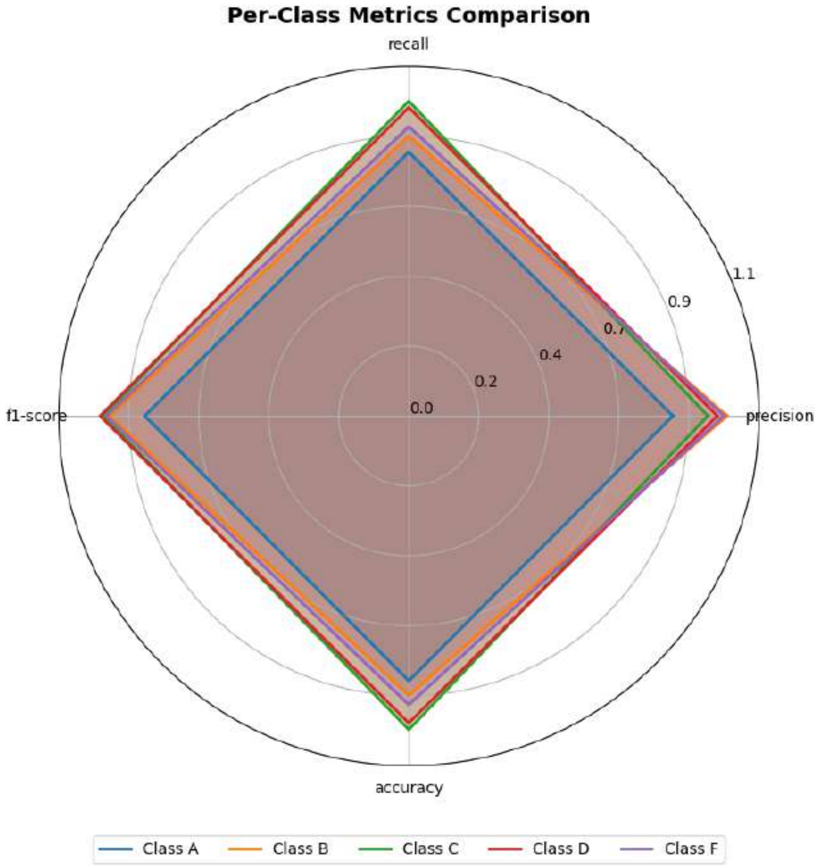

Figure 6.
Radar chart comparing per-class performance metrics (accuracy, precision, recall, and F1-score) for each grade category (A–F).

4.2.2. Error Analysis of the Hybrid Model

The predictive accuracy of the proposed hybrid ensemble classifiers and the error distribution have been discussed in terms of the confusion matrix, which is shown in Figure 7. The confusion matrix lists the correctly and incorrectly classified instances by grade category (A-F), giving us a more detailed picture of the model's performance across all class boundaries.

The results indicate a significant diagonal dominance level, demonstrating high classification accuracy and low intra-group confusability. The model accurately identified most students at the C level, with 796 accurate predictions and D level (n =595), thus the model was at least useful in identifying average performing middle-of-the-road students. A few misclassifications were also found between sequential classes, including 26 class B to C and 20 class C to D; this overlap indicates that some performance boundaries are less clearly defined in contiguous grade levels, and thus predictive confidence can be low.

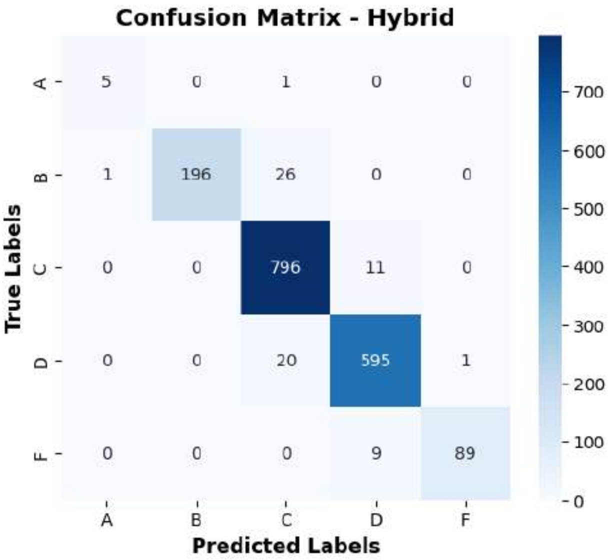

Figure 7.
Confusion matrix of the hybrid stacked ensemble model.

