MSTS Methods in Science and Technology Studies

futuristech PT. Teknologi Futuristik Indonesia

Date of publication March 15, 2026, date of current version March 15, 2026.
Digital Object Identifier 10.64539/msts.v2i1.2026.441

Article

Interpretable Academic Outcome Prediction Using Explainable Boosting Machines

Godfrey Perfectson Oise[1,*], Felix Oshiorenoya Uloko[2], Kevin Chinedu Pius[1], Enovwo Eferoba–Idio[1], Michael Uyiosa Edobor[3], Evans Mintah[4], Osahon Ukpebor[5], Oludare Sokoya[6], Tejiri Jessa[4]

[1] Department of Computing, Wellspring University, Benin City, Edo State, Nigeria; e-mail: godfrey.oise@wellspringuniversity.edu.ng (G. P. Oise), kevin.pius@wellspringuiversity.edu.ng (K. C. Pius), eferoba-idio.enovwo@wellspringuniversity.edu.ng (E. Eferoba-Idio).
[2] Department Computer Science, Veritas University, Abuja, Bwari 901101, Federal Capital Territory, Nigeria; e-mail: ulokof@veritas.edu.ng (F. O. Uloko).
[3] Department of Computer Science, University of Benin, Edo State, Nigeria; u.edobor@yahoo.com (M. U. Edobor).
[4] College of Business, Westcliff University, Irvine, CA 92614, United States; e-mail: mintah1@gmail.com (E. Mintah), tejirijessa@gmail.com (T. Jessa).
[5] Department of Computer and Information Science, University of the Cumberlands, Williamsburg, KY 40769, United States; e-mail: iukpebor@gmail.com (O. Ukpebor).
[6] College of Business, Engineering & Technology, National University, San Diego, CA 92123, United States; e-mail: daresokoya@gmail.com (O. Sokoya).
* Correspondence Author

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Abstract: Predictive analytics has become an important component of learning analytics in higher education, enabling institutions to identify academic risks and support student success through data-driven decision making. However, many existing academic outcome prediction models rely on complex black-box machine learning techniques that provide high predictive performance but limited transparency and interpretability. The lack of explainability restricts the practical adoption of such models in educational environments where accountability, trust, and ethical decision-making are essential. This study proposes an interpretable machine learning framework for multi-class academic outcome prediction using the Explainable Boosting Machine (EBM), a glass-box model that combines the predictive power of ensemble boosting with the transparency of generalized additive models. The proposed framework was evaluated using a publicly available Student Performance and Learning Behavior dataset consisting of 6,519 student records containing academic, behavioral, and demographic attributes. Academic outcomes were formulated as a four-class classification task: Distinction, Pass, Fail, and Withdrawn. Model performance was assessed using multiple evaluation metrics including accuracy, precision, recall, F1-score, and ROC–AUC analysis. Experimental results demonstrate that the proposed EBM model achieves an accuracy of 88% and an overall ROC–AUC score of 0.91, indicating strong predictive capability across outcome categories. Furthermore, the model provides native interpretability through feature contribution functions and SHAP-based explanations, enabling both global and instance-level understanding of predictions. The results demonstrate that reliable academic outcome prediction can be achieved without sacrificing interpretability, providing a transparent and trustworthy decision-support framework for educational analytics.

Keywords: Explainable Machine Learning; Academic Outcome Prediction; Learning Analytics; Explainable Boosting Machine; Model Interpretability.

1. Introduction

The growing availability of educational data has significantly accelerated the adoption of learning analytics and predictive modeling in higher education. Institutions increasingly rely on data-driven approaches to understand student behavior, monitor academic progress, and

predict learning outcomes, enabling timely interventions and informed decision-making [1]. Among these applications, academic outcome classification plays a critical role in identifying students' final performance categories, such as distinction, pass, fail, or withdrawal, based on academic and learning-related indicators [2]. Accurate classification of academic outcomes enables early identification of at-risk students, facilitates targeted academic support, and informs institutional planning and policy formulation. A wide range of machine learning techniques has been explored for academic outcome prediction, including decision trees, support vector machines, artificial neural networks, random forests, and gradient boosting models [3]. While many of these approaches achieve strong predictive performance, they are often treated as black-box systems, offering limited insight into how predictions are generated or which factors drive classification decisions.

In educational settings where predictive outcomes may influence academic progression, student support strategies, and institutional policy [4], the lack of interpretability poses significant challenges for transparency, trust, accountability, and practical adoption. Educators and administrators may be reluctant to rely on models whose internal logic cannot be easily understood, validated, or explained to stakeholders [5]. Interpretable machine learning has therefore gained increasing attention in educational data mining and learning analytics.

Transparent models allow educators and stakeholders to understand the relationship between input features and predicted outcomes, validate model behavior against domain knowledge, and communicate results effectively to non-technical audiences. In addition, interpretability supports ethical and responsible AI practices by enabling the detection of potential biases, spurious correlations, and unfair decision patterns. However, achieving interpretability without sacrificing predictive accuracy remains a key challenge, particularly for complex multi-class classification tasks involving heterogeneous academic, behavioral, and demographic features [2].

The Explainable Boosting Machine (EBM) has recently emerged as a promising glass-box modeling approach that addresses this challenge [6], [7]. EBM combines the predictive power of gradient boosting with the transparency of generalized additive models, producing feature-level contribution functions that are inherently interpretable. Unlike post hoc explanation techniques applied to black-box models [8]. EBM provides native interpretability, enabling direct inspection of both global feature effects and instance-level prediction logic. This dual capability allows domain experts to assess model behavior more reliably and ensures that predictive decisions remain aligned with pedagogical expectations and institutional objectives.

In this study, we propose an interpretable academic outcome classification framework based on the Explainable Boosting Machine and evaluate its effectiveness using a publicly available Student Performance and Learning Behavior dataset comprising 6,519 student records. The dataset includes academic, behavioral, and demographic attributes, with a multi-class target variable representing four academic outcomes: Distinction, Pass, Fail, and Withdrawn. Model performance is assessed using a comprehensive set of evaluation metrics, including accuracy, balanced accuracy, macro-averaged F1-score, and class-wise precision and recall, complemented by confusion matrix analysis and multi-class ROC curves. In addition to predictive performance, model transparency is examined through EBM's inherent feature contribution functions and complementary SHAP-based local explanation analysis to provide both global and local interpretability [9].

The main contributions of this work are threefold. First, it demonstrates the effectiveness of a glass-box machine learning model for multi-class academic outcome classification. Second, it provides a fully interpretable framework that enables a clear understanding of how academic, behavioral, and demographic features influence predicted outcomes. Third, it offers empirical evidence that high predictive performance can be achieved without resorting to opaque black-box models, thereby supporting the broader adoption of explainable machine learning techniques in educational analytics [10].

