Journal of Computer Science

Research Article

Hybrid Soft Voting Ensemble of XGBoost and DNN for At-Risk Student Performance Prediction

Eugene Wan[1], Po Chan Chiu[1], Mohammad bin Hossin[1], Hamizan Sharbini[1], King Kuok Kuok[2] Noor Hazlini Borhan[1] and Chih How Bong[1]

[1]Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Malaysia
[2]Faculty of Engineering, Computing and Science, Swinburne University of Technology Sarawak Campus, 93350 Kuching, Malaysia

Article history
Received: 13-01-2026
Revised: 29-03-2026
Accepted: 13-04-2026

Corresponding Author:
Po Chan Chiu
Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Malaysia
Email: pcchiu@unimas.my

Abstract: Early identification of at-risk students in higher education is important for timely academic intervention, yet conventional prediction methods often struggle with data imbalance and limited model precision. This study proposes a hybrid soft voting ensemble model that integrates Extreme Gradient Boosting (XGBoost) and Deep Neural Network (DNN) to enhance multi-class student grade prediction (A-F classification) and at-risk student identification. This proposed approach is evaluated using two datasets: a publicly available Kaggle Student Performance Dataset and a real-world dataset collected from a Database Concept and Design course at Universiti Malaysia Sarawak (UNIMAS). Both datasets undergo comprehensive pre-processing, including class imbalance handling using SMOTE and feature normalization using StandardScaler. Comparative evaluations were conducted against baseline models, including KNN, SVM, XGBoost and DNN, with all models optimised via hyperparameter tuning. Experimental results demonstrate that the proposed hybrid ensemble model outperforms the baseline models, achieving an accuracy of 77.37% and a macro F1-score of 74.50% on Dataset 1, and an accuracy of 74.13% with a macro F1-score of 81.53% on Dataset 2. The ensemble specifically demonstrates better sensitivity in detecting minority "at-risk" categories (Grades F and D). This study highlights the effectiveness of hybrid ensemble learning in improving predictive performance and supporting data-driven educational decision-making for early intervention in higher education.

Keywords: At-Risk Student Performance Prediction, Machine Learning, Predictive Analytics, Hybrid Soft Voting Ensemble

Introduction

Student academic performance is a central concern for Higher Education Institutions (HEIs), serving as a critical metric in evaluating institutional success and supporting student progression. A critical factor influencing the success of these institutions is the academic performance of students, where poor performance or high dropout rates can pose significant challenges (Quinn and Gray, 2019). In Malaysia, attrition among undergraduates is particularly high during their first semester, study shows that 14% of first-year students in Malaysian private universities leave within their initial semester, commonly due to academic pressure and poor program alignment (Sangodiah et al., 2015). These challenges emphasize the need for proactive mechanisms that can identify at-risk students before failure occurs. In this study, “at-risk” defined as students projected to achieve a final course grade of D and F. Early and accurate identification of this cohort allows solid intervention scenarios, such as triggering automated alerts to academic advisors, scheduling mandatory tutoring sessions, or providing personalized academic counselling.

However, the existence of imbalanced data sets is one of the challenges for using machine learning to predict student performance. In many cases, most students perform well, while a smaller subset may struggle. This imbalance skews the predictions, leading models to underperform in identifying students at risk of failure (Bujang et al., 2023). Addressing this issue through oversampling techniques will

enhance the reliability and accuracy of the performance predictions, making them more useful in real-world educational environments (Vaishnavi et al., 2023).

Traditional approaches to student performance prediction rely on limited data types such as final exam scores or demographic information. These methods often fail to adapt to dynamic learning contexts or provide timely feedback to both students and educators. In response, Machine Learning (ML) and Educational Data Mining (EDM) techniques have emerged as powerful tools to predict student academic performance, as presented in Table 1.