Especially for the F category, which contains most of the students of interest (89 right predictions), it was highly accurate to identify students at risk. Also, very few cross-category errors were observed for A and B, which confirms that the model can generalize over all grades. In general, the confusion matrix confirms the equilibrium and stable prediction of the hybrid stacked ensemble model, indicating its effectiveness in academic outcome prediction.

4.2.3. Feature Importance Analysis

To explain how the hybrid ensemble model made decisions, feature importance analysis was conducted by averaging the importance scores over all base learners. Figure 8 shows the top contributors of the most important predictors to classify students' grades based on each sub-activity.

Of all the input variables, Total_Score was found to be the most dominant feature, indicating that it has a key importance in guiding the sum total academic performance. The next two stronger contributors are Projects_Score and Final_Score, which it is also evident that the performance on projects and final exam plays key roles in the overall score. Midterm_Score, Assignments_Avg, and Quizzes_Avg had low importance values, which are secondary and helpful to predict the grades. Predictors such as Participation_Score and Extracurricular_Activities_Yes were minor, indicating that the personal qualities about engagement in class or extracurricular activities might have a less decisive role than academic assessment information.

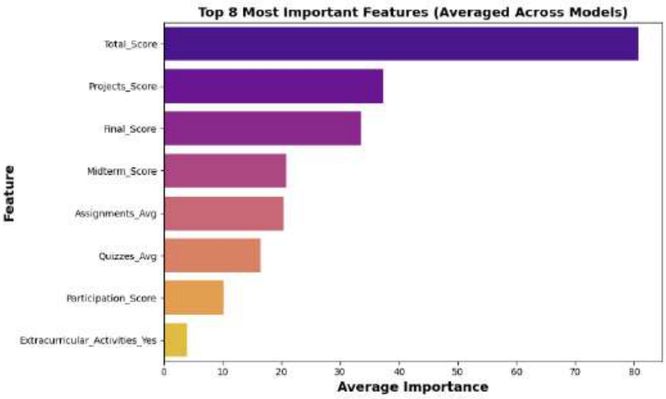

Figure 8.
Top 8 most important features averaged across base models.

The highly concentrated importance of continuous assessment variables (Total_Score, Projects_Score, and Final_Score) justifies the educational insight that persistent academic evaluation components could be taken as the most significant indicators of students' success. These results support the interpretability and pedagogical adherence of the predictive process performed by the hybrid ensemble model.

4.2.4. Distribution Consistency Between Actual and Predicted Grades

In order to assess the general consistency and congruity between the model predictive outcomes and the ground-truth labels, a Kernel Density Estimation (KDE) comparison was carried out. The KDE curves indicated in Figure 9 are the probability density curves of the real grades class and the predicted grades class in the case of numerical encoding. The fact that the two curves are very close to each other implies that the hybrid ensemble model is effective in the manner that it implies the underlying class distribution, but does not result in the prediction of grades with much bias and/or distortion.

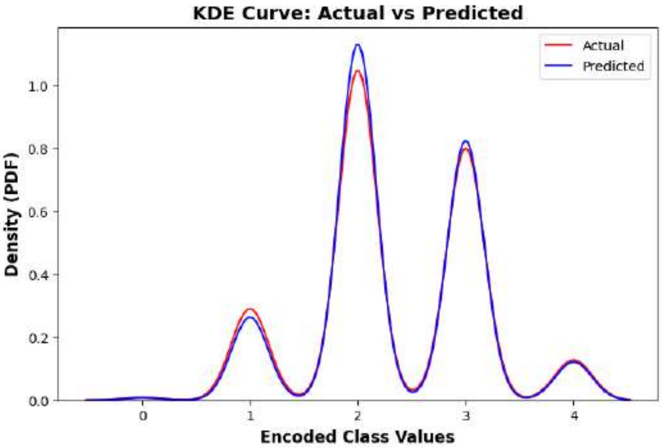

Figure 9.
KDE curve illustrating the probability density functions of actual and predicted grade class values.

Similarities are found between both curves in their respective peaks at corresponding class values, especially the ranges in the mid-range. This confirms the fact that the probability structure as predicted is too close to actual performance trends in the world. Minor density amplitude deviation is found on the lower and upper boundaries of grades, which could be caused by a slight imbalance of classes or an overlapping performance characteristic of high and poor-performing students. However, these almost-tight coincidences between the distribution of the red (true) and the blue (predicted) distributions are an indication that the hybrid model shows high fidelity to the distribution, meaning its predictions will be statistically consistent with how the actual data behaves.

