Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 47 No. 11
Nov 2026

A Hybrid Weighted Probability Ensemble of Random Forest, XGBoost and Artificial Neural Network for Student Academic Failure Prediction

Babandi Usman [1*], Abdullahi Hassan Birnin Kudu [2], Salim Ahmad [3], Zahraddeen Sufyan [4]

[1] Department of Computer Science, Federal University Dutse, Jigawa State, Nigeria
[1] Department of Cybersecurity, Federal University Dutse, Jigawa State, Nigeria
[3] Department of Information Technology, Federal University Dutse, Jigawa State, Nigeria
[4] Department of Computer Science, Federal University Dutse, Jigawa State, Nigeria
* Corresponding Author Email: Babandee@gmail.com

Abstract

Identifying students who are likely to fail before failure becomes irreversible is a central concern of higher education institutions. Conventional academic monitoring is largely retrospective, and many data-driven predictors rely on a single learning algorithm, ignore predictor redundancy, and are judged on accuracy alone. This study develops and evaluates a Hybrid Weighted Ensemble (HWE) that fuses the failure probabilities of three heterogeneous learners Random Forest (RF), Extreme Gradient Boosting (XGBoost) and an Artificial Neural Network (ANN) for binary prediction of student academic failure, and benchmarks it against each constituent learner. An integrated educational dataset of 5,000 student records and 25 variables (demographic, admission, institutional, academic-history, assessment, attendance, study-behaviour and learning-management-system engagement variables) was used; the target was failure status (2,807 non-failure, 2,193 failure). Preprocessing comprised data inspection, median imputation of eight numerical variables with fewer than 6% missing values, categorical encoding, predictor–target separation and a stratified 80:20 train/test split (4,000/1,000). Predictors were screened with Random Forest feature importance, Recursive Feature Elimination (RFE) and Mutual Information (MI). The base learners' failure probabilities were combined as PHWE = 0.40·PRF + 0.40·PXGB + 0.20·PANN with a 0.50 decision threshold, and all models were evaluated on the same held-out set using accuracy, precision, recall, F1-score, ROC-AUC and confusion matrices. Random Forest achieved 68.4% accuracy, 66.5% precision, 56.5% recall, 61.1% F1 and 75.4% ROC-AUC; XGBoost 67.4% accuracy and 74.1% ROC-AUC; and the ANN 63.1% accuracy and 67.3% ROC-AUC. The HWE obtained the best scores on every reported metric: 68.7% accuracy, 66.9% precision, 56.7% recall, 61.4% F1 and 75.8% ROC-AUC (confusion matrix: 438 true negatives, 123 false positives, 190 false negatives, 249 true positives), i.e. gains of +0.3 pp accuracy and +0.4 pp ROC-AUC over Random Forest. Previous GPA was the most influential predictor (RF importance 0.114), and the three selection methods converged on Previous GPA, Attendance Rate, Failed Courses, Entry Score and Study Hours; RFE retained ten predictors. Probability-level fusion of heterogeneous learners yields a small but consistent improvement over the strongest single model, with the largest practical value lying in the integrated, reproducible pipeline (preprocessing, multi-method feature screening, weighted fusion and multi-metric evaluation) rather than in a large accuracy gain. Recall (≈57%) indicates that a considerable share of at-risk students is still missed, so the framework should be used as decision support alongside professional academic judgement.

Keywords: Student failure prediction, educational data mining, ensemble learning, weighted probability fusion, Random Forest, XGBoost, artificial neural network, feature selection.

1. Introduction

Advances in artificial intelligence, machine learning, educational data mining (EDM) and learning analytics have changed how higher-education institutions collect and use information for academic and administrative decisions [1]. Student information systems, learning management systems (LMS), electronic examination platforms and attendance-monitoring tools generate large volumes of data on students' academic histories, assessment performance, attendance, engagement and study behaviour [2,3]. Analyzed systematically, these data can support the early identification of

students who are likely to experience academic difficulty. Academic success is a major indicator of educational quality and institutional effectiveness, whereas persistent failure leads to repeated courses, delayed graduation, withdrawal, higher institutional cost and lower retention [4]. Traditionally, failure has been identified through examination results, continuous assessment, attendance records and lecturer observation. These approaches are largely reactive: support may start only after a student has accumulated poor scores or failed courses [5]. Supervised machine learning offers a proactive alternative by learning patterns from historical records and applying them to new students, and algorithms such as Decision Trees, Support Vector Machines, Random Forest, gradient boosting, XGBoost and Artificial Neural Networks have been widely applied to performance, failure and dropout prediction [6,7].