The study contributes to the growing body of research on explainable artificial intelligence in education by presenting a transparent, accurate, and practically deployable approach for academic outcome classification. By bridging the gap between interpretability and predictive performance, the proposed framework advances the development of trustworthy, accountable, and student-centered learning analytics systems capable of supporting data-driven decision-making in modern higher education.

The remainder of this paper is organized as follows. Section 2 reviews related literature on academic outcome prediction and interpretable machine learning in educational analytics. Section 3 presents the research methodology, including dataset description, pre-processing procedures, model development, and evaluation metrics. Section 4 reports the experimental results and discusses the predictive performance and interpretability analysis of the proposed framework. Finally, Section 5 concludes the study and outlines potential directions for future research.

2. Literature Review

The application of learning analytics and educational data mining has grown substantially over the past decade, driven by the increasing availability of large-scale educational datasets and advances in machine learning. Researchers have widely explored predictive modeling techniques to analyze student behavior, forecast academic performance, and classify learning outcomes to support early interventions and institutional decision-making [11].

Academic outcome classification, in particular, has been recognized as a critical task for identifying students at risk of failure or withdrawal and for enabling proactive academic support strategies. Early studies in this domain primarily employed traditional statistical and machine learning methods such as logistic regression, naïve Bayes classifiers, k-nearest neighbors, and decision trees [12]. These models offered a degree of transparency and ease of interpretation, making them attractive for educational settings. For example, decision tree-based approaches have been used to model student progression and identify key performance drivers due to their rule-based structure and intuitive visualization capabilities. However, such models often struggle to capture complex nonlinear relationships and interactions among heterogeneous academic and behavioral features, limiting their predictive performance in large and diverse datasets [13].

To overcome these limitations, more advanced machine learning models have been increasingly adopted. Support vector machines, artificial neural networks, random forests, and gradient boosting algorithms have demonstrated strong predictive accuracy in various academic outcome prediction tasks [14], [15]. Ensemble learning methods, in particular, have shown superior performance by combining multiple weak learners to improve generalization and robustness. Random forests and boosting-based models have been widely reported to outperform single classifiers in predicting student grades, dropout risk, and final academic standing. Despite these performance gains, such models are typically treated as black boxes, offering little transparency into how predictions are generated or which features are most influential in driving outcomes [16]. The lack of interpretability in black-box models has emerged as a significant concern in educational applications. Predictive systems are increasingly used to inform high-stakes decisions, including academic probation, scholarship allocation, and personalized learning interventions. Without clear explanations, educators and administrators may find it difficult to trust model outputs, justify decisions to students, or verify that predictions are aligned with pedagogical principles and institutional policies [17], [18]. Moreover, opaque models raise ethical and regulatory concerns related to fairness, accountability, and potential bias, particularly when demographic or socioeconomic variables are included as predictors.

In response to these challenges, the field has witnessed growing interest in explainable artificial intelligence (XAI) and interpretable machine learning for educational analytics. Post hoc explanation techniques such as SHAP (SHapley Additive exPlanations) [19], LIME (Local Interpretable Model-Agnostic Explanations), and partial dependence plots have been widely used to interpret predictions from black-box models. Several studies have demonstrated that SHAP-based feature importance analysis can provide valuable insights into the factors influencing student performance predictions, highlighting the dominant role of variables such as exam scores, attendance, and assignment completion. However, post hoc explanations are inherently approximations of complex models and may not always provide faithful or stable interpretations, especially in highly nonlinear or high-dimensional settings [20], [21].

To address the limitations of post hoc explainability, researchers have increasingly explored glass-box models that are interpretable by design. Generalized additive models (GAMs), rule-based learners, and sparse linear models have been revisited in educational contexts due to their transparency and ease of interpretation. While these models offer native explainability, they often sacrifice predictive performance when applied to complex, real-world educational datasets. This trade-off between accuracy and interpretability has remained a central challenge in the design of trustworthy educational prediction systems. The Explainable Boosting Machine (EBM) has recently emerged as a promising approach that bridges this gap between transparency and performance. EBM is a boosted generalized additive model that learns flexible, nonlinear feature contribution functions while retaining full interpretability [22]. By incorporating limited pairwise interaction terms, EBM can capture moderate feature dependencies without compromising transparency. Unlike deep learning or traditional ensemble methods, EBM produces human-readable feature effect plots that allow direct inspection of how each input variable influences the predicted outcome. EBM has been successfully applied in several high-stakes domains, including healthcare risk prediction, financial credit scoring, and fraud detection, where interpretability is critical for trust and regulatory compliance [23].

Despite its demonstrated potential, the application of EBM in educational data mining remains relatively limited. Most existing studies in academic outcome prediction continue to rely on black-box ensemble models supplemented with post hoc explanation techniques. Only a small number of works have explored inherently interpretable boosting-based models for student performance classification, and even fewer have systematically evaluated their effectiveness in multi-class academic outcome classification settings. Furthermore, many prior studies focus on binary classification tasks (e.g., pass/fail or dropout prediction), leaving a gap in understanding how interpretable models perform when distinguishing among multiple academic outcome categories such as distinction, pass, fail, and withdrawal. [24] Examines an AI-driven Educational Data Mining framework for student academic performance classification, combining machine learning and explainable AI techniques. Among several evaluated models, Extreme Gradient Boosting achieved the highest accuracy (83%).

Explainability analyses revealed that curriculum evaluation, course-related factors, and economic variables across semesters significantly influence academic performance. The work also introduces a decision-support system to provide personalized recommendations, with future extensions focusing on real-time data and mobile integration. [9] Applies machine learning techniques to predict vocational undergraduate students' final exam performance and identify key influencing factors. Multiple models were evaluated using Ridge Regression as a baseline, with Random Forest delivering the strongest predictive results. Model interpretability analyses using SHAP revealed that academic-related features, particularly monthly exam scores, admission scores, and self-study time, were the most influential, while demographic variables had a limited impact.

Further analysis using Partial Dependence Plots and Kernel Density Estimation highlighted performance differences between high- and low-achieving students, providing actionable insights for targeted academic interventions in vocational education. [8] investigates the enhancement of Performance Factors Analysis (PFA), a widely used Knowledge Tracing approach for adaptive educational systems, by integrating ensemble machine learning techniques. While existing PFA extensions primarily focus on pedagogical improvements through behavioral analysis, this work emphasizes technical enhancements in predictive accuracy. [25] Ensemble models Random Forest, AdaBoost, and XGBoost are incorporated into a novel PFA framework and evaluated across three datasets. Experimental results demonstrate that XGBoost consistently outperforms the other models and significantly improves student performance prediction compared to the original PFA algorithm, highlighting the effectiveness of ensemble learning for advanced knowledge tracing [26], [27].