4.2.5. ROC-AUC Evaluation for Multiclass Classification

In further evaluation of the evidence of the discriminative capability of the proposed hybrid ensemble model, the Receiver Operating Characteristic (ROC) analysis has been conducted on each of the five grade categories (A-F) separately. Figure 10 represents the ROC curves to visualize the trade-off between the true positive rate (sensitivity) and the false positive rate of the various classification threshold levels.

The hybrid model recorded close-perfect Area Under the Curve (AUC) scores in all classes of A (0.998), B (0.998), C (0.997), D (0.998), and F (0.999), which is the best in student grade types. The micro-averaged AUC was 0.998, which validates the high performance across the whole dataset. The fact that the AUC values remain consistently high shows that the model is great in distinguishing between various academic performance levels, and there is a low degree of overlap between the probabilities of the various classes as predicted.

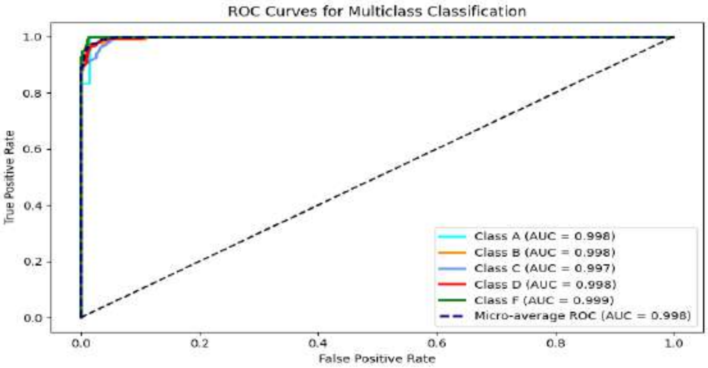

Figure 10.
ROC curves for multiclass grade prediction.

The large tissue of all ROC curves toward the upper-left side of the plot indicates a high degree of sensitivity and specificity, which bears out the ability of the hybrid ensemble to support generalization and its ability to put up with non-proportional multiclass educational data.

4.3. SHAP-Based Feature Interpretability for the Hybrid Model

SHapley Additive exPlanations (SHAP) analysis was performed on all the base learners and meta-model to increase the interpretability and to disclose the internal processes of decision-making of the hybrid ensemble. SHAP gives a game-theoretic, unified structure to measure the contribution of each feature to individual predictions, and it has global and local interpretability.

Figure 11 shows the importance of the aggregated SHAP feature of the meta-model (Logistic Regression) computed based on the original names of the features across all the base learners. The mean absolute SHAP values of all grade classes are presented in the stacked bar format, which illustrates the greatest influence of what features have on the final ensemble prediction. The qualities of Total, Projects, Final score all show substantial positive effects and this indicates that they play a dominant role in defining the level of academic performance. Moderate contributions are also found in the Assignments_Avg and Midterm_Score and other variables, including Participation_Score and Extracurricular_Activities_Yes, have smaller SHAP values, which is in line with their role as secondary predictors of the feature influence, which was previously explored in the feature importance analysis.

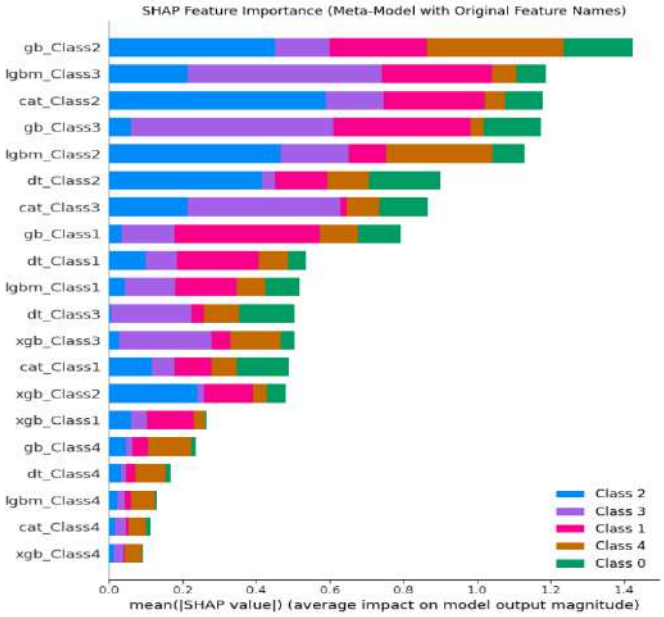

Figure 11.
SHAP feature importance plot.

In addition, the color segregation reflects SHAP contributions per the individual classes (grade A-F), indicating that the same feature may affect different outputs differently in grade sharing distributions. This highlights the subtle nature of the hybrid model by balancing features differently by academic level to make predictions that are more context sensitive.