Student failure is nevertheless a difficult learning problem because outcomes are rarely determined by one factor; they reflect interactions among previous achievement, current assessment, attendance, study effort, platform engagement, failed courses and admission characteristics [5,8]. Three limitations recur in the literature. First, many studies depend on a single algorithm, which restricts robustness and generalization [9,10]. Second, comprehensive feature-selection frameworks are used relatively rarely, leaving redundant or noisy variables in the model [11]. Third, models are often judged by accuracy only, although the two error types have different educational consequences: a false negative is an at-risk student who receives no support. Ensemble learning addresses the first limitation by combining complementary learners [12,13], but an ensemble does not automatically outperform its best member and must be tested against its own components.

This paper therefore reports the design and evaluation of a Hybrid Weighted Ensemble for student failure prediction. The contributions are: (i) a complete, reproducible pipeline from raw educational data to prediction, including systematic preprocessing and three complementary feature-screening procedures; (ii) a weighted probability-fusion ensemble of RF, XGBoost and ANN, described as a weighted soft-voting scheme rather than stacking; (iii) a like-for-like comparison of the ensemble and its three constituents on the same holdout data with five metrics and confusion-matrix analysis; and (iv) an identification of the predictors that matter most for failure in this dataset. The study is predictive rather than causal: an influential predictor is not claimed to be a cause of failure.

2. Related Work

Recent studies on student performance prediction show a move towards ensemble learning, in which Random Forest, XGBoost, LightGBM, CatBoost and hybrid ensembles are combined to improve robustness and generalization [9,10]. Almazroei [14] compared Logistic Regression, Random Forest, XGBoost, a Multi-Layer Perceptron and bagging ensembles on seven integrated educational datasets with stratified cross-validation, SMOTE balancing and five performance metrics, and found XGBoost to be the strongest single classifier with bagging ensembles competitive. Awadelkarim and Albalawi [15] reported that an explainable stacked ensemble outperformed standalone tree-based and neural models in precision, recall, F1-score and ROC-AUC. A second trend is the integration of multi-source data LMS logs, demographic, behavioral and online-activity variables instead of examination records alone [4,16].

Beyond education, ensemble strategies such as bagging, majority and soft voting, and stacking have been used to combine diverse learners [17–19]. Weighted ensembles go further by assigning each learner an importance proportional to its predictive capability, so that stronger learners influence the decision more while weaker but different learners still contribute. Random Forest contributes stability through bootstrap aggregation [12]; XGBoost contributes sequential error correction and strong performance on tabular data [20,21]; and neural networks can represent nonlinear interactions that tree-based models may miss.

The gaps motivating the present work are the reliance on single models, the limited use of combined feature selection strategies, the evaluation of models with accuracy alone, and the scarcity of studies that test an ensemble against each of its own constituents under identical conditions.

3. Materials and Methods

Research Design and Framework

The study follows a pragmatic, quantitative and experimental design: models are trained, tested and compared under a common experimental

environment, and usefulness is judged by the reliability of the predictions. The sequential process is dataset acquisition and preparation, data inspection, cleaning, missing-value treatment, categorical encoding, target separation, feature selection, model development, ensemble construction, prediction and performance evaluation. The overall framework (Fig. 1) has six layers: educational data, data preparation and preprocessing, feature selection, predictive modelling, hybrid weighted ensemble with evaluation, and an explainable-AI layer (SHAP, LIME and ALE; reported separately).