Table 1: Related existing work comparison

Author	Data Source	Algorithms Used	Best algorithm	Evaluation Metrics	Student Performance Prediction
(Farissi et al., 2020)	Kaggle, Kalboard 360 (Open-Source Dataset)	ANN, DT, RF, Bagging, Voting, Boosting	GA with RF (F1-Score 81.18%)	G-Mean, F1-Score, AUC, Precision, TPR, TNR	-High, Medium, Low
(Pujianto et al., 2020)	Kaggle, Kalboard 360 (Open-Source Dataset)	C4.5, KNN	DT (Accuracy 71.09%)	Accuracy, Precision, Recall	-High, Medium, Low
(Sarker et al., 2024)	Randomly generated synthetic dataset	DT, KNN, NB, NN, RF	RF with Gini Index (Accuracy 96.45%)	Accuracy, Precision, Recall, F1-score (F-Measure), Cohen's Kappa	- Good (60%- 100%), Average (50%-59%), and Poor (0%-49%)
(Ayienda et al., 2021)	Kaggle, Portuguese Secondary Education Student Performance Dataset (Open-Source Dataset)	SVM, MLP, LR, KNN, NB, WVC	WVC (Accuracy 97.6%)	Accuracy, Precision, Recall, F1-Score, AUC	Excellent, Good, Fair
(Lim et al., 2019)	Kaggle, Portuguese Secondary Education Student Performance Dataset (Open-Source Dataset)	C4.5, NB, NBT, LibSVM	LibSVM and C4.5 for Portuguese Dataset (Accuracy 92.9%), C4.5 for Mathematics Dataset (Accuracy 91.4%)	Accuracy	Binary: Fail, Pass classification
(Bujang et al., 2021)	Malaysia Polytechnic First Semester Course Grades Dataset (2016-2019) (Real Dataset)	J48, SVM, NB, kNN, LR, RF	RF + SMOTE (Accuracy 99.5%)	Accuracy, Precision, Recall, F-Measure	Exceptional (A+), Excellent (A), Distinction (A-, B+, B), Pass (B-, C+, C, C-, D+, D) and Fail (E, E-, F)
(Yan, 2019)	Student performance data (1986 to 2019) (Real Dataset)	XGBoost, DT, RF, SVM, LR	XGBoost (R^2 score 0.993)	R^2, MAE, RMSE	- Predicting continuous student scores (regression-based approach)
(Hakkal and Lahcen, 2024)	8 datasets: ASSISTments Intelligent Tutoring System dataset, Knowledge Discovery and Data Mining, KDD Cup Challenge 2010 dataset, Statics dataset and Moodle-morocco dataset	XGBoost, Item Response Theory, Performance Factor Analysis (PFA), DAS3H	XGBoost-enhanced PFA (AUC ≈ 0.88)	AUC, accuracy	- Binary: Wrong, correct classification
(Nabil et al., 2021)	Undergraduate Academic Performance Dataset (2006-2020) (Real Dataset)	DNN, DT, LR, SVC, KNN, RF, GB	DNN + SMOTE (Accuracy 89%)	Accuracy, Precision, Recall, F1-score, Time, Classification error	-Excellent (85%-100%), Very Good (75%-84%), Good (65%-74%), Poor (50%-64%) and Fail (<50%)
(Wen and Juan, 2023)	OULA dataset (Open-Source Dataset)	DNN, Autoencoder, FNN	DNN (Autoencoder + FNN) (Accuracy > 80%)	Accuracy, Precision, Recall, F1-score	- Distinction, Pass, Fail, Withdraw
(Adil et al., 2023)	UCI, Student Performance Dataset (Open-Source Dataset)	DT, RF, LR, KNN, XGB, DNN	DNN (R-squared 99.97%)	R^2-score, MAE, MSE, RMSE	-Predicting continuous student marks (regression-based approach)

The existing research has explored a range of ML algorithms for student performance prediction, including k-Nearest Neighbours (KNN: Pujianto et al., 2020; Sarker et al., 2024), Support Vector Machines (SVM: Ayienda et al., 2021), Extreme Gradient Boosting (XGBoost: Yan, 2019; Hakkal and Lahcen, 2024), and Deep Neural Networks (DNN: Nabil et al., 2021; Wen and Juan, 2023; Adil et al., 2023). However, many existing models are either binary classification (Lim et al., 2019; Hakkal and Lahcen, 2024) or general performance prediction using high, medium and low categories (Farissi et al., 2020; Pujianto et al., 2020; Sarker et al., 2024; Ayienda et al., 2021), especially in cases where low-performing student grades (e.g., D or F) are underrepresented. While these models have demonstrated different levels of success, they often struggle to generalize across diverse student populations or suffer performance degradation on minority class labels, particularly those representing at-risk students. In this paper, a hybrid soft voting ensemble that integrates XGBoost and DNN is proposed to improve early identification of at-risk students, a strategy that has not been discussed comprehensively enough in previous educational research. This hybrid ensemble model utilizes an optimized weighting ratio to prioritize XGBoost's proficiency in structured data while utilizing DNN model to capture non-linear behavioural patterns. The hybrid model is evaluated on two datasets which are an open source Kaggle Student Performance Dataset and a dataset collected from the Database Concept and Design course at Universiti Malaysia Sarawak (UNIMAS). Both datasets undergo pre-processing steps including Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance and improve the macro F1-score. The overall modelling follows the student performance predictive framework. The main contributions of this paper include:

- Hybrid soft voting ensemble model that integrates XGBoost and DNN for multi-class student grade prediction (A–F) with enhanced performance across all classes
- Student performance prediction framework that enables early identification of at-risk students to support timely and targeted academic interventions
- Comprehensive evaluation of model performance using various metrics (Accuracy, Precision, Recall, F1-score)

Related Work

Numerous studies have investigated different machine learning models for predicting student performance. Farissi et al. (2020) incorporated Genetic Algorithm (GA)-based feature optimization to address high-dimensional data, where the GA-RF model achieved the best performance with an F1 score of 81.18% compared to Decision Tree (DT), Artificial Neural Network (ANN), Random Forest (RF), Voting, Bagging, and Boosting. The study highlights the effectiveness of GA-based feature selection in improving prediction accuracy for student performance at high, medium and low levels. However, there is concern that the dataset will be affected by imbalanced class distribution, which can bias traditional classifiers towards the majority class and reduce the reliability of predicting minority outcomes.