Recent research has also highlighted the importance of combining global and local interpretability in educational analytics. Global explanations provide an overall understanding of feature importance and model behavior across the entire dataset, while local explanations offer instance-level insights into individual predictions. Integrating both perspectives enables educators to understand not only which factors are generally influential but also why specific students are assigned to particular outcome categories [28]. However, most existing approaches rely on post hoc explanation frameworks layered on top of black-box models, rather than leveraging models that are inherently transparent at both global and local levels.

The existing literature demonstrates substantial progress in academic outcome prediction using machine learning, with ensemble and deep learning models achieving strong predictive performance. However, the widespread reliance on black-box models poses significant challenges for transparency, trust, and ethical deployment in educational contexts [29]. While post hoc explanation techniques provide partial solutions, they do not fully address the need for faithful and stable interpretability. The Explainable Boosting Machine offers a compelling alternative by combining native interpretability with competitive predictive accuracy. Nevertheless, its application to multi-class academic outcome classification remains underexplored, and empirical evidence on its effectiveness in educational analytics is still limited. Motivated by these gaps, this study investigates the use of an Explainable Boosting Machine based framework for interpretable multi-class academic outcome classification, aiming to demonstrate that high predictive performance can be achieved without sacrificing transparency. By systematically evaluating both predictive accuracy and model interpretability, this work contributes to the development of trustworthy, explainable, and practically deployable learning analytics systems for modern higher education.

3. Methodology

This study adopts a quantitative, experimental research design to develop and evaluate an interpretable framework for multi-class academic outcome prediction using the Explainable Boosting Machine (EBM). The methodological workflow consists of dataset acquisition, data preprocessing, model development, performance evaluation, and interpretability analysis. The overall objective is to achieve reliable predictive performance while maintaining full model transparency for educational decision-support applications.

Figure 1 illustrates the methodological stages of the study, beginning with dataset acquisition from the Student Performance and Learning Behavior dataset. The workflow proceeds through data pre-processing, including data cleaning, feature engineering, and encoding. The Explainable Boosting Machine (EBM) model is then developed and trained for multi-class academic outcome prediction. Model performance is evaluated using classification metrics and ROC analysis. Finally, interpretability analysis is conducted using EBM feature contribution functions and SHAP explanations to provide both global and instance-level insights before deriving the final research findings and conclusions.

3.1. Dataset Description

Experiments were conducted using the Student Performance and Learning Behavior dataset from Kaggle, containing 6,519 student records. The dataset includes academic, behavioral, and demographic features describing students' learning activities. The target variable represents four academic outcome classes: Distinction, Pass, Fail, and Withdrawn, formulated as a multi-class classification task.

To clarify the relationship between the predictor variables and the target label, it is important to note that the

Table 1. Description of Dataset Variables.

Feature	Type	Description
Study Hours	Numerical	Number of hours a student spends studying per week.
Attendance	Numerical	Student's attendance rate expressed as a percentage.
Resources	Categorical	Access to learning resources (0 = No, 1 = Yes).
Extracurricular	Categorical	Participation in extracurricular activities (0 = No, 1 = Yes).
Motivation	Categorical	Student's motivation level represented as categorical ratings.
Internet	Categorical	Availability of internet access for learning (0 = No, 1 = Yes).
Gender	Categorical	Student's gender (e.g., 0 = Male, 1 = Female).
Age	Numerical	Age of the student in years.
Learning Style	Categorical	Preferred learning style encoded using categorical codes.
Online Courses	Numerical	Number of online courses taken by the student.
Discussions	Numerical	Level of participation in academic discussions.
Assignment Completion	Numerical	Percentage of assignments completed by the student.
Exam Score	Numerical	Score obtained in course examinations.
EduTech	Categorical	Use of educational technology tools for learning (0 = No, 1 = Yes).
Stress Level	Categorical	Student's stress level represented using categorical ratings.

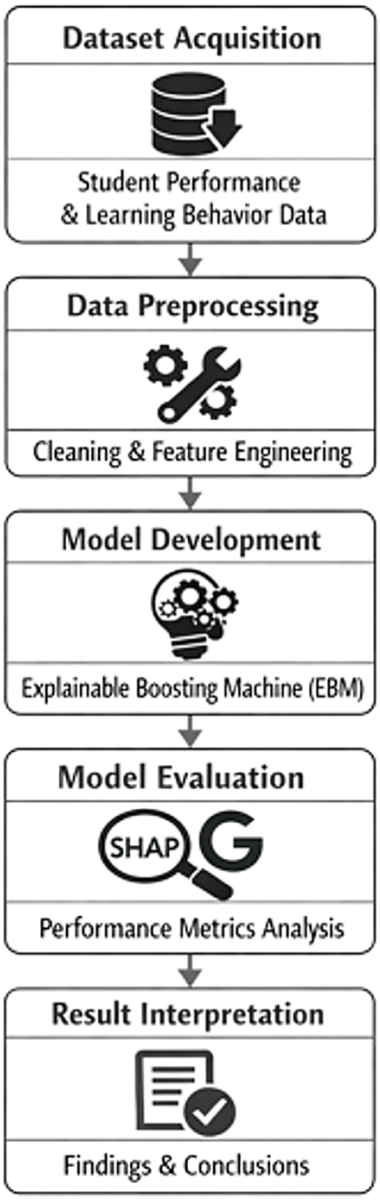

Figure 1. Research Workflow of the Proposed Interpretable Academic Outcome Prediction Framework.

academic outcome categories used in this study (Distinction, Pass, Fail, and Withdrawn) are provided as predefined categorical labels within the dataset and are not directly computed from the ExamScore variable used in the model features. The ExamScore attribute represents an individual academic assessment indicator that reflects examination performance during the course, whereas the final academic outcome variable represents the overall student status determined by the dataset's labeling scheme. Therefore, ExamScore functions as a predictive academic feature rather than the basis for constructing the target variable. This distinction ensures that the modeling process does not involve target leakage, and the model learns meaningful relationships between academic, behavioral, and contextual features and the final academic outcome.

Table 1 presents the variables used in the Student Performance and Learning Behaviour dataset for academic outcome prediction. The dataset includes numerical features that capture students' academic engagement and performance, such as *Study Hours, Attendance, Online Courses, Discussions, Assignment Completion,* and *Exam Score*. It also contains several categorical variables representing contextual and behavioural factors, including *Resources, Extracurricular participation, Motivation, Internet access, Learning Style, EduTech usage,* and *Stress Level*, as well as demographic attributes such as *Age* and *Gender*. The target variable, labelled Outcome, represents the final academic classification of students into four categories: *Distinction, Pass, Fail,* and *Withdrawn*. Together, these variables provide a comprehensive representation of students' academic behaviour and learning environment, enabling the development of a multi-class predictive model.