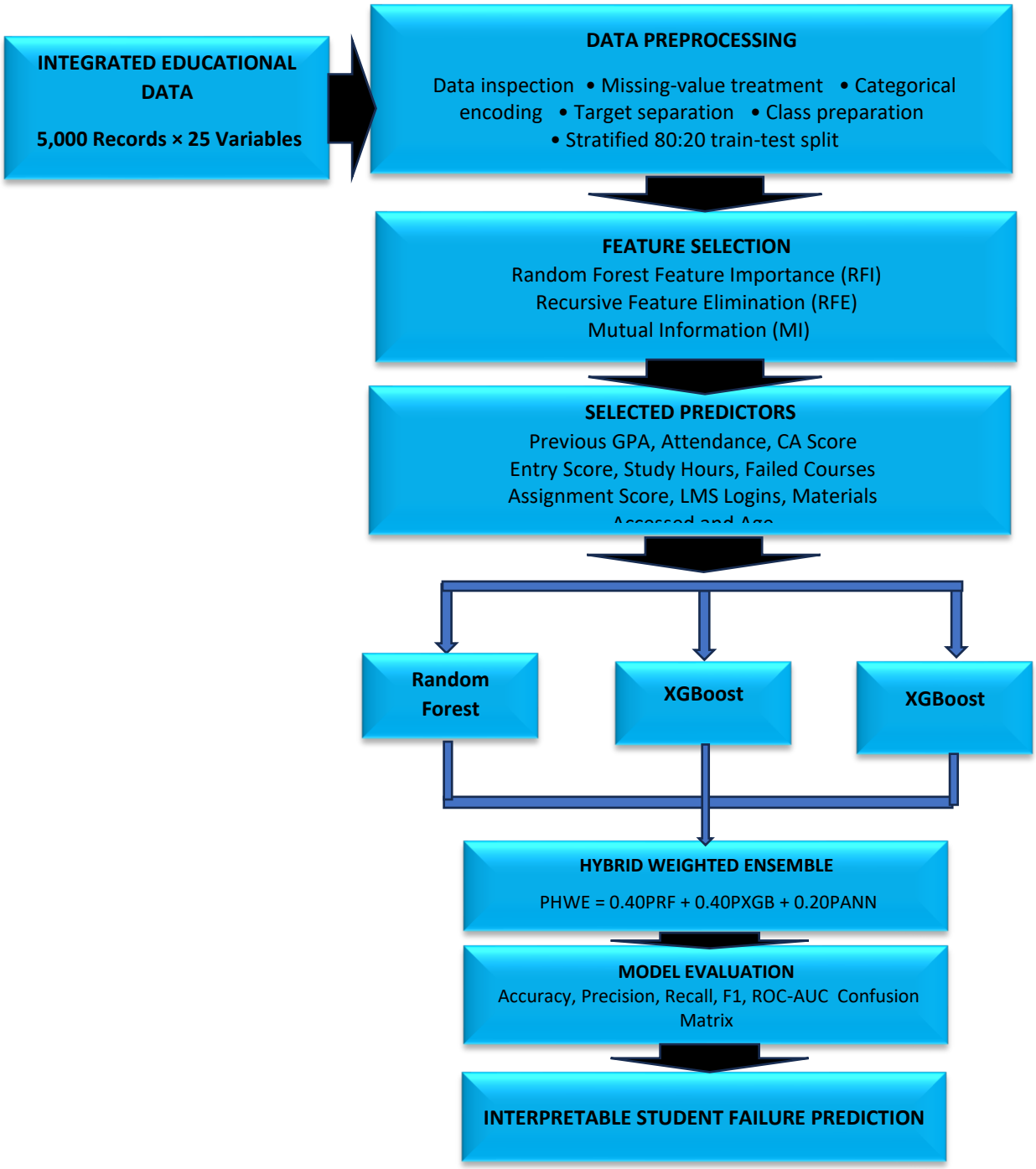

Fig. 1. Overall framework of the proposed hybrid ensemble learning approach for student failure prediction

Dataset and Target Variable

The integrated educational dataset contains 5,000 student records and 25 variables (Table 1) spanning demographic, admission, institutional, programme, academic-history, assessment, attendance, digital-learning engagement and study-behaviour information. Traditional academic variables (previous GPA, failed courses, continuous assessment,

assignment performance, entry score) are thus combined with behavioral and engagement variables (attendance, study hours, LMS logins, learning-material access, quiz attempts, assignment submissions). The dependent variable is failure status, a binary outcome $Y \in \{0, 1\}$, where 0 denotes non-failure and 1 failure; the prediction task is $X \rightarrow Y$, with X the student-level predictors. The target is separated from the predictor matrix before training to avoid leakage.

Table 1. Variables of the educational dataset

No.	Variable	Description	Role
1	student_id	Unique identification number	Identifier
2	Age	Age of student	Predictor
3	Gender	Gender category	Predictor
4	Location	Geographical location	Predictor
5	admission_type	Type of admission	Predictor
6	entry_score	Admission/entry score	Predictor
7	Institution	Institution type	Predictor
8	Department	Academic department	Predictor
9	Programme	Academic programme	Predictor
10	Level	Current academic level	Predictor
11	Semester	Current semester	Predictor
12	academic_session	Academic session	Predictor
13	Region	Regional classification	Predictor
14	previous_gpa	Previous Grade Point Average	Predictor
15	failed_courses	Number of previously failed courses	Predictor
16	credit_load	Number of registered credits	Predictor
17	assignment_score	Assignment performance	Predictor
18	ca_score	Continuous assessment score	Predictor
19	attendance_rate	Percentage attendance	Predictor
20	lms_logins	Number of LMS logins	Predictor
21	materials_accessed	Learning materials accessed	Predictor
22	quiz_attempts	Number of quiz attempts	Predictor
23	assignment_submissions	Number of assignments submitted	Predictor
24	study_hours	Estimated study hours	Predictor
25	failure_status	Student failure status	Target

Data Preprocessing

Preprocessing (Fig. 2) comprised data inspection (dimensions, variable types, missing-value structure, target distribution), missing-value treatment, categorical encoding, predictor–target separation, class-distribution assessment and a stratified training/testing partition. Missing observations occurred in numerical variables only (Age, Entry Score, Previous GPA, Assignment Score, CA Score, Attendance Rate, LMS Logins and Study Hours) and were below 6% for each variable. Instead of deleting records, median imputation was applied, because the median is resistant to the extreme values that educational variables may contain. The categorical variables (Gender, Location, Admission Type, Institution, Department, Programme, Level, Semester, Academic

Session and Region) were converted to numerical representations because the selected algorithms require numerical input.