Pujianto et al. (2020) utilised the C4.5 Decision Tree and k-Nearest Neighbour (KNN) algorithms to predict student performance into high, medium, and low categories. To address class imbalance, the study applied SMOTE, generating synthetic instances for minority classes. Performance was evaluated using 10-fold cross-validation and metrics such as accuracy, recall, and precision. The C4.5 Decision Tree outperformed KNN, achieving an accuracy of 74.09% compared to KNN's 69.68%. This research demonstrates the effectiveness of the C4.5 algorithm for predicting student performance and highlights the importance of pre-processing techniques like SMOTE to mitigate the challenges of imbalanced datasets. However, the use of broad performance categories limits the ability to detect critically low-performing learners and may bias predictions toward majority groups.

Furthermore, Sarker et al. (2024) analysed student academic performance using Random Forest with Gini Index (RF-GI). The RF-GI achieved the highest accuracy of 96.45% compared to baseline models. This research highlights how well RF detects consistent performers and acknowledges the absence of students' socioeconomic and demographic features as a limitation. In addition, Ayienda et al. (2021) proposed a hybrid model based on a Weighted Voting Classifier (WVC) to predict student performance in three categories: Fair, good, and excellent. The study integrates five machine learning algorithms, including SVM, Multi-Layer Perceptron (MLP), Logistic Regression (LR), KNN, and Naïve Bayes (NB), with a weighted voting classifier and 10-fold cross-validation, achieving an accuracy of 97.6%. Despite strong results, Ayienda et al. (2021) acknowledge that a major challenge lies in identifying the most appropriate features and algorithms. This indicates that the accuracy may be influenced by feature selection methods and algorithm setups, which could limit the effectiveness of the approach across various datasets.

Lim et al. (2019) explored the use of dimensionality reduction to enhance student performance prediction, particularly focusing on identifying students likely to fail. The study utilised datasets from Portuguese secondary schools and UCI Repository. The datasets included academic, demographic, social, and school-related attributes, modelled as binary classifications for "Pass" and "Fail" grades. The proposed C4.5 Decision Tree demonstrated the highest overall accuracy of 92.9% for the Portuguese language and 91.4% for mathematics surpassing the baseline models. The study reveals that dimensionality reduction such as wrapper improved true positive rates for the "Fail" grade

from 70.0% to 76.0% in the Portuguese language and from 86.2% to 88.5% in mathematics. These findings underscore the value of dimensionality reduction in improving minority class prediction accuracy in imbalanced educational datasets. However, this study lacks fine-grained multi-class grade prediction across A–F categories.

Research conducted by Bujang et al. (2021) proposed multi-class machine learning model to predict student grades by categorizing performance into five levels, ranging from exceptional (A+) to fail. The study achieved the highest f-measure of 99.5% when the proposed SFS (SMOTE + Feature Selection) model combined with RF. The study reveals that the proposed model can improve prediction performance in imbalanced multi-class classification for student grade prediction.

In the study by Yan (2019), the researcher applied XGBoost to predict continuous student performance scores, rather than performing classification tasks such as A–F grade prediction. The study demonstrates that XGBoost outperformed other models including RF, SVM, DT, Lasso and Elastic Net. However, the study does not address the identification of at-risk students, which limits the model's capability to provide early warning alerts. Hakkal and Lahcen (2024) further compared logistic regression-based models such as Item Response Theory (IRT), Performance Factor Analysis (PFA), and DAS3H with XGBoost. The proposed XGBoost-enhanced PFA achieved superior performance, with AUC values reaching up to 0.88. However, the study is limited to binary classification, which underperforms in identifying at-risk students.

Furthermore, various studies have utilised DNN (Nabil et al., 2021; Wen and Juan, 2023; Adil et al., 2023) to predict student performance. Nabil et al. (2021) aimed to evaluate the effectiveness of deep learning models in predicting student performance, particularly in challenging courses such as Data Structures, where failure rates are high. Using resampling techniques such as SMOTE, ADASYN, and SMOTE-ENN to balance the dataset, the proposed DNN achieved an accuracy of 89%, outperforming other algorithms such as DT, LR, KNN, RF, Support Vector Classifiers (SVC) and Gradient Boosting (GB). Additionally, Wen and Juan (2023) performed multi-class student performance prediction using four categories (Distinction, Pass, Fail, Withdraw). The proposed DNN (Autoencoder + FNN) outperformed other models in predictive accuracy. Another study by Adil et al. (2023) employed DNN to improve academic outcomes. The DNN shows the best performance compared to others, achieving an R-squared of 99.97% and the lowest error rates (MAE = 0.45, MSE = 0.05). However, this study places insufficient emphasis on minority-class performance, which may reduce the reliability of detecting low-performing students.