The independent interpretive behavior of each of these 5 base models, LightGBM, XGBoost, CatBoost, Gradient Boosting, and Decision Tree, was also analyzed by creating an individual SHAP summary plot. Total_Score, Final_Score kept on topping the list of most predictive ones across all learners, but with some marginal differences in the contribution of the secondary features (e.g., quizzes and assignments), the complementary character of the ensemble stacking was revealed. Taken together, these interpretability findings support the transparency and explainability of the hybrid model and allow educators and analysts to follow the reasoning behind a certain prediction to particular academic aspects.

4.4. SHAP-Based Interpretability for Individual Base Learners

In addition to the meta-model analysis, visualizations using SHAP (SHapley Additive exPlanations) technique were created for each base learner — LightGBM, XGBoost, CatBoost, Gradient Boosting, and Decision Tree in order to explore their individual feature contribution patterns. These beeswarm plots (Figures 12–16) present the average and instance-level effect of each input variable on model output across all prediction categories.

As illustrated in Figure 12, it is evident from the LightGBM explainer that socioeconomic and demographic indicators such as Family_Income_Level and Parent_Education_Level have the highest SHAP values, suggesting a moderate but consistent impact to affect the prediction of student performance. As LightGBM excels in learning structural relationships, because it centers around categorical features such as Department and Internet_Access_at_Home_Yes, one would naturally think that its weight of contribution lies in modeling the contextual diversity more than the academic performance.

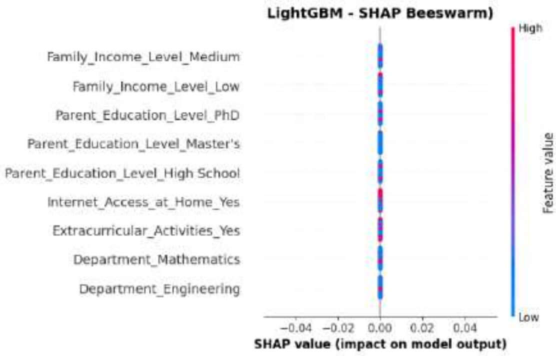

Figure 12.
SHAP beeswarm plot for the LightGBM classifier.

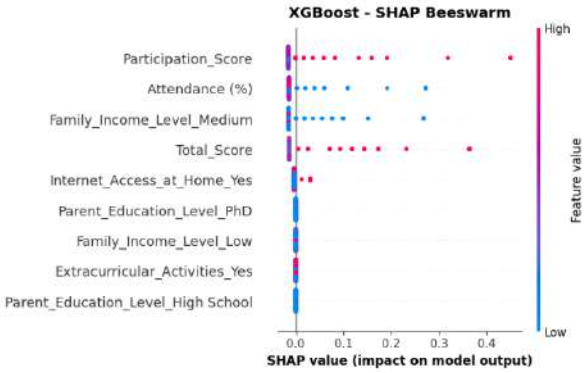

Figure 13.
SHAP beeswarm plot for the XGBoost classifier.

On the other hand, XGBoost has a stronger relationship with performance-related factors, as identified in Figure 13, such as Participation_Score, Attendance (%), and Total_Score. The very high SHAP values of these features suggest that the engagement and continuous assessment components are strong demarcators of predicted grades. This is consistent with the gradient-regularized learning strategy of XGBoost, which is effective at discovering subtle feature–target dependencies that are present in structured tabular data.

The Gradient Boosting model SHAP beeswarm (Figure 14) shows that Total_Score is the most important feature by far, with all other predictors taking a backseat. This indicates that the model is likely to strongly lean on overall academic performance measures. But this is also the point that highlights its fairly low diversity of interpretations compared to tree boosting algorithms (like CatBoost or LightGBM).

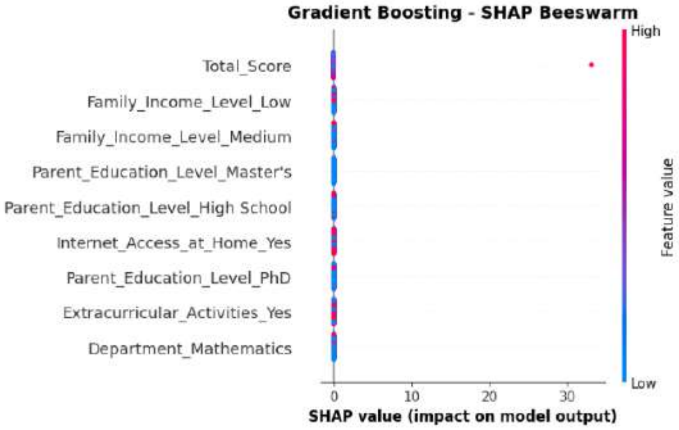

Figure 14.
SHAP beeswarm plot for the Gradient Boosting classifier.