Fig. 2. Data preprocessing flowchart

The modelling dataset contained 2,807 non-failure (56.14%) and 2,193 failure (43.86%) observations; this is far from a severely skewed problem, but class distribution was still considered in preparation and evaluation because failure cases are of primary practical interest. The data were divided 80:20 with stratified sampling, giving 4,000 training and 1,000 independent test observations with preserved class proportions.

Feature Selection

Three complementary techniques were used to reduce dependence on the assumptions of any single one: Random Forest feature importance (embedded; importance proportional to the impurity reduction attributable to a predictor), Recursive Feature Elimination (wrapper; features are repeatedly ranked and the least informative removed until the desired number remains) and Mutual Information, MI(X;Y) (filter; the statistical dependency between a predictor and failure status, able to capture nonlinear dependence). Predictors that were consistently supported by the three methods formed the basis for model development.

Base Learners

Three algorithms with different learning mechanisms were developed as base learners (Table 2) and trained on the same prepared training set and evaluated on the same test set. Random Forest aggregates decision trees trained on bootstrap samples and random predictor subsets and can model nonlinear relationships and interactions (for instance between attendance, previous GPA, study hours and failed courses) without explicit specification. XGBoost builds trees sequentially, each correcting the errors of its predecessors. The ANN has an input layer receiving the selected predictors, hidden layers that learn nonlinear relationships through $z = \Sigma(wi\ xi) + b$ followed by an activation function, and an output layer producing the failure probability.

Table 2. Base learning algorithms

Model	Learning paradigm	Principal strength	Role in framework
Random Forest	Bagging / tree ensemble	Robustness and nonlinear modelling	Base learner + feature importance
XGBoost	Gradient boosting	Sequential error correction; strong on tabular data	Base learner
ANN	Neural learning	Nonlinear representation and interaction learning	Base learner

Hybrid Weighted Ensemble

Each base learner outputs a failure-class probability: P_RF, P_XGB and P_ANN. The ensemble probability is their weighted sum

$P_{HWE} = 0.40\ P_{RF} + 0.40\ P_{XGB} + 0.20\ P_{ANN}$, with $0.40 + 0.40 + 0.20 = 1.00$ (1)

(Table 3). Random Forest and XGBoost perform strongly and are given the larger weights, while the ANN, which performed comparatively lower under the experimental conditions, is retained at 0.20 as a complementary nonlinear learner. This differs from equal-weight soft voting, P = (P_RF + P_XGB + P_ANN)/3, in which every model contributes equally. Because the base-model outputs are fused at the

probability level with fixed weights and no meta-learner is trained, the scheme is a hybrid weighted probability ensemble, not stacking. A common threshold of 0.50 is applied:

$\hat{Y} = 1$ if $P_{HWE} \geq 0.50$; $\hat{Y} = 0$ if $P_{HWE} < 0.50$ (2)

Table 3. Ensemble weight allocation

Base learner	Weight	Contribution
Random Forest	0.40	40%
XGBoost	0.40	40%
ANN	0.20	20%
Total	**1.00**	**100%**

Evaluation

Models were compared on the same held-out data with accuracy, precision, recall, F1-score, ROC-AUC and the confusion matrix, since accuracy alone does not describe a binary classifier whose two error types have different educational implications. With TP, TN, FP and FN denoting true/false positives/negatives (failure is the positive class):

Accuracy = (TP + TN)/(TP + TN + FP + FN); Precision = TP/(TP + FP) (3)

Recall = TP/(TP + FN); F1 = 2 × (Precision × Recall)/(Precision + Recall) (4)

Recall is particularly important here because a false negative is a genuinely at-risk student the system fails to identify; precision reflects how few students are wrongly flagged; ROC-AUC measures discrimination across all thresholds (≈1.0 strong, ≈0.50 random).

Experimental Procedure and Implementation

The procedure (dataset loading → inspection → missing-value analysis and median imputation → categorical encoding → target separation → class preparation → feature selection → stratified 80:20 split → RF, XGBoost and ANN training → probability prediction (P_RF, P_XGB, P_ANN) → weighted fusion (P_HWE) → classification → evaluation and confusion-matrix analysis) was implemented in Python on Google Collaboratory using Pandas and NumPy for data handling, Scikit-learn for preprocessing, Random Forest, RFE, the ANN and evaluation, XGBoost for gradient boosting, and Matplotlib for visualization. Reliability is supported by using identical data and test conditions for all models, defining and separating the target before training, reporting multiple metrics, using several feature-selection methods and reserving an independent test set.