Figure 2 presents the correlation heatmap illustrating the relationships between the input variables and the academic outcome variable. The visualization highlights the strength and direction of associations among the academic,

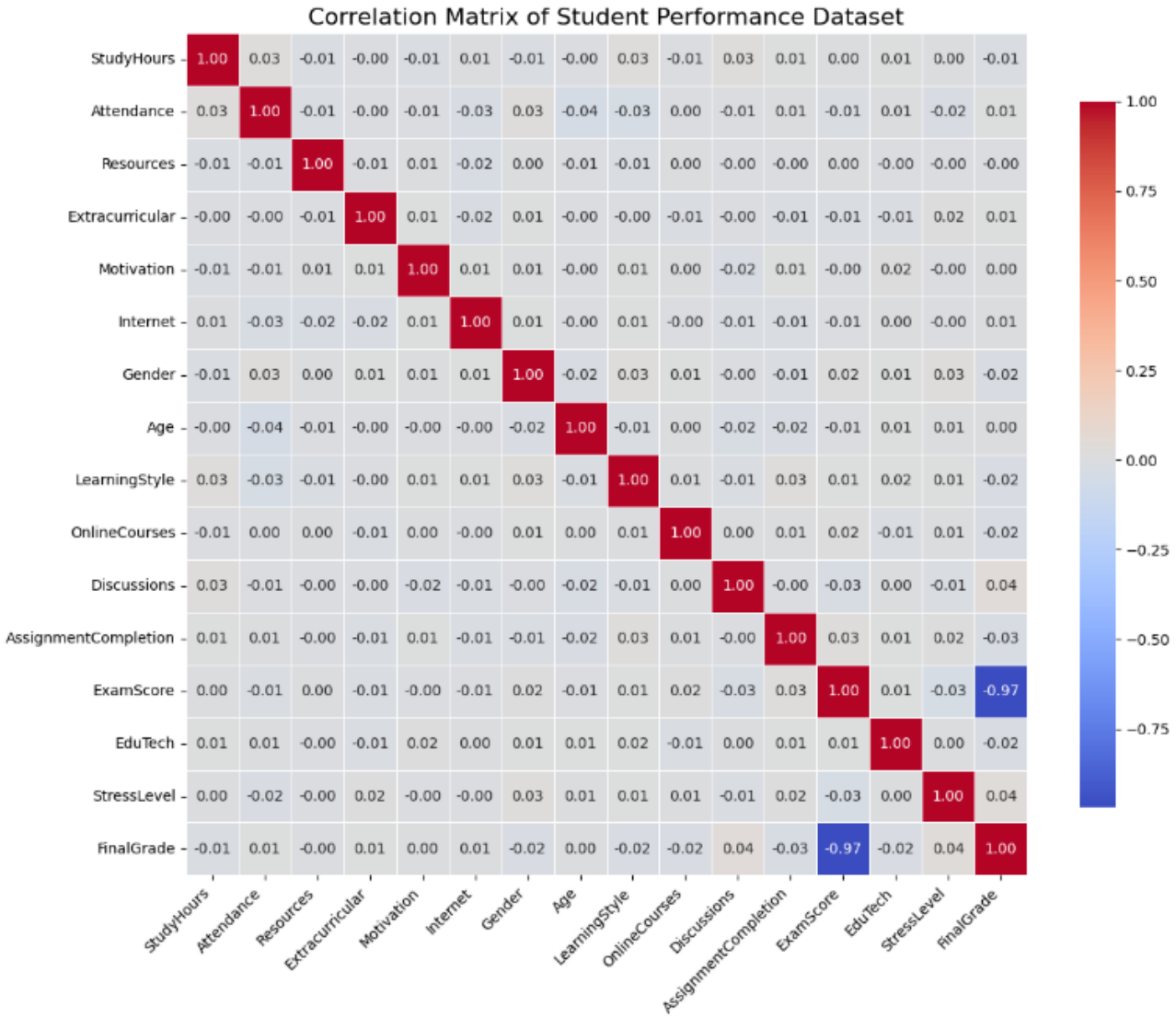

Figure 2. Correlation matrix of the variables used in the student performance dataset.

behavioral, and demographic attributes used in the study. ExamScore shows the strongest correlation with the academic outcome, indicating its dominant influence on student performance. In contrast, behavioral variables such as attendance and study hours exhibit relatively weak associations, while demographic features show minimal correlations, suggesting that academic outcomes are primarily driven by learning-related factors rather than demographic characteristics.

3.2. Data Pre-processing

Prior to model training, several preprocessing steps were applied to ensure data quality and consistency. Missing values in numerical variables were handled using median imputation to reduce sensitivity to extreme values, while categorical variables were imputed using their most frequent category. Outliers in continuous features, including examination scores and study hours, were identified using interquartile range (IQR) analysis and clipped to reasonable bounds to prevent undue influence on model learning. Categorical variables were transformed into numerical representations using appropriate encoding schemes based on their cardinality and semantic meaning. Although the Explainable Boosting Machine is relatively robust to feature scaling due to its internal binning mechanism, continuous variables were standardized to promote numerical stability.

3.3. Model Development

The study employed the Explainable Boosting Classifier (EBC), an interpretable model that combines boosting with additive feature contributions and limited pairwise interactions. Hyperparameters were set to balance accuracy and interpretability: learning rate = 0.01, max 10 interaction terms, 256 bins for continuous features, and 500 boosting rounds. Uniform class weighting was used due to balanced outcome classes, ensuring stable training and transparent, human-readable feature explanations.

The complete training workflow of the proposed Explainable Boosting Machine model is summarized in Algorithm 1, which describes the pre-processing procedures, boosting iterations, interaction learning, and evaluation steps involved in the modeling process.

Algorithm 1. EBM Training Workflow (Pseudocode).

Student dataset $D = \{X, y\}$, where X denotes features and y denotes labels;
Hyperparameters $\theta = \{lr, T, B, R\}$, where lr is the learning rate, T is the number of interactions, B is the maximum number of bins, and R is the maximum number of boosting rounds. Trained Explainable Boosting Machine (EBM) model M.
Begin

Step 1: *Data Preprocessing: Handle missing values in X using median imputation for numerical features and mode imputation for categorical features. Detect and clip outliers in continuous features using the interquartile range (IQR). Encode categorical variables into numerical representations, standardize continuous features, and split D into the training set D_{train} and test set D_{test} using stratified sampling*

Step 2: *Model Initialization Initialize EBM model M with hyperparameters θ Set the intercept term $\beta_0 \leftarrow 0$ Initialize feature contribution functions $f_i(x_i) \leftarrow 0$ for all features x_i*

Step 3: *Boosting Iterations Compute residuals:*

$$residuals = y_{train} - M.predict(X_{train})$$

Fit weak learner $h_i(x_i)$ to residuals Update feature contribution:

$$f_i(x_i) \leftarrow f_i(x_i) + lr \times h_i(x_i)$$

Step 4: *Learn Pairwise Interactions (Optional) Select feature pair (x_i, x_j) Fit interaction learner $h_{ij}(x_i, x_j)$ to residuals Update interaction contribution:*

$$f_{ij}(x_i, x_j) \leftarrow f_{ij}(x_i, x_j) + lr \times h_{ij}(x_i, x_j)$$

Step 5: *Update model M using updated f_i and f_{ij} Check convergence or early stopping criteria break*

Step 6: *Model Evaluation Evaluate M on D_{test} using accuracy, precision, recall, F1-score, and ROC-AUC*

End

Table 2. Configuration of the Explainable Boosting Classifier.