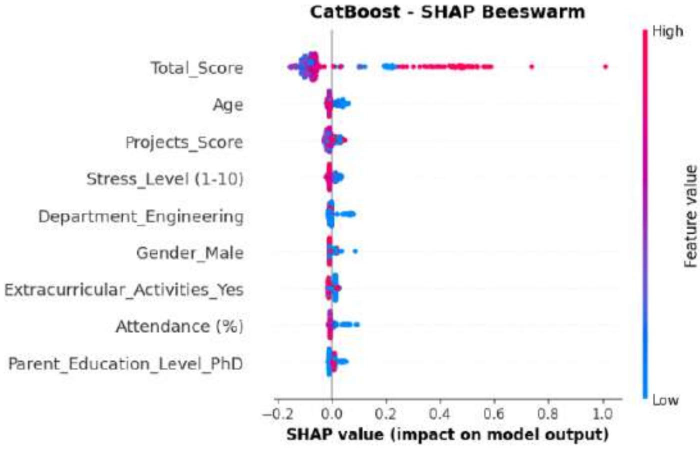

Figure 15.
SHAP beeswarm plot for the CatBoost classifier.

As shown in Figure 15, CatBoost includes a wider distribution of feature importance, with Total_Score, Projects_Score, and Stress_Level (1-10) being more influential. The model is also sensitive to Age and Gender_Male, indicating that it has the ability to capture complex individual-level patterns. CatBoost may have advantages in the interpretive perspective, and is suited specifically to mixed-type educational data.

The SHAP output of the Decision Tree model (Figure 16) shows a kind of light pictorial representation with the main focus on Total_Score and Study_Hours_per_Week as drivers. Simple as the Decision Tree may be, it helps the hybrid ensemble deliver interpretable low-complexity decision boundaries that increase model diversity.

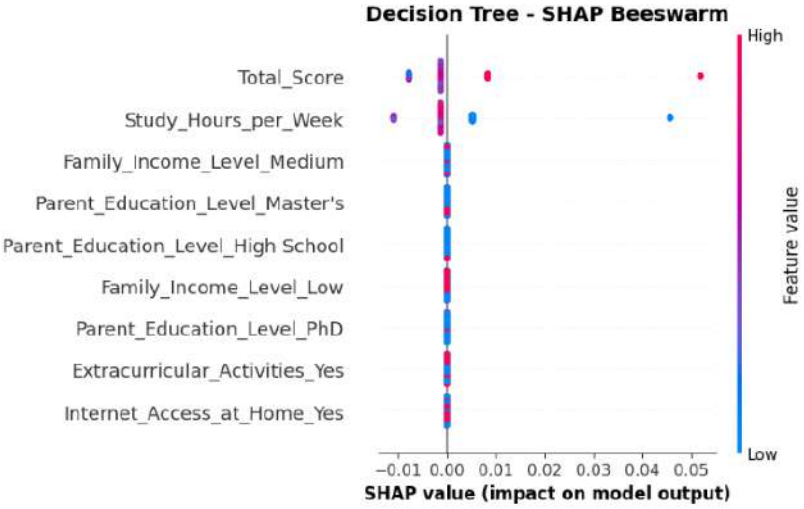

Figure 16.
SHAP beeswarm plot for the Decision Tree classifier.

Taken together, these SHAP-based analyses verify that each base learner is emphasizing different predictive aspects/forms – from demographic factors (LightGBM) to behavioral engagement (XGBoost, CatBoost), yet their ensemble within the hybrid stacking architecture maintains a compromise between interpretability and robustness of prediction. The complementary learning behavior between models increases the ability of the hybrid model to generalize and fit with real-world educational testing structures.

4.5. Comparative Analysis and Discussion

To measure the efficiency of the hybrid approach, it was compared with a number of benchmark models available from past works, as given in Table 6. Luo, et al. [39] used a Random Forest Classifier to predict students' performance with an accuracy of 85.3%. However, the performance of their model was based mostly on decision-tree bagging, which has issues with high-dimensional or correlated features. In comparison, our proposed hybrid model received an accuracy rate of 96%, which represents a much higher increase of 10.7%, due to the joint combination of multiple boosting learners promoting feature discrimination and generalization.