4. Results and Discussion

Dataset Overview and Descriptive Statistics

Table 4: Dataset Characteristics

Attribute	Value
Total records	5,000
Total variables	25 (14 numerical, 11 categorical)
Target variable	Failure status
Non-failure cases	2,807 (56.14%)
Failure cases	2,193 (43.86%)

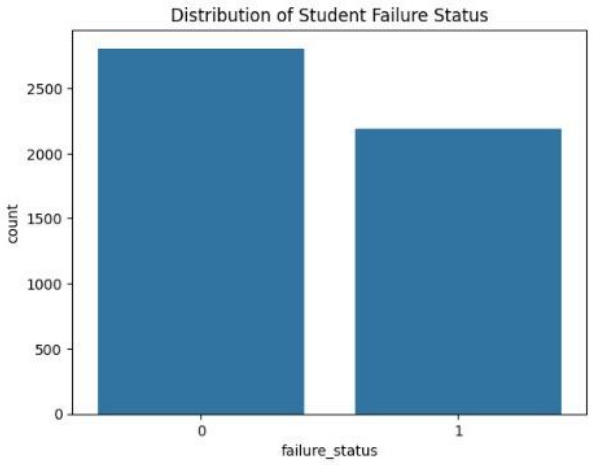

Fig. 3. Distribution of failure and non-failure cases (0 = non-failure, 1 = failure)

The two classes are reasonably balanced and failure accounts for a substantial share (43.86%) of outcomes, which reinforces the need for tools that support early intervention; stratified partitioning kept the class proportions identical in the training and test sets. Descriptive statistics (Table 5) show substantial variation in prior achievement, attendance, study behaviour and digital engagement: mean Previous GPA was 2.91 (minimum 0.64), students had on average 1.70 failed courses (range 0–7), mean attendance was 73.50% (35– 100%), study hours ranged from 0 to 18.51 and LMS logins from 5 to 69. This variation allows the algorithms to learn meaningful relationships with failure status.

Table 5. Descriptive statistics of the key numerical variables

Variable	Mean	Std. dev.	Min.	Max.
Age	26.00	5.49	17	35
Entry Score	220.90	34.90	120	320
Previous GPA	2.91	0.67	0.64	5.00
Failed Courses	1.70	1.27	0	7
Credit Load	22.54	5.12	15	30
Assignment Score	66.25	11.69	28.14	100
CA Score	63.94	12.98	15	100
Attendance Rate	73.50	11.06	35	100
LMS Logins	36.57	9.20	5	69
Materials Accessed	34.19	10.97	0	72
Quiz Attempts	7.45	3.24	0	19
Assignment Submissions	4.95	2.13	0	13
Study Hours	8.36	2.82	0	18.51

Preprocessing Results

Missing values were found in eight numerical variables, each below 6% (Table 6, Fig. 3). Median imputation produced a complete numerical feature matrix while retaining all 5,000 observations; the categorical variables were encoded and the data were split 80:20 (4,000 training, 1,000 test).

Table 6. Missing-value distribution and treatment

Variable	Missing values	Approx. %	Treatment
Age	298	5.96%	Median imputation
Entry Score	297	5.94%	Median imputation
Previous GPA	296	5.92%	Median imputation
Assignment Score	293	5.86%	Median imputation
CA Score	296	5.92%	Median imputation
Attendance Rate	296	5.92%	Median imputation
LMS Logins	297	5.94%	Median imputation
Study Hours	297	5.94%	Median imputation

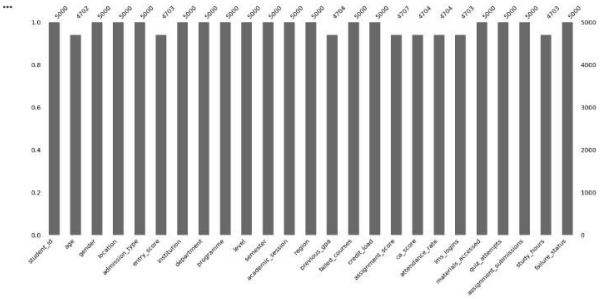

Fig. 4. Non-null record counts per variable before imputation (out of 5,000 records)

Predictors Used for Modelling

The three feature-selection procedures produced a consistent picture. Random Forest importance ranked Previous GPA first (0.114), followed by Attendance Rate, CA Score, Entry Score, Study Hours, Failed

Courses, Assignment Score, LMS Logins and Materials Accessed. RFE selected ten predictors (Table 7), and Mutual Information found the strongest dependency with failure status for Previous GPA, Failed Courses, Attendance Rate, Entry Score and Study Hours. The selected set covers academic history (Previous GPA, Failed Courses), current performance (CA Score, Assignment Score), admission background (Entry Score), engagement (Attendance Rate, Study Hours), digital learning behaviour (LMS Logins, Materials Accessed) and demography (Age). This multidimensional structure supports treating failure as

a multifactorial problem and using a heterogeneous ensemble.