Parameter	Value
Model	Explainable Boosting Classifier (EBC)
Learning Rate	0.01
Number of Interactions	10
Maximum Bins	256
Maximum Rounds	500

Table 3. Classification Report of the Proposed EBM Model.

Class	Precision	Recall	F1-score	Support
Distinction	0.90	0.88	0.89	1784
Pass	0.82	0.90	0.86	1541
Fail	0.90	0.84	0.87	1684
Withdrawn	0.90	0.91	0.91	1510
Accuracy			0.88	6519
Macro avg	0.88	0.88	0.88	6519
Weighted avg	0.88	0.88	0.88	6519

3.4. Model Training and Evaluation

The performance of the proposed model was evaluated using several standard multi-class classification metrics derived from the confusion matrix. Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

Recall:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

F1-score:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

ROC–AUC: The Area Under the Receiver Operating Characteristic Curve (ROC–AUC) measures the model's ability to discriminate between classes across different classification thresholds.

3.5. Model Interpretability Analysis

Interpretability was examined using two complementary approaches. The Explainable Boosting Machine (EBM) provides inherent global explanations via feature contribution functions, showing how feature values influence predicted academic outcomes. To complement this, SHAP-based techniques were applied for instance-level explanations, quantifying each feature's contribution to individual predictions. Integrating global and local analyses offers a comprehensive framework, enabling a clear understanding of both overall model behavior and the factors driving specific student predictions.

4. Results and Discussion

This section presents the experimental results obtained from the proposed Explainable Boosting Machine (EBM) model and discusses the predictive performance and interpretability outcomes. The analysis includes model configuration, quantitative performance evaluation, confusion matrix analysis, ROC curve assessment, and interpretability insights derived from feature importance and SHAP explanations.

Table 2 summarizes the key configuration parameters used for the Explainable Boosting Classifier (EBC). The model was trained with a learning rate of 0.01, enabling

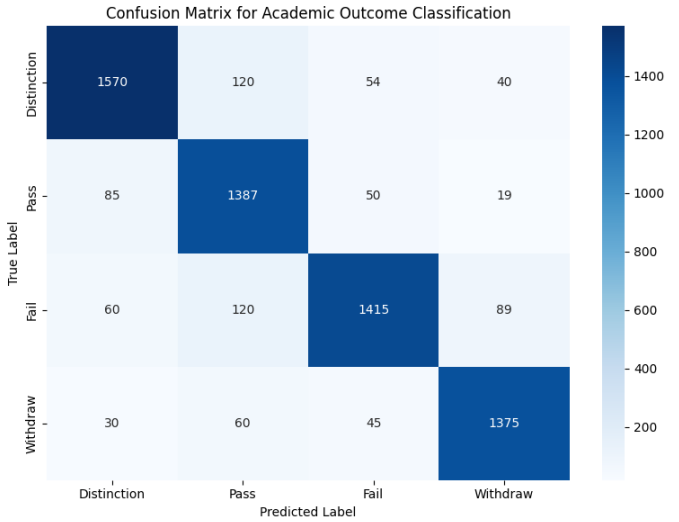

Figure 3. Confusion matrix for academic outcome classification.

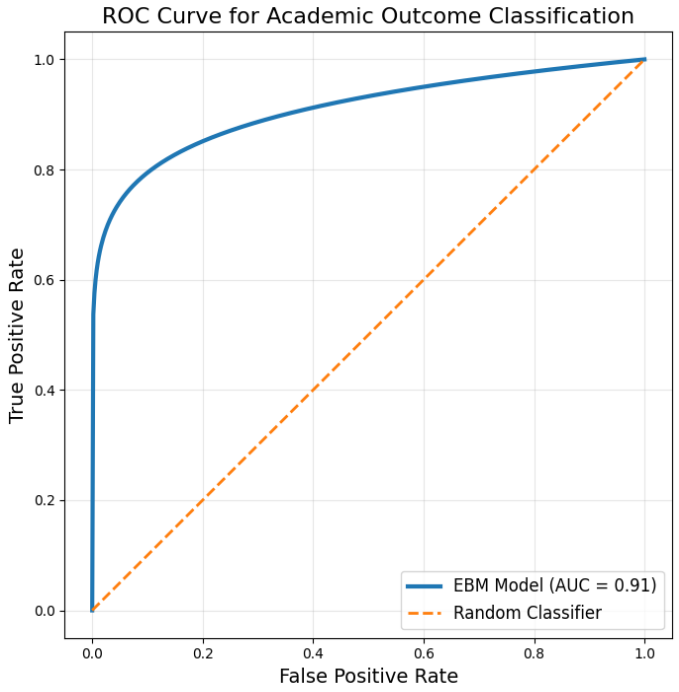

Figure 4. Predictive Performance Evaluation Using ROC Analysis.

gradual and stable updates during training. A total of 10 interaction terms were allowed, supporting the capture of limited but interpretable feature interactions. The maximum number of bins (256) was used to discretize continuous features with sufficient granularity, while the model was trained for up to 500 boosting rounds to ensure convergence. Collectively, these settings balance predictive performance with interpretability, aligning with the objectives of explainable learning analytics models.

Table 3 presents the classification performance of the proposed model across the four outcome categories: Distinction, Pass, Fail, and Withdrawn. The model achieves an overall accuracy of 0.88 on 6,519 instances**,** demonstrating strong predictive capability. Performance across the classes is balanced, with precision, recall, and F1-scores

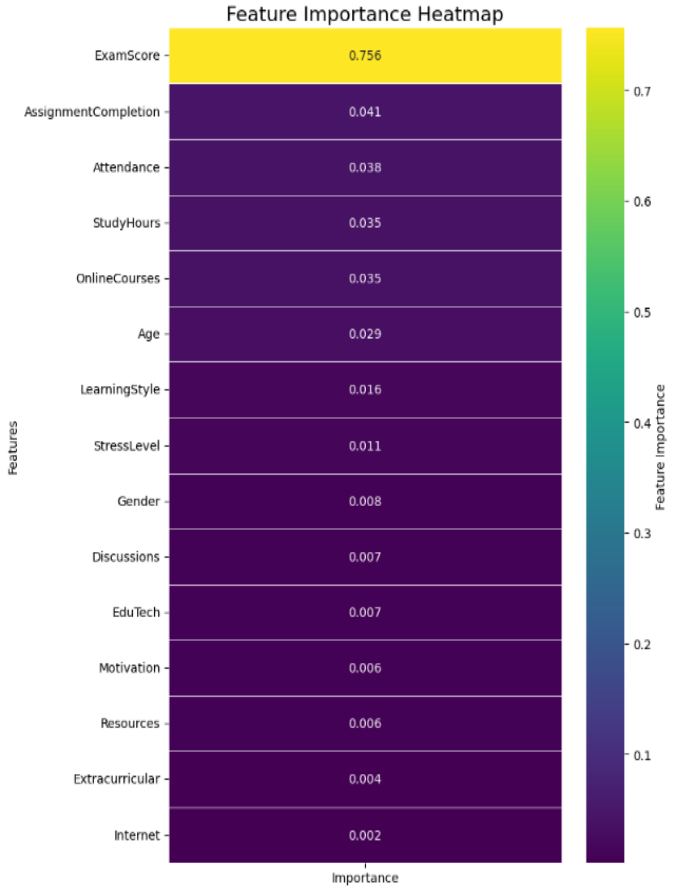

Figure 5. Feature Importance Heatmap Based on Model Explanations.