Hafdi and El Kafhali [40] used an LSTM network, with consideration of time dependencies in sequential data, achieving a good accuracy which is 94%. Nonetheless, deep learning techniques, such as LSTM, can require a significant amount of data and computation. In addition, our ensemble model achieved a higher accuracy than this baseline, even though we reduced the training complexity and gained more interpretable results based on SHAP.

Table 6.
Comparative analysis of the proposed hybrid model with existing studies.

Reference	Method	Accuracy (%)
Luo, et al. [39]	Random Forest Classifier	85.3
Hafdi and El Kafhali [40]	LSTM	94
Iatrellis, et al. [41]	K-Means Cluster	83.2
Perez and Perez [42]	Naïve Bayes Classifier	83.77
Proposed	Hybrid ML Classifier	96

Iatrellis, et al. [41] utilized a K-Means clustering approach that accurately classified students based on learning styles (83.2%). Although useful for unsupervised exploration, this method is deficient in predictive accuracy. The hybrid model exhibited a boost in accuracy by 12.8%, highlighting the superior performance of supervised ensemble learning in learning nonlinear feature relations as well as complex class borders.

Perez and Perez [42] used Naïve Bayes classification, with an accuracy of 83.77% on predicting student performance. Despite its efficiency, the assumption of independence among the features restricts its generalizing ability. The proposed hybrid model performed better than this approach by 12.23% indicating robustness in handling correlated predictors and exploiting the strength of multi-models to improve accuracy performance.

On the whole, the proposed hybrid machine learning classifier was able to record an accuracy of 96%, and was robustly better than reference models by 2% at least up to more than 12%. Such advances reveal the predictive robustness, overfitting control, and interpretability through inbuilt SHAP-based analysis of the ensemble, thus making it a more explainable and teacher-friendly predictive framework for analyzing academic performance.

5. Conclusion

The study showed that a stacked model, which is a set of tree-based gradient boosters with a logistic-regression meta-model, is capable of making accurate and well-calibrated predictions of multi-class PLO achievement and remains interpretable. In the stratified testing, the hybrid method has outperformed the single model by more than an order of magnitude, and has superior discrimination to the common base, especially with the mid-range and failing (important in early intervention) groups. SHAP results offer practical levels of transparency as they assign predictions to Total, Project, and Final scores, and other supporting variables are engagement and contextual values, making the model. behavior is very important to pedagogical intuition. These results imply that explicable hybrid AI could aid in data-driven curriculum development, individualized academic course advice, and outcome assessment quality in outcome-based instruction. Future efforts need to confirm the framework using bigger dataset sizes, and stronger temporal dynamics (longitudinal, course-level) to generate more generalizable results and to increase feature coverage (e.g., affective/cognitive measures) in order to increase applicability and implementation.

References

[1] A. Bettahi, F. Z. Belouadha, and H. Harroud, "AI and EDM: Revolutionizing Global Education and Crafting a Personalized Digitally Advanced Learning," in *Internationalization of Higher Education and Digital Transformation: Insights from Morocco and Beyond*, A. Adoui Ed. Cham: Springer Nature Switzerland, 2024, pp. 243-258.

[2] S. Kaspi and S. Venkatraman, "Data-driven decision-making (DDDM) for higher education assessments: A case study," *Systems*, vol. 11, no. 6, p. 306, 2023. https://doi.org/10.3390/systems11060306

[3] J. Jovanović, M. Saqr, S. Joksimović, and D. Gašević, "Students matter the most in learning analytics: The effects of internal and instructional conditions in predicting academic success," *Computers & Education*, vol. 172, p. 104251, 2021. https://doi.org/10.1016/j.compedu.2021.104251

[4] S. Li, *Learning analytics enhanced online learning support*. London, UK: Routledge, 2023.

[5] J. Mahisha, A. B, M. Edwin, G. P, and K. Ravichandran, "Identifying learning difficulties at an early stage in education with the help of artificial intelligence models and predictive analytics," *International Research Journal of Multidisciplinary Scope*, vol. 05, pp. 1455-1461, 2024. https://doi.org/10.47857/irjms.2024.v05i04.01821

[6] A. Parhizkar, G. Tejeddin, and T. Khatibi, "Student performance prediction using datamining classification algorithms: Evaluating generalizability of models from geographical aspect," *Education and Information Technologies*, vol. 28, no. 11, pp. 14167-14185, 2023. https://doi.org/10.1007/s10639-022-11560-0

[7] M. Domladovac, "Comparison of neural network with gradient boosted trees, random forest, logistic regression and SVM in predicting student achievement," presented at the 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO), 2021.

[8] C. Wang, L. Chang, and T. Liu, *Predicting student performance in online learning using a highly efficient gradient boosting decision tree*. Cham: Springer International Publishing, 2022.