Table 7. Predictors selected by Recursive Feature Elimination

Rank	Selected predictor	Rank	Selected predictor
1	Previous GPA	6	Failed Courses
2	Attendance Rate	7	Assignment Score
3	CA Score	8	LMS Logins
4	Entry Score	9	Materials Accessed
5	Study Hours	10	Age

Performance of the Base Learners

Random Forest: On the 1,000 test observations the Random Forest achieved 68.4% accuracy, 66.5% precision, 56.5% recall, 61.1% F1-score and 75.4% ROC-AUC (Table 8). Its confusion matrix (Table 9, Fig. 4) shows 436 non-failure and 248 failure cases classified correctly (684 of 1,000), with 125 false positives and 191 false negatives. The ROC-AUC of 75.4% indicates good discrimination, but the recall of 56.5% means that a considerable proportion of failing students is not identified, which matters for intervention.

Table 8. Random Forest performance on the test set

Metric	Result
Accuracy	68.4%
Precision	66.5%
Recall	56.5%
F1-score	61.1%
ROC-AUC	75.4%

Table 9. Confusion matrix of the Random Forest model

Actual / Predicted	Non-failure	Failure
Non-failure	436	125
Failure	191	248

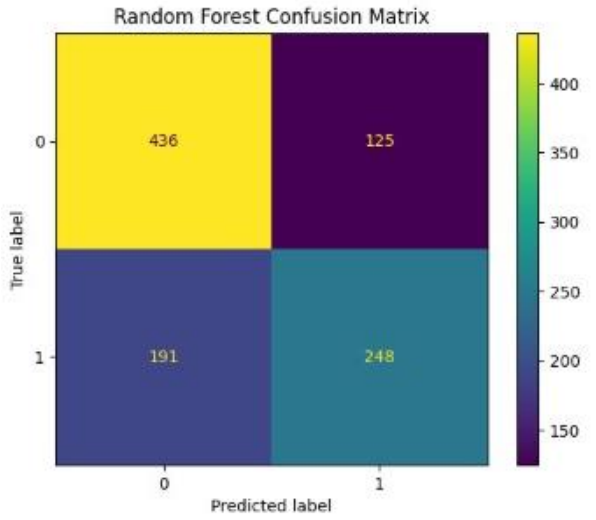

Fig. 5. Confusion matrix of the Random Forest model (0 = non-failure, 1 = failure)

XGBoost: XGBoost reached 67.4% accuracy and 74.1% ROC-AUC, slightly below Random Forest but competitive. Because its learning mechanism (sequential boosting) differs from Random Forest, its predictions may carry complementary information, which justifies retaining it in the ensemble.

ANN: The ANN gave the lowest standalone performance, 63.1% accuracy and 67.3% ROC-AUC; a convergence warning was reported during training, a plausible contributor to the weaker result. Performance of neural networks depends on dataset size, feature representation, architecture and optimization, so this result does not imply that neural networks are unsuitable for educational prediction. The ANN was kept because an ensemble does not require every member to be the strongest model; its contribution is diversity from a different learning paradigm.

Performance of the Hybrid Weighted Ensemble

Fusing the three probability outputs with the 0.40/0.40/0.20 weights gave 68.7% accuracy, 66.9% precision, 56.7% recall, 61.4% F1-score and 75.8% ROC-AUC (Table 10). The confusion matrix (Table 11, Fig. 5) shows 438 true negatives, 123 false positives, 190 false negatives and 249 true positives; (438 + 249)/1,000 = 68.7% of observations were classified correctly, and one more failure case was identified (249 versus 248 for Random Forest).

Table 10. Performance of the Hybrid Weighted Ensemble

Metric	Result
Accuracy	68.7%
Precision	66.9%
Recall	56.7%
F1-score	61.4%
ROC-AUC	75.8%

Table 11. Confusion matrix of the Hybrid Weighted Ensemble

Actual / Predicted	Non-failure	Failure
Non-failure	438	123
Failure	190	249

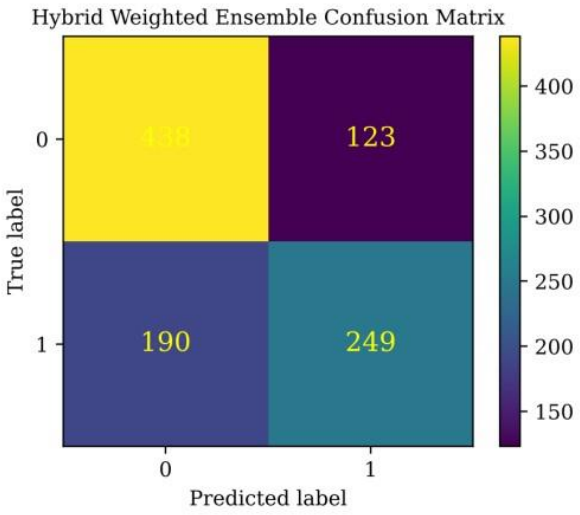

Fig. 6. Confusion matrix of the Hybrid Weighted Ensemble on the 1,000-record test set (0 = non-failure, 1 = failure)