Despite its potential, DNNs often struggle with the uninformative features found in tabular data (c et al., 2022). Katuwal and Yang (2026) proposed hybrid framework combining XGBoost and DNN to overcome the weaknesses of standalone models for efficient prediction in early-stage disease detection. However, the role of hybrid XGBoost and DNN approaches in handling imbalanced data for at-risk student identification remains largely unexplored.

Methods

Student Performance Prediction Framework

This section describes the methodology adopted in this study for developing and evaluating a hybrid ensemble learning model. It comprises five phases including Data Selection, Data Pre-processing, Data Transformation, Model Development, Training and Tuning, and Model Evaluation. Fig. 1 shows the proposed hybrid soft voting ensemble framework for student performance prediction and at-risk student identification.

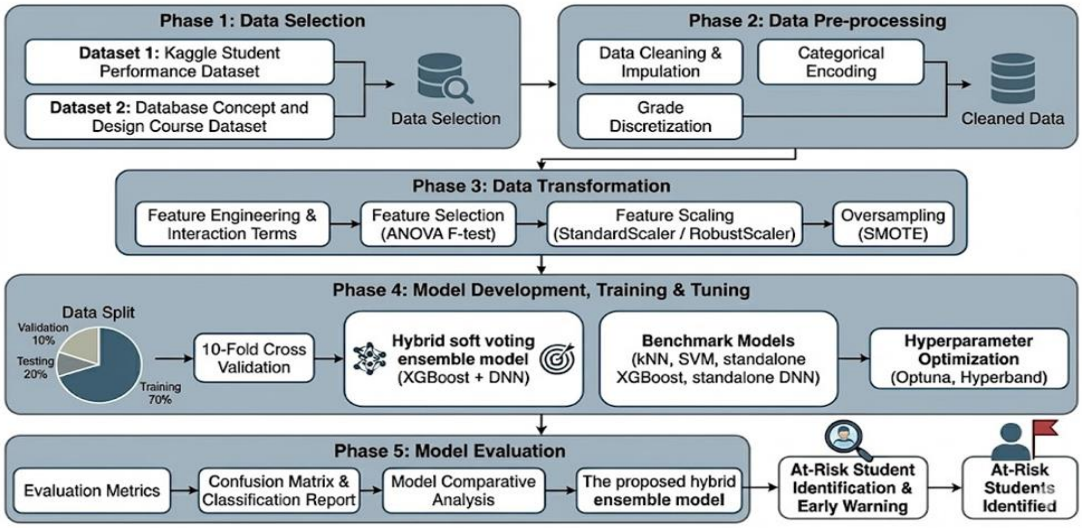


Fig. 1: The proposed hybrid soft voting ensemble framework for student performance prediction and at-risk student identification

Phase 1: Data Selection

The data selection phase utilized two datasets for predicting student performance. The Kaggle Student Performance Dataset, sourced from Kaggle (Myrick, 2024), merges two CSV files (student-por.csv, student-mat.csv) into 1,044 instances with 33 attributes, including demographic (e.g., age, sex), socioeconomic (e.g., parental education), academic (e.g., G1, G2), and behavioural (e.g., absences) features, with the continuous final grade (G3, 0–20) as the target variable. Table 2 presents the student-related features and target variable of the Kaggle Student Performance Dataset (Dataset 1).

The Database Concept and Design course dataset consists of 991 instances collected from three cohorts at Universiti Malaysia Sarawak (UNIMAS), with eight attributes. The course dataset comprises numerical assessment scores (e.g., Assignment 1, Quiz) and a categorical target variable, Grade (A–F). Table 3 summarizes the assessment features of the Database Concept and Design Course Dataset (Dataset 2). These datasets were chosen for their complementary features, enabling comprehensive model development for grade prediction.

Phase 2: Data Pre-processing

The pre-processing phase cleaned Dataset 1 and Dataset 2 for transformation and modelling to ensure data quality and prevent data leakage. For Dataset 1, which initially contain 1,044 instances, numerical columns (e.g., age, absences) were checked for missing values and resolved using median imputation. Duplicates were removed and outliers in age and absences were excluded using the Interquartile Range method. The continuous G3 target variable which was measured on a 0-20 scale, was discretised into five different grade categories which are Grade F (0–7) will be labelled as “0”, Grade D (8–9) as “1”, Grade C (10–12) as “2”, Grade B (13–15) as “3”, and Grade A (16–20) as “4”. These specific thresholds were chosen to align with standard Malaysian higher education grading rubrics, where a score below 40% are considered as fail (Grade F). These categories follow the standard institutional letter-grade distributions (F, D, C, B, and A) to ensure the predictive model’s outputs are practically interpretable for educators. Categorical attributes (e.g., sex, address) were encoded with Scikit-learn’s OrdinalEncoder. Following these steps, Dataset 1 was reduced to 950 instances.