[9] 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, no. 1, p. 26879, 2025. https://doi.org/10.1038/s41598-025-12353-4

[10] A. Joshi, P. Saggar, R. Jain, M. Sharma, D. Gupta, and A. Khanna, "CatBoost — an ensemble machine learning model for prediction and classification of student academic performance," *Advances in Data Science and Adaptive Analysis*, vol. 13, no. 03n04, p. 2141002, 2021. https://doi.org/10.1142/S2424922X21410023

[11] N. A. Butt, Z. Mahmood, K. Shakeel, S. Alfarhood, M. Safran, and I. Ashraf, "Performance prediction of students in higher education using multi-model ensemble approach," *IEEE Access*, vol. 11, pp. 136091-136108, 2023. https://doi.org/10.1109/ACCESS.2023.3336987

[12] K. M. Sujon *et al.*, "The effects of imbalanced datasets on machine learning algorithms in predicting student performance," *International Journal on Informatics Visualization*, vol. 8, no. 3-2, pp. 1599-1605, 2024.

[13] M. Kumar, V. Bhardwaj, D. Thakral, A. Rashid, and M. T. B. Othman, "Ensemble learning based model for student's academic performance prediction using algorithms," *Ingenierie des Systemes d'Information*, vol. 29, no. 5, pp. 1925–1935, 2024. https://doi.org/10.18280/isi.290524

[14] R. Dey and R. Mathur, *Ensemble learning method using stacking with base learner, a comparison*. Singapore: Springer Nature Singapore, 2023.

[15] W. C. Choi, C. T. Lam, and A. J. Mendes, "Analyzing the interpretability of machine learning prediction on student performance using shapley additive exPlanations," presented at the 2024 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE), 2024.

[16] H. Sahlaoui, E. A. A. Alaoui, A. Nayyar, S. Agoujil, and M. M. Jaber, "Predicting and interpreting student performance using ensemble models and shapley additive explanations," *IEEE Access*, vol. 9, pp. 152688-152703, 2021. https://doi.org/10.1109/ACCESS.2021.3124270

[17] A. Kowalska, R. Banasiak, J. Stańdo, M. Wróbel-Lachowska, A. Kozłowska, and A. Romanowski, "Study on using machine learning-driven classification for analysis of the disparities between categorized learning outcomes," *Electronics*, vol. 11, no. 22, p. 3652, 2022. https://doi.org/10.3390/electronics11223652

[18] M. Goyal, C. Gupta, and V. Gupta, "A meta-analysis approach to measure the impact of project-based learning outcome with program attainment on student learning using fuzzy inference systems," *Heliyon*, vol. 8, no. 8, p. e10248, 2022. https://doi.org/10.1016/j.heliyon.2022.e10248

[19] N. Zaki, S. Turaev, K. Shuaib, A. Krishnan, and E. Mohamed, "Automating the mapping of course learning outcomes to program learning outcomes using natural language processing for accurate educational program evaluation," *Education and Information Technologies*, vol. 28, no. 12, pp. 16723-16742, 2023. https://doi.org/10.1007/s10639-023-11877-4

[20] H. Pallathadka, A. Wenda, E. Ramirez-Asís, M. Asís-López, J. Flores-Albornoz, and K. Phasinam, "Classification and prediction of student performance data using various machine learning algorithms," *Materials Today: Proceedings*, vol. 80, pp. 3782-3785, 2023. https://doi.org/10.1016/j.matpr.2021.07.382

[21] M. Adnan *et al.*, "Predicting at-risk students at different percentages of course length for early intervention using machine learning modelsa," *IEEE Access*, vol. 9, pp. 7519-7539, 2021. https://doi.org/10.1109/ACCESS.2021.3049446

[22] M. Yağcı, "Educational data mining: Prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, p. 11, 2022. https://doi.org/10.1186/s40561-022-00192-z

[23] P. Nayak, S. Vaheed, S. Gupta, and N. Mohan, "Predicting students' academic performance by mining the educational data through machine learning-based classification model," *Education and Information Technologies*, vol. 28, no. 11, pp. 14611–14637, 2023. https://doi.org/10.1007/s10639-023-11706-8

[24] R. Lamb, K. Neumann, and K. A. Linder, "Real-time prediction of science student learning outcomes using machine learning classification of hemodynamics during virtual reality and online learning sessions," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100078, 2022. https://doi.org/10.1016/j.caeai.2022.100078

[25] J. Kabathova and M. Drlik, "Towards predicting student's dropout in university courses using different machine learning techniques," *Applied Sciences*, vol. 11, no. 7, p. 3130, 2021. https://doi.org/10.3390/app11073130