Comparative Performance

Table 12. Comparative performance of all models on the test set

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Random Forest	68.4%	66.5%	56.5%	61.1%	75.4%
XGBoost	67.4%	n/r	n/r	n/r	74.1%
ANN	63.1%	n/r	n/r	n/r	67.3%
Hybrid Weighted Ensemble	68.7%	66.9%	56.7%	61.4%	75.8%

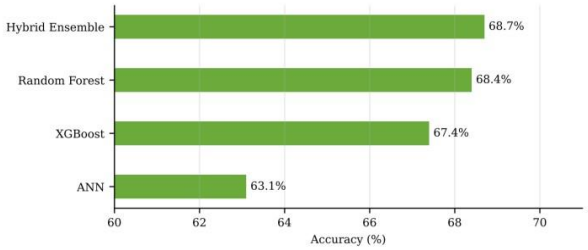

Fig. 7. Comparative accuracy of the models (x-axis truncated at 60%)

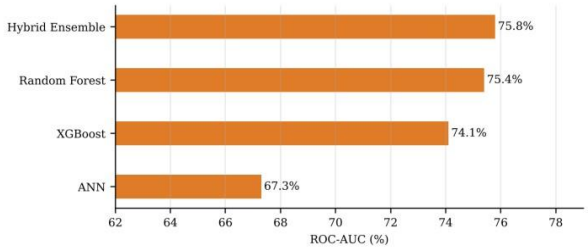

Fig. 8. Comparative ROC-AUC of the models (x-axis truncated at 62%)

Table 13. Random Forest versus Hybrid Weighted Ensemble

Metric	Random Forest	Hybrid Ensemble	Improvement
Accuracy	68.4%	68.7%	+0.3 pp
Precision	66.5%	66.9%	+0.4 pp
Recall	56.5%	56.7%	+0.2 pp
F1-score	61.1%	61.4%	+0.3 pp
ROC-AUC	75.4%	75.8%	+0.4 pp

5. Discussion

Four findings emerge. First, the Hybrid Weighted Ensemble achieved the best overall performance, 68.7% accuracy and 75.8% ROC-AUC. Second, Random Forest remained extremely competitive (68.4%, 75.4%). Third, XGBoost was slightly below Random Forest, and fourth, the ANN was lowest. The improvement of the ensemble over Random Forest is small +0.3 pp accuracy and +0.4 pp ROC-AUC but it is consistent across all five metrics. It can be attributed conceptually to the different learning mechanisms of the constituents: Random Forest captures nonlinear relationships through aggregated trees, XGBoost through sequential gradient boosting and the ANN through a neural nonlinear representation, and the weighted fusion allows the final decision to draw on all three while giving more influence to the stronger tree-based learners. The result should not be read as showing that combining models automatically guarantees superior performance; it shows that, under this configuration, probability-level integration of heterogeneous learners gave a modest advantage over the strongest single classifier.

The contribution of the work therefore lies in the combination of heterogeneous learners, multi-method feature screening, weighted probability fusion and multi-metric evaluation, rather than in raw accuracy. The predictors that matter Previous GPA, Attendance Rate, Failed Courses, Entry Score, Study Hours, assessment scores and LMS engagement indicate that an early-warning system should consider academic history, current assessment, physical participation and digital engagement together, and that continuous-assessment data can flag difficulty before final examinations. Practically, with a recall of about 57%, 190 of the 439 failing students in the test set were still not flagged by the ensemble, so predictions are decision-support information and not automatic decisions about students.

6. Conclusion

A Hybrid Weighted Ensemble that fuses Random Forest, XGBoost and ANN probabilities (0.40/0.40/0.20) was developed and evaluated for student failure prediction on 5,000 records and 25 variables. After median imputation, encoding, stratified 80:20 partitioning and three-method feature screening, the ensemble obtained the highest accuracy (68.7%), precision (66.9%), recall (56.7%), F1-score (61.4%) and ROC-AUC (75.8%), improving modestly on Random Forest (68.4% accuracy, 75.4% ROC-AUC), with XGBoost (67.4%, 74.1%) and the ANN (63.1%, 67.3%) behind. Previous GPA, Attendance Rate, Failed Courses, Entry Score and Study Hours were the most consistently important predictors.

References

[1] J, K. S. (2024). The role of learning analytics in improving student outcomes. 31–35.