Table 2: Student related variables for Kaggle student performance dataset (Dataset 1)

Category	Feature	Description
Demographics	school	student's school
	sex	student’s sex
	age	student’s age
	address	student's home address type
	famsize	family size
Parental Background	Pstatus	parent's cohabitation status
	Medu	mother's education
	Fedu	father’s education
	Mjob	mother's job
	Fjob	father’s job
School Preferences	reason	reason to choose this school
	guardian	student’s guardian
Academic Attributes	traveltime	Home to school travel time
	studytime	Weekly study time
	failures	Number of pass class failure
	schoolsup	Extra educational support
	famsup	Family educational support
	paid	Extra paid classes within the course subject (Math or Portuguese)
	activities	Extra-curricular activities
Extracurricular	nursery	Attended nursery school
	higher	Wants to take higher education
	internet	Internet access at home
	romantic	With a romantic relationship
Behavioural	famrel	Quality of family relationships
	freetime	Free time after school
	goout	Going out with friends
	Dalc	Workday alcohol consumption
	Walc	Weekend alcohol consumption
	health	Current health status
Attendance	absences	Number of school absences
Grades	G1	First-period grade
	G2	Second-period grade
	G3	Final grade (Output Target)

Table 3: Features of the Database Concept and Design Course Dataset (Dataset 2)

Feature	Description
Assignment 1	First assignment score
Assignment 2	Second assignment score
Lab Test	Lab test score
Project	Project score
Quiz	Quiz score
Final Exam	Final examination score
Total Mark	Sum of assessment scores
Grade	Final grade (Output Target)

For Dataset 2, which initially comprised 991 instances, missing values in Quiz and Assignment 1 were imputed with median values using SimpleImputer, duplicates were removed, and outliers were retained unless unrealistic after boxplot inspection. The Grade variable was mapped to five categories, Grade F will be labelled as "0", Grade D as "1", Grade C as "2", Grade B as "3", and Grade A as "4". This unified classification mapping ensures that the numeric labels (0–4) used in the confusion matrices and evaluation reports throughout this study consistently represent the same academic performance levels across both datasets. No encoding was needed for numerical attributes, but empty columns were flagged for removal. After all pre-processing steps, Dataset 2 contained 985 instances.

Phase 3: Data Transformation

The transformation phase optimized both datasets by preparing input features for machine learning. Using the 950 cleaned instances of Dataset 1, irrelevant features (school, higher) and Final grade (Output Target) were dropped, reducing the number of attributes to just 30. Scikit-learn's SelectKBest with ANOVA F-test was utilised as a feature selection mechanism. This study uses the univariate statistical test to rank the importance of variables based on their individual statistical significance to the target variable, ensuring that only attributes with a meaningful impact are retained (Moussa et al., 2024). A total of 10 features were selected for the upcoming phase: Medu, Fedu, reason, studytime, failures, Dalc, Walc, absences, G1, and G2. Additionally, a G1_G2 interaction term was created and added to the feature set, resulting of 11 features. Since G1 and G2 represent consecutive first and second-period grades, their mathematical interaction captures a student's 'academic momentum'. This engineered feature provides critical sequential context, allowing the model to differentiate between a consistently average student and a student experiencing a sudden academic decline prior to the final evaluation. Numerical features were standardized using StandardScaler.

For Dataset 2, the Total Mark, Final Exam and Final grade (Output Target) were excluded, remaining the five features (Assignment 1, Assignment 2, Lab Test, Project, and Quiz). All features were retained after ANOVA F-test as this dataset contains only five input features and all features were retained and selected. Following, an Assignment1_Assignment2 interaction term added, resulting in a feature set of 6 attributes. The F-test scores were used solely for ranking and interpretability rather than dimensionality reduction.

Phase 4: Model Development, Training and Tuning

In this phase, hold-out method and 10-fold stratified cross-validation were applied to both datasets. Initially, hold-out method was applied to both datasets. For Dataset 1, which contains 950 instances, the data was split into a 70% training set (665 samples), a 10% validation set (95 samples), and a 20% testing set (190 samples). To address the class imbalance, SMOTE was applied to the training set, increasing it from 665 to 1,290 samples. For Dataset 2, containing 985 instances, the data followed the same split into 70% training set (689 samples), a 10% validation set (99 samples), and a 20% testing set (197 samples). After applying SMOTE, the training set for Dataset 2 increased from 689 to 1,265 samples.

The study also implemented 10-fold stratified cross-validation. This method divides the dataset into ten equal subsets. In each of the ten iterations, nine subsets are used for training and one subset is used for testing. For Dataset 1, this means 855 samples were used for training and 95 samples for testing in each fold. After applying SMOTE within the loop, the training set increased from 855 to 1,660 samples. Similarly for Dataset 2, 887 samples were used for training and 98 for testing per fold. To address the imbalance classes, SMOTE was applied to the training subsets within each iteration where the training set increased from 887 to 1,629 samples.