[26] F. Giannakas, C. Troussas, I. Voyiatzis, and C. Sgouropoulou, "A deep learning classification framework for early prediction of team-based academic performance," *Applied Soft Computing*, vol. 106, p. 107355, 2021. https://doi.org/10.1016/j.asoc.2021.107355

[27] S. Hussain and M. Q. Khan, "Student-performulator: Predicting students' academic performance at secondary and intermediate level using machine learning," *Annals of Data Science*, vol. 10, no. 3, pp. 637-655, 2023. https://doi.org/10.1007/s40745-021-00341-0

[28] F. Qiu *et al.*, "Predicting students' performance in e-learning using learning process and behaviour data," *Scientific Reports*, vol. 12, no. 1, p. 453, 2022. https://doi.org/10.1038/s41598-021-03867-8

[29] T. Al-Shehari and R. A. Alsowail, "An insider data leakage detection using one-hot encoding, synthetic minority oversampling and machine learning techniques," *Entropy*, vol. 23, no. 10, p. 1258, 2021. https://doi.org/10.3390/e23101258

[30] G. Sailasya and G. Kumari, "Analyzing the performance of stroke prediction using ML classification algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 12, 2021. https://doi.org/10.14569/IJACSA.2021.0120662

[31] K. M. Al-Gethami, M. T. Al-Akhras, and M. Alawairdhi, "Empirical evaluation of noise influence on supervised machine learning algorithms using intrusion detection datasets," *Security and Communication Networks*, vol. 2021, no. 1, p. 8836057, 2021.

[32] M. M. Ahsan, M. A. P. Mahmud, P. K. Saha, K. D. Gupta, and Z. Siddique, "Effect of data scaling methods on machine learning algorithms and model performance," *Technologies*, vol. 9, no. 3, p. 52, 2021. https://doi.org/10.3390/technologies9030052

[33] G. Li, J. Cui, H. Fu, and Y. Sun, "Light GBM and GA based algorithm for predicting and improving students' performance in higher education in the context of big data," presented at the 2024 7th International Conference on Education, Network and Information Technology (ICENIT), 2024.

[34] A. Joshi, P. Saggar, R. Jain, M. Sharma, D. Gupta, and A. Khanna, "CatBoost- an ensemble machine learning model for prediction and classification of student academic performance," *Advances in Data Science and Adaptive Analysis*, vol. 13, 2021. https://doi.org/10.1142/S2424922X21410023

[35] E. Mashagba, F. Al-Saqqar, and A. Al-Shatnawi, "Using gradient boosting algorithms in predicting student academic performance," presented at the 2023 International Conference on Business Analytics for Technology and Security (ICBATS), 2023.

[36] I. Muraina, E. Aiyegbusi, and S. Abam, "Decision tree algorithm use in predicting students' academic performance in advanced programming course," *International Journal of Higher Education Pedagogies*, vol. 3, pp. 13-23, 2023. https://doi.org10.33422/ijhep.v3i4.274

[37] W. Wang, J. Zhang, and B. Hu, *Meta-learning with logistic regression for multi-classification*. Singapore: Springer Singapore, 2022.

[38] Y. Guan, F. Wang, and S. Song, "Interpretable machine learning for academic performance prediction: A SHAP-based analysis of key influencing factors," *Innovations in Education and Teaching International*, pp. 1-20, 2025. https://doi.org/10.1080/14703297.2025.2532050

[39] Y. Luo, X. Han, and C. Zhang, "Prediction of learning outcomes with a machine learning algorithm based on online learning behavior data in blended courses," *Asia Pacific Education Review*, vol. 25, no. 2, pp. 267-285, 2024. https://doi.org/10.1007/s12564-022-09749-6

[40] Z. S. Hafdi and S. El Kafhali, "A comparative evaluation of machine learning methods for predicting student outcomes in coding courses," *AppliedMath*, vol. 5, no. 2, p. 75, 2025. https://doi.org/10.3390/appliedmath5020075

[41] O. Iatrellis, I. K. Savvas, P. Fitsilis, and V. C. Gerogiannis, "A two-phase machine learning approach for predicting student outcomes," *Education and Information Technologies*, vol. 26, no. 1, pp. 69-88, 2021. https://doi.org/10.1007/s10639-020-10260-x

[42] J. G. Perez and E. S. Perez, "Predicting student program completion using Naïve Bayes classification algorithm," *International Journal of Modern Education and Computer Science*, vol. 12, no. 3, p. 57, 2021.