[2] Nonso Sunday, O., & Timothy Umeifekwem, U. (2023). Addressing digital technology gap challenges: The Nigerian experience. *Nigerian Journal of Social Development*, 11(1), 95–100.

[3] Simon, P. D., Jiang, J., Fryer, L. K., King, R. B., & Frondozo, C. E. (2025). An assessment of learning management system use in higher education: Perspectives from a comprehensive sample of teachers and students. *Technology, Knowledge and Learning*, 30(2), 741–767. https://doi.org/10.1007/s10758-024-09734-5

[4] Hetigan, J., Abrenica, M. C., & Lolong, L. (2025). Leveraging predictive analytics on LMS logs to examine the impact of engagement on academic performance among college students enrolled in Centro Escolar University. *International Journal of Research and Scientific Innovation*, 12(2321), 563–573. https://doi.org/10.51244/IJRSI

[5] Bello, A., Ajibade, I., Wonuola, I., & Ekweli, D. (2025). Developing a predictive model for student academic performance using machine learning techniques.

[6] Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021). Educational data mining techniques for student performance prediction: Method review and comparison analysis. *Frontiers in Psychology*, 12, 698490. https://doi.org/10.3389/fpsyg.2021.698490

[7] Lou, Y., & Colvin, K. (2025). Performance prediction using educational data mining techniques: a comparative study. *Discover Education*, 4. https://doi.org/10.1007/s44217-025-00502-w

[8] Edmond, V., Sada, S., & Osijirin, A. (2025). Predictive analytics using machine learning models on undergraduate students' performance of the Federal University of Allied Health Sciences, Enugu, Nigeria in Introduction to Computing Science. *Saudi Journal of Engineering and*

Technology, 10. https://doi.org/10.36348/sjet.2025.v10i07.004

[9] Rehman, M., Iftikhar, A., Muhammad, S., & Ahmed, R. (2025). Student academic performance prediction using ensemble learning methods. *Journal of ICT, Design, Engineering, and Technological Science*, 9. https://doi.org/10.33150/JITDETS-9.1.2

[10] Alzahrani, M. R. (2024). Predicting student performance using ensemble models and learning analytics techniques. Preprints. https://doi.org/10.20944/preprints202406.1100.v1

[11] Kalita, E., Sunday, S., Gaftandzhieva, S., et al. (2025). Educational data mining: a 10-year review. *Discover Computing*. https://doi.org/10.1007/s10791-025-09589-z

[12] Ibrahim, M. (2023). Evolution of Random Forest from decision tree and bagging: A bias-variance perspective. *Dhaka University Journal of Applied Science and Engineering*, 7(1). https://doi.org/10.3329/dujase.v7i1.62888

[13] Ngo, G., Beard, R., & Chandra, R. (2022). Evolutionary bagging for ensemble learning. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2022.08.055

[14] Almazroei, E. E. (2026). Ensemble machine learning framework with SHAP and LIME for accurate early prediction of student success in online learning environments. *Scientific Reports*.

[15] Awadelkarim, A. M., & Albalawi, H. (2026). An explainable hybrid machine learning model for data-driven assessment and enhancement of program learning outcomes in higher education. *International Journal of Innovative Research and Scientific Studies*, 9(3), 84–105. https://doi.org/10.53894/ijirss.v9i3.11393

[16] Wang, Y. (2025). Artificial intelligence in student management systems to enhance academic performance monitoring and intervention. *Scientific Reports*, 15(1), 35122. https://doi.org/10.1038/s41598-025-19159-4

[17] Azad, M., Hasan, T., & Mikhail, N. (2025). A novel ensemble learning method using majority based voting of multiple selective decision trees. *Computing*, 107(1). https://doi.org/10.1007/s00607-024-01394-8

[18] Kokate, S., & Chetty, M. S. R. (2021). Credit risk assessment of loan defaulters in commercial banks using voting classifier ensemble learner machine learning model. *International Journal of Safety and Security Engineering*, 11(5), 565–572. https://doi.org/10.18280/IJSSE.110508

[19] Kalabarige, L. R., Rao, R. S., Abraham, A., & Gabralla, L. A. (2022). Multilayer stacked ensemble learning model to detect phishing websites. *IEEE Access*, 10, 79543–79552. https://doi.org/10.1109/ACCESS.2022.3194672

[20] Ali, Z. A., Abduljabbar, Z. H., Tahir, H. A., Sallow, A. B., & Almufti, S. M. (2023). Exploring the power of eXtreme Gradient Boosting algorithm in machine learning: a review. 12(2).

[21] Id, M. N., Hossain, A., Id, C., Arefeen, F., & Id, M. (2023). A tree based eXtreme Gradient Boosting (XGBoost) machine learning model to forecast the annual rice production in Bangladesh. *PLOS ONE*. https://doi.org/10.1371/journal.pone.0283452