generally ranging between 0.84 and 0.91. The Withdraw class shows the highest performance with an F1-score of 0.91**,** while the Pass class records the lowest precision (**0.82**) but maintains high recall (0.90**).** The macro and weighted averages of 0.88 further indicate consistent classification performance across all categories. Overall, the results demonstrate that the model provides reliable and balanced predictions of student academic outcomes.

Figure 3 show the Confusion matrix for academic outcome classification using the Explainable Boosting Machine (EBM) model. The matrix shows the distribution of predicted and actual student outcome classes across Distinction, Pass, Fail, and Withdrawn categories. Most predictions appear along the diagonal elements, indicating correct classifications, while a small number of off-diagonal entries represent misclassifications between related outcome categories. The model achieved an overall accuracy of approximately 88%, which is consistent with the performance metrics reported in Table 3.

Figure 4 shows the one-vs-rest receiver operating characteristic (ROC) curves for the four academic outcome classes. All curves lie well above the diagonal reference line, indicating strong discriminative capability across classes. The overall ROC–AUC value of 0.91 confirms that the model effectively distinguishes between outcome categories under varying classification thresholds. The consistent shape of the ROC curves across classes suggests

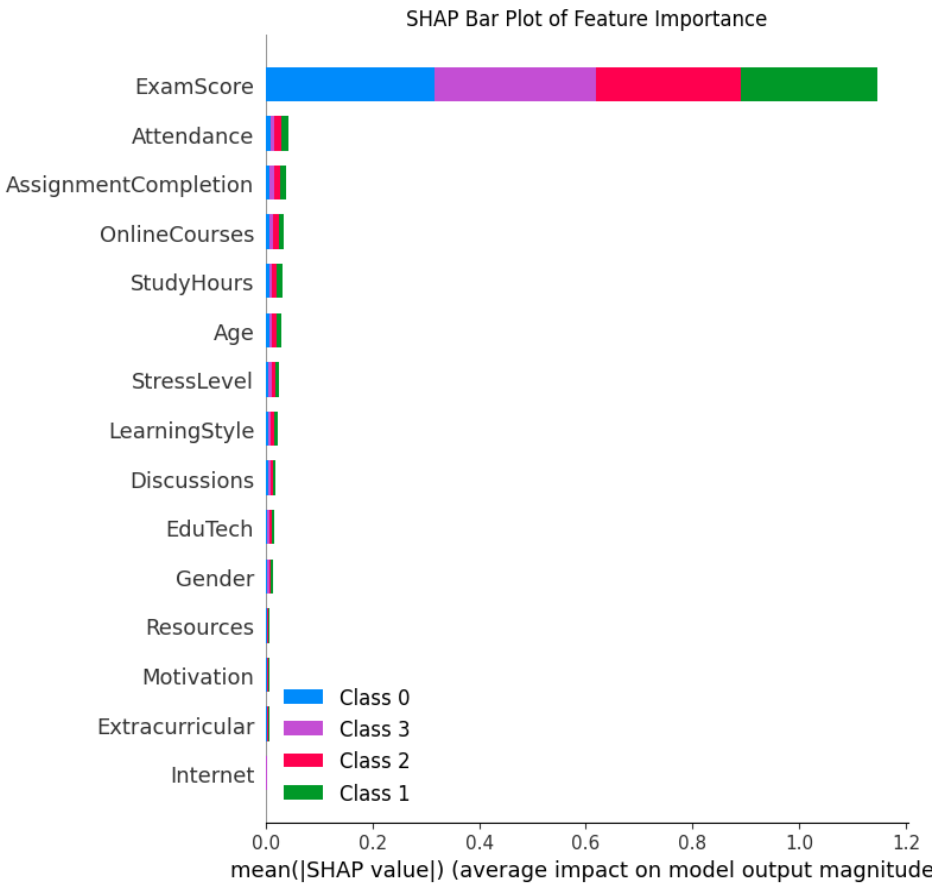

Figure 6. SHAP-Based Feature Importance Across Classes.