Given the importance of reliable student performance prediction, the selection of an appropriate predictive model is critical, particularly for supporting the early identification of at-risk students and enabling timely academic intervention. To address this, this study proposes a hybrid ensemble model that integrates XGBoost and DNN using a weighted soft voting strategy. The proposed model combined the predicted class probabilities generated by each base learner to compute a final weighted average, which determines the grade classification.

Models including KNN, SVM, XGBoost and DNN were set as the baseline models for comparison with the proposed model. Hyperparameter optimization such as Optuna and Hyperband was conducted using two distinct methods, chosen based on the architectural requirements and computational costs of the algorithms (Chiu et al., 2021; Kumar et al., 2025). Optuna was

utilized for the traditional ML models: KNN (Hasanshahi et al., 2026; Efendi et al., 2026), SVM (Ahmed et al., 2026; Efendi et al., 2026), and XGBoost (Al-Salih et al., 2026; Villar and de Andrade, 2024) because of its proven efficiency in exploring mixed-parameter spaces. Recent comparative studies (Ahmed et al., 2026; Al-Salih et al., 2026) show that Optuna demonstrates better efficiency in hyperparameter tuning when benchmarked against traditional approaches like GridSearchCV. This makes it the ideal choice for navigating the discrete and continuous search spaces of gradient-boosting and kernel-based models. Conversely, Hyperband (Li et al., 2018) was specifically selected for tuning the DNN to manage its high computational cost and faster model training.

Furthermore, the weighting strategy for the soft-voting ensemble was determined through an ensemble weighting optimization method that utilized a grid search technique to maximize the macro F1-score (Dhar, 2021). This process involved testing different weight combinations in increments of 0.1 to find the most effective balance between XGBoost and the DNN for each specific dataset. Based on this optimization, a weight ratio of 0.7 for XGBoost and 0.3 for DNN was identified as the optimal result for Dataset 1, whereas the default weight distribution provided the optimal balance for Dataset 2. The final classification is obtained by computing the optimized weighted average of the predicted class probabilities from both base learners to provide the final output.

Phase 5: Model Evaluation

The model evaluation phase assessed the performance of the hybrid ensemble (XGBoost + DNN) and baseline models (KNN, SVM, XGBoost, DNN) for predicting student grades (A–F) on both Dataset 1 and 2. Accuracy, precision, recall, and F1-score were calculated using the confusion matrix to evaluate multi-class classification performance, with F1-score emphasized for imbalanced grade distributions. Additionally, the confusion matrix table is used to visualise the classification performance of each predictive model. Fig. 2 presents the confusion matrix used for student grade prediction.

The performance metrics of the confusion matrix is determined using accuracy, precision recall and F1-score. Accuracy measures the proportion of correct predictions as shown in Equation 1. Precision is the ratio of correctly classified instances to total classified instances as shown in Equation 2. Recall is the ratio of correctly classified instances to total relevant instances as shown in Equation 3. The F1-score combines precision and recall as their harmonic mean as shown in Equation 4.

		Predicted Label				
		F	**D**	**C**	**B**	**A**
True Label	**F**	FF	FD	FC	FB	FA
	D	DF	DD	DC	DB	DA
	C	CF	CD	CC	CB	CA
	B	BF	BD	BC	BB	BA
	A	AF	AD	AC	AB	AA

Fig. 2: Confusion matrix for student grade prediction classification

$$Accuracy = \frac{(AA+BB+CC+DD+FF)}{\sum N} \quad (1)$$

$$Precision = \frac{1}{5}\left(\frac{AA}{AA+BA+CA+DA+FA} + \frac{BB}{AB+BB+CB+DB+FB} + \frac{CC}{AC+BC+CC+DC+FC} + \frac{DD}{AD+BD+CD+DD+FD} + \frac{EE}{AF+BF+CF+DF+FF}\right) \quad (2)$$

$$Recall = \frac{1}{5}\left(\frac{AA}{AA+AB+AC+AD+AF} + \frac{BB}{BA+BB+BC+BD+BF} + \frac{CC}{CA+CB+CC+CD+CF} + \frac{DD}{DA+DB+DC+DD+DF} + \frac{EE}{FA+FB+FC+FD+FF}\right) \quad (3)$$

$$F1-Score = 2\left(\frac{Precision \cdot Recall}{Precision+Recall}\right) \quad (4)$$

Experiment

In this study, five machine learning models (KNN, SVM, XGBoost, DNN, and DNN + XGBoost) were evaluated for predicting student grades (A-F) on two datasets. The aim was to identify the most effective model for detecting at-risk students in HEIs.

Experimental Setup

Experiments were conducted on the Kaggle Student Performance Dataset (Dataset 1, 950 samples after pre-processing) and the Database Concept and Design Course Dataset (Dataset 2, 985 samples after pre-processing) using Google Colab, leveraging its cloud-based GPU acceleration for efficient training. This research was built on Python 3.12, using stack of python libraries Pandas, NumPy, Scikit-learn, TensorFlow, XGBoost and Imbalanced-learn were used for pre-processing and modelling.