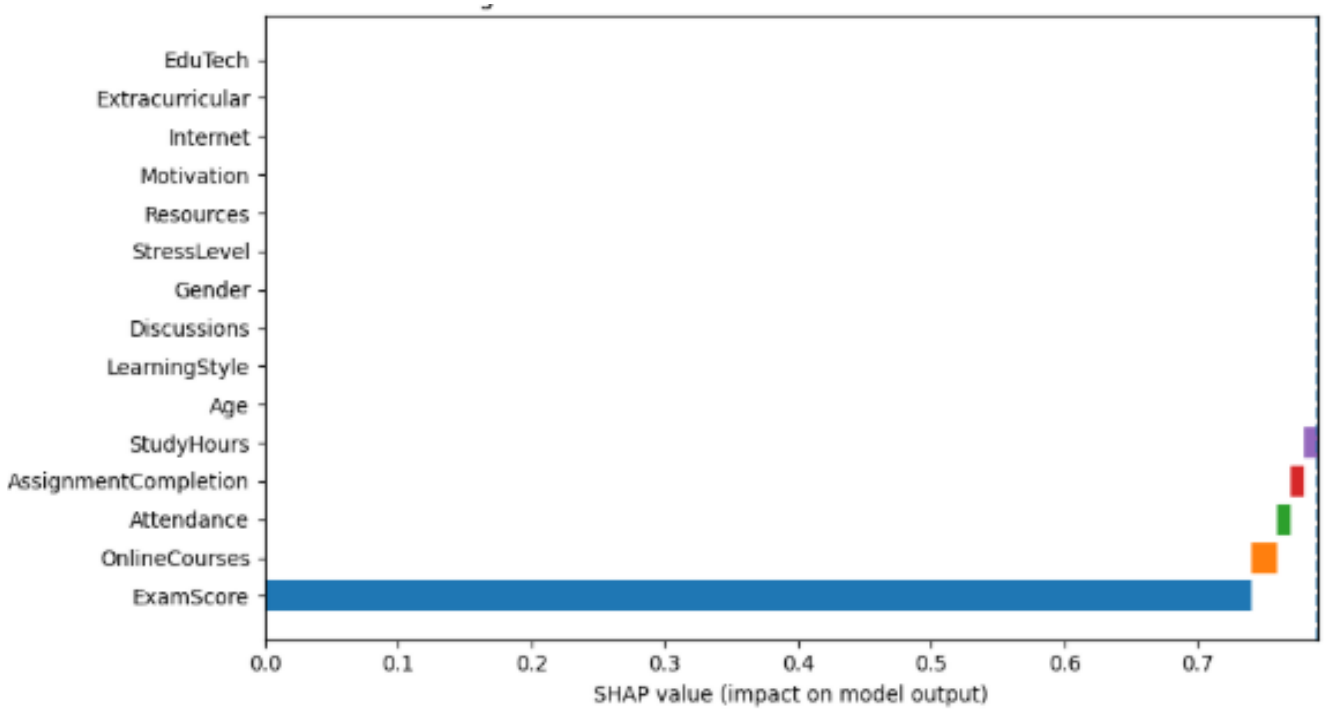

Figure 7. SHAP Waterfall Plot for Individual Prediction.

stable and robust predictive behavior, further validating the effectiveness of the proposed EBM framework for multi-class academic outcome prediction.

Figure 5 shows a heatmap of feature importance from the model's interpretability analysis. Exam Score is the dominant predictor, behavioral features contribute moderately, and demographic factors have minimal impact. This pattern aligns with educational theory, indicating that predictions are driven by academically relevant features and reinforcing the model's transparency and ethical use.

Figure 6 presents the SHAP-based global feature importance across the four outcome classes using mean absolute SHAP values. Exam Score emerges as the dominant predictor, contributing substantially more to the model's predictions than other variables. Behavioral features show moderate influence, while demographic attributes contribute minimally. This consistency indicates that the model primarily relies on academically relevant indicators, supporting the interpretability and validity of the proposed framework.

Figure 7 shows a SHAP waterfall plot for an individual EBM prediction, decomposing the output into feature contributions. Exam Score dominates the prediction, while behavioral, contextual, and demographic features contribute minimally. This instance-level explanation aligns with the global SHAP analysis, highlighting that predictions are primarily driven by examination performance, which supports transparent decision-making for individual students.

The predictive performance of the proposed EBM framework compares favorably with previous studies on academic outcome prediction. Prior research using ensemble learning models such as Random Forest and XGBoost has reported accuracy values ranging between approximately 80% and 87% for student performance classification tasks. In contrast, the proposed EBM model achieves an accuracy of 88% while maintaining full model transparency. Unlike many high-performing black-box models, the EBM framework provides interpretable feature contribution functions that allow educators to understand how academic and behavioral factors influence predicted outcomes. This combination of competitive predictive accuracy and inherent interpretability highlights the practical value of the proposed approach for educational decision-support systems.

5. Conclusion

This study presents an interpretable machine learning framework for multi-class academic outcome prediction using the Explainable Boosting Machine (EBM). The model achieved strong and balanced performance (accuracy 88%, AUC 0.91) across four outcome categories: Distinction, Pass, Fail, and Withdrawn. It provides both global and instance-level interpretability through feature contribution functions and SHAP explanations, enabling clear insights into the factors driving predictions. Examination score emerged as the dominant predictor, behavioral features contributed moderately, and demographic variables had minimal influence, consistent with educational theory. These results demonstrate that high predictive accuracy can be achieved without black-box models, supporting transparent, accountable, and ethical deployment of AI in educational decision-making. The proposed EBM-based framework offers a practical, trustworthy solution for academic outcome prediction and advances the application of explainable AI in learning analytics.

6. Declarations

6.1. Author Contributions

Godfrey Perfectson Oise: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Writing – Original Draft, Visualization; **Felix Oshiorenoya Uloko:** Methodology, Formal Analysis, Validation, Writing – Review & Editing; **Kevin Chinedu Pius:** Investigation, Data Curation, Resources, Writing – Original Draft; **Enovwo Eferoba–Idio:** Investigation, Data Curation, Resources, Writing – Original Draft; **Michael Uyiosa Edobor:** Resources, Data Curation, Investigation, Writing – Review & Editing; **Evans Mintah:** Supervision, Formal Analysis, Writing – Review & Editing; **Osahon Ukpebor:** Supervision, Project Administration, Writing – Review & Editing; **Oludare Sokoya**: Validation, Visualization, Writing – Review & Editing; **Tejiri Jessa:** Resources, Data Curation, Writing – Review & Editing; All authors have read and approved the final version of the manuscript.

6.2. Institutional Review Board Statement

Not applicable.

6.3. Informed Consent Statement

Not applicable.

6.4. Data Availability Statement

The dataset used in this study is publicly available and can be accessed at: https://www.kaggle.com/datasets/adilshamim8/student-performance-and-learning-style.

6.5. Acknowledgment

Not applicable.

6.6. Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

7. References

[1] R. O. Mensah, K. D. Amponsah, P. Adiza Babah, and H. S. Jibril, "Factors affecting students' academic performance and teachers' efficiency in Ghana; a case study of Wa senior high school," *Cogent Arts Humanit.*, vol. 11, no. 1, Dec. 2024. https://doi.org/10.1080/23311983.2024.2412944.

[2] Á. Kocsis and G. Molnár, "Factors influencing academic performance and dropout rates in higher education," *Oxf. Rev. Educ.*, vol. 51, no. 3, pp. 414–432, May 2025. https://doi.org/10.1080/03054985.2024.2316616.

[3] A. Mappadang, K. Khusaini, M. Sinaga, and E. Elizabeth, "Academic interest determines the academic performance of undergraduate accounting students: Multinomial logit evidence," *Cogent Business & Management*, vol. 9, no. 1, Dec. 2022. https://doi.org/10.1080/23311975.2022.2101326.

[4] E. Ahmed, "Student Performance Prediction Using Machine Learning Algorithms," *Applied Computational Intelligence and Soft Computing*, vol. 2024, 2024. https://doi.org/10.1155/2024/4067721.

[5] A. Alhothali, M. Albsisi, H. Assalahi, and T. Aldosemani, "Predicting Student Outcomes in Online Courses Using Machine Learning Techniques: A Review," *Sustainability (Switzerland)*, vol. 14, no. 10, May 2022. https://doi.org/10.3390/SU14106199.

[6] S. Alraddadi, S. Alseady, and S. Almotiri, "Prediction of students academic performance utilizing hybrid teaching-learning based feature selection and machine learning models," *2021 International Conference of Women in Data Science at Taif University, WiDSTaif 2021*, Mar. 2021. https://doi.org/10.1109/WIDSTAIF52235.2021.9430248.

[7] B. Albreiki, N. Zaki, and H. Alashwal, "A systematic literature review of student' performance prediction using machine learning techniques," *Educ. Sci. (Basel).*, vol. 11, no. 9, Sep. 2021. https://doi.org/10.3390/EDUCSCI11090552.

[8] A. Asselman, M. Khaldi, and S. Aammou, "Enhancing the prediction of student performance based on the machine learning XGBoost algorithm," *Interactive Learning Environments*, vol. 31, no. 6, pp. 3360–3379, Aug. 2023. https://doi.org/10.1080/10494820.2021.1928235.

[9] 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, Jul. 2025. https://doi.org/10.1080/14703297.2025.2532050.

[10] G. perfectson Oise, Ejenarhome Otega Prosper, Augustine Osazee Airhiavbere, and Agwam Gladys Ifeoma, "Student Success Prediction in Digital Learning Environments," *Journal Of Digital Learning and Distance Education*, vol. 4, no. 6, pp. 1697–1707, Nov. 2025. https://doi.org/10.56778/jdlde.v4i6.592.

[11] S. A. Oyedotun, O. P. Ejenarhome, and G. P. Oise, "Learning Analytics and Predictive Modeling: Enhancing Student Success through Data-Driven Insights," *Journal of Science Research and Reviews*, vol. 2, no. 3, pp. 42–51, Jul. 2025. https://doi.org/10.70882/josrar.2025.v2i3.77.

[12] G. Oise, Ejenarhome Otega Prosper, Oyedotun Samuel Abiodun, and Onwuzo Chioma Julia, "Evaluating the Impact of Blended Learning Models on Higher Education Outcomes: A Multidimensional Analysis," *Journal Of Digital Learning and Distance Education*, vol. 4, no. 2, pp. 1507–1519, Jul. 2025. https://doi.org/10.56778/jdlde.v4i2.535.

[13] G. G. James, G. P, Oise, E. G. Chukwu, N. A. Michael, W. F. Ekpo, and P. E. Okafor, "Optimizing Business Intelligence System Using Big Data and Machine Learning," *Journal of Information Systems and Informatics*, vol. 6, no. 2, pp. 1215–1236, Jun. 2024. https://doi.org/10.51519/journalisi.v6i2.631.

[14] D. Alboaneen, M. Almelihi, R. Alsubaie, R. Alghamdi, L. Alshehri, and R. Alharthi, "Development of a Web-Based Prediction System for Students' Academic Performance," *Data (Basel).*, vol. 7, no. 2, Feb. 2022. https://doi.org/10.3390/DATA7020021.

[15] M. Abou Naaj, R. Mehdi, E. A. Mohamed, and M. Nachouki, "Analysis of the Factors Affecting Student Performance Using a Neuro-Fuzzy Approach," *Educ. Sci. (Basel).*, vol. 13, no. 3, Mar. 2023. https://doi.org/10.3390/EDUCSCI13030313.

[16] A. Alhassan, B. Zafar, and A. Mueen, "Predict students' academic performance based on their assessment grades and online activity data," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, 2020. https://doi.org/10.14569/IJACSA.2020.0110425.

[17] K. Vimarsha, S. P. S. Prakash, K. Krinkin, and Y. A. Shichkina, "Student Performance Prediction: A Co-Evolutionary Hybrid Intelligence model," *Procedia Comput. Sci.*, vol. 235, pp. 436–446, 2024. https://doi.org/10.1016/J.PROCS.2024.04.043.

[18] Y. Alshamaila *et al.*, "An automatic prediction of students' performance to support the university education system: a deep learning approach," *Multimed. Tools Appl.*, vol. 83, no. 15, pp. 46369–46396, May 2024. https://doi.org/10.1007/S11042-024-18262-4.

[19] G. Oise and S. Konyeha, "Environmental impacts in e-waste management using deep learning," *Discover Artificial Intelligence*, vol. 5, no. 1, p. 210, Aug. 2025. https://doi.org/10.1007/s44163-025-00376-9.

[20] S. Amin, M. I. Uddin, A. A. Alarood, W. K. Mashwani, A. Alzahrani, and A. O. Alzahrani, "Smart E-Learning Framework for Personalized Adaptive Learning and Sequential Path Recommendations Using Reinforcement Learning," *IEEE Access*, vol. 11, pp. 89769–89790, 2023. https://doi.org/10.1109/ACCESS.2023.3305584.

[21] E. Alyahyan and D. Düştegör, "Predicting academic success in higher education: literature review and best practices," *International Journal of Educational Technology in Higher Education*, vol. 17, no. 1, Dec. 2020. https://doi.org/10.1186/S41239-020-0177-7.

[22] K. Kesgin, S. Kiraz, S. Kosunalp, and B. Stoycheva, "Beyond Performance: Explaining and Ensuring Fairness in Student Academic Performance Prediction with Machine Learning," *Applied Sciences*, vol. 15, no. 15, p. 8409, Jul. 2025. https://doi.org/10.3390/app15158409.

[23] C. Li and Z. Cao, "Deep learning-based AI model for predicting academic success and engagement among physical higher education students," *Sci. Rep.*, vol. 15, no. 1, p. 45471, Nov. 2025. https://doi.org/10.1038/s41598-025-29000-7.

[24] Md. M. Islam, F. H. Sojib, Md. 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, Jun. 2025. https://doi.org/10.1016/j.teler.2025.100203.

[25] L. Aluso and J. O. Enyejo, "Using XGBoost and Time-Series Forecasting to Predict Student Academic Trajectories in Educational Analytics Platforms," *Int. J. Innov. Sci. Res. Technol.*, p. 143, Dec. 2025. https://doi.org/10.38124/ijisrt/25dec159.

[26] D. W. J. Anson, "Personas of plagiarism: The construction of the 'plagiarist' in Australian university subreddits," *Linguistics and Education*, vol. 69, p. 101050, Jun. 2022. https://doi.org/10.1016/j.linged.2022.101050.

[27] T. Binali, C.-C. Tsai, and H.-Y. Chang, "University students' profiles of online learning and their relation to online metacognitive regulation and internet-specific epistemic justification," *Comput. Educ.*, vol. 175, p. 104315, Dec. 2021. https://doi.org/10.1016/j.compedu.2021.104315.

[28] K. A. Bird, B. L. Castleman, Z. Mabel, and Y. Song, "Bringing Transparency to Predictive Analytics: A Systematic Comparison of Predictive Modeling Methods in Higher Education," *AERA Open*, vol. 7, Jan. 2021. https://doi.org/10.1177/23328584211037630.

[29] C. L. Huang, C. Wu, and S. C. Yang, "How students view online knowledge: Epistemic beliefs, self-regulated learning and academic misconduct," *Comput. Educ.*, vol. 200, p. 104796, Jul. 2023. https://doi.org/10.1016/j.compedu.2023.104796.