Hyperparameter Tuning

The list of hyperparameter search space and optimal values for Dataset 1 and Dataset 2 are presented in Tables 4-5, respectively.

Table 4: Hyperparameter search range and selected values for Dataset 1.

Model	Hyperparameter	Search Range	Optimal Value
KNN	n_neighbors	[3, 9]	3
	weights	{uniform, distance}	distance
	metric	{euclidean, manhattan}	Manhattan
SVM	C	[0.1, 10.0]	9.6954
	Kernel	{rbf, linear}	RBF
	gamma	{scale, auto}	auto
XGBoost	max_depth	[2, 4]	4
	min_child_weight	[3, 10]	4
	gamma	[0.1, 0.5]	0.1004
	reg_alpha	[1.0, 2.0]	1.4825
	reg_lambda	[1.0, 2.0]	1.4571
	learning_rate	[0.005, 0.05]	0.0473
	n_estimators	[100, 300]	300
DNN	units (layer 1)	[64, 128]	64
	L2 regularization (layer 1)	[0.01, 0.05]	0.032780
	dropout (layer 1)	[0.4, 0.6]	0.5
	units (layer 2)	[32, 64]	32
	L2 regularization (layer 2)	[0.01, 0.05]	0.026123
	dropout (layer 2)	[0.4, 0.6]	0.5
	units (layer 3)	[16, 32]	16
	learning_rate	[1e-5, 1e-3]	8.054449e-4
Hybrid Ensemble	Estimators	XGBoost, DNN	XGBoost, DNN
	Voting Type	Soft voting	Soft voting
	Weights	XGBoost: 0.7, DNN: 0.3	XGBoost: 0.7, DNN:0.3

Table 5: Hyperparameter search range and selected values for Dataset 2

Model	Hyperparameter	Search Range	Optimal Value
KNN	n_neighbors	[3, 15]	7
	weights	{uniform, distance}	distance
	metric	{euclidean, manhattan}	euclidean
SVM	C	[0.1, 10.0]	7.7749
	Kernel	{rbf, linear}	RBF
	gamma	{scale, auto}	auto
XGBoost	max_depth	[3, 10]	3
	min_child_weight	[1, 10]	3
	gamma	[0.0, 0.3]	0.4665
	reg_alpha	[0.0, 1.0]	1.4710
	reg_lambda	[0.0, 1.0]	1.5431
	learning_rate	[0.01, 0.3]	0.0435
	n_estimators	[100, 500]	200
DNN	units (layer 1)	[32, 128]	96
	L2 regularization (layer 1)	[0.001, 0.05]	0.026397
	dropout (layer 1)	[0.1, 0.5]	0.4
	units (layer 2)	[16, 64]	48
	L2 regularization (layer 2)	[0.001, 0.05]	0.006610
	dropout (layer 2)	[0.1, 0.5]	0.4
	units (layer 3)	[8, 32]	16
	learning_rate	[1e-5, 1e-3]	7.121132e-4
Hybrid Ensemble	Estimators	XGBoost, DNN	XGBoost, DNN
	Voting Type	Soft voting	Soft voting
	Weights	default	default

Results and Discussion

Model Effectiveness

Comparative analysis was conducted on the five models using hold-out method and 10-fold stratified cross validation. As shown in Table 6, the proposed hybrid ensemble achieved the best results on Dataset 1 using hold-out method, with an accuracy of 77.37% and a macro F1-score of 74.50%, while maintaining moderate standard deviation. Among the baseline models, XGBoost achieved the second-best performance (accuracy: 76.84%, macro F1-score: 73.46%), followed by SVM (73.68%, 67.69%), which outperformed DNN (67.89%, 66.48%) and KNN

(65.79%, 62.92%). KNN exhibited the lowest variability but the weakest performance. As shown in the confusion matrix in Fig. 3, the hybrid ensemble shows significantly fewer false positives in the grade "B" and "C" categories (Classes 3 and 2) than the baseline models.

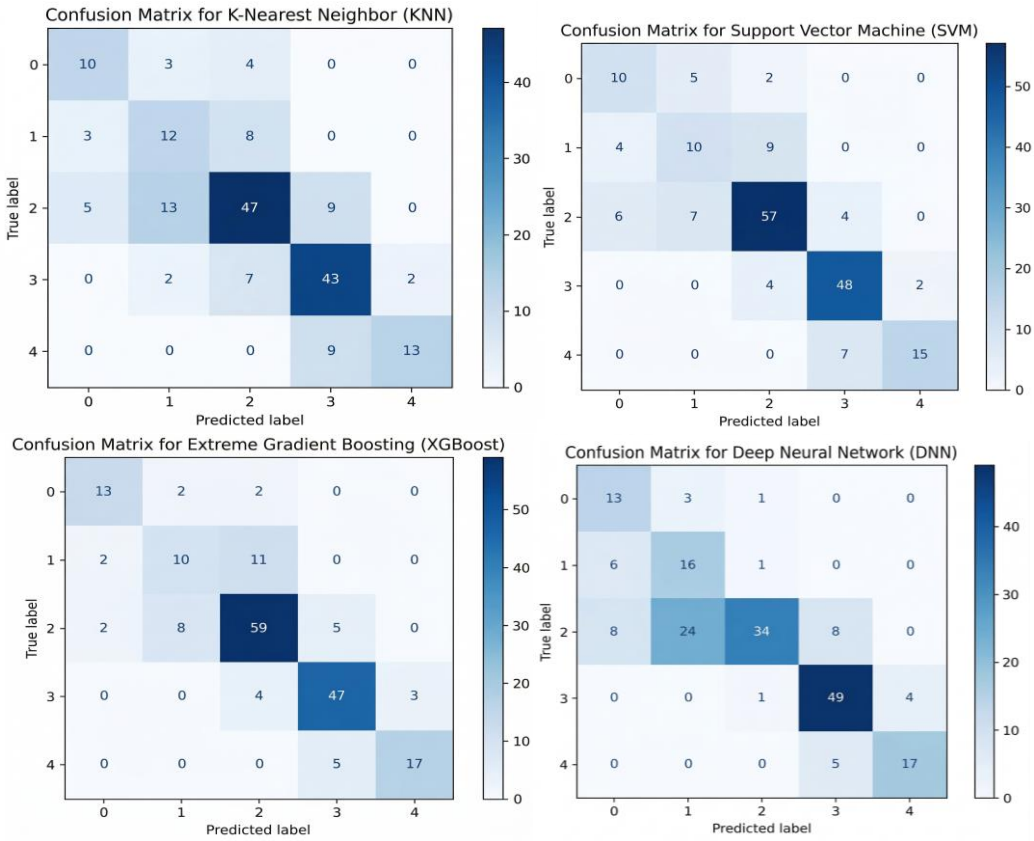
Following the hold-out evaluation, 10-fold stratified cross validation was performed to evaluate their stability and generalizability across different data subsets (see Table 7). The results further validate the hybrid ensemble as the best-performing model, with a macro F1-score of 74.31% and a macro recall of 75.69%, while KNN recorded the lowest scores.

Table 6: Overall model performance on Dataset 1 (Hold-out method)

Model	KNN	SVM	XGB	DNN	Hybrid Ensemble
Train Accuracy	99.92 ± 0.04	96.28 ± 0.46	91.71 ± 0.72	84.26 ± 0.84	91.32 ± 0.70
Validation Accuracy	62.11 ± 2.95	58.95 ± 2.91	66.32 ± 4.54	61.05 ± 4.73	68.42 ± 4.50
Test Accuracy	65.79 ± 2.29	73.68 ± 4.45	76.84 ± 3.04	67.89 ± 3.99	77.37 ± 3.47
Precision (Macro)	64.79 ± 3.03	68.84 ± 5.78	74.31 ± 4.40	67.45 ± 3.87	75.47 ± 4.02
Precision (Weighted)	67.62 ± 2.20	74.15 ± 4.18	76.41 ± 3.10	76.44 ± 3.57	77.24 ± 3.20
Recall (Macro)	62.65 ± 3.10	67.28 ± 5.05	72.80 ± 3.50	72.00 ± 4.28	73.67 ± 4.11
Recall (Weighted)	65.79 ± 2.29	73.68 ± 4.45	76.84 ± 3.04	67.89 ± 3.99	77.37 ± 3.47
F1-Score (Macro)	62.92 ± 2.90	67.69 ± 4.99	73.46 ± 3.73	66.48 ± 3.88	74.50 ± 3.16
F1-Score (Weighted)	66.14 ± 2.24	73.68 ± 4.26	76.55 ± 3.12	68.18 ± 3.95	77.25 ± 3.46

Table 7: Overall model performance on Dataset 1 (10-fold stratified cross validation)

Model	KNN	SVM	XGB	DNN	Hybrid Ensemble
Train Accuracy	99.90 ± 0.03	94.79 ± 0.34	91.93 ± 0.44	85.22 ± 2.08	90.95 ± 0.42
Test Accuracy	61.37 ± 6.30	70.11 ± 3.37	72.74 ± 3.35	73.48 ± 5.23	73.47 ± 4.03
Precision (Macro)	60.63 ± 5.94	67.47 ± 4.48	70.94 ± 4.61	65.80 ± 5.80	72.98 ± 4.98
Precision (Weighted)	62.61 ± 5.86	71.70 ± 3.35	73.44 ± 3.44	74.26 ± 5.19	74.52 ± 3.70
Recall (Macro)	58.80 ± 6.15	67.59 ± 3.57	70.52 ± 4.54	70.45 ± 5.66	75.69 ± 4.08
Recall (Weighted)	61.37 ± 6.30	70.11 ± 3.37	72.74 ± 3.35	73.48 ± 5.23	73.47 ± 4.03
F1-Score (Macro)	59.70 ± 6.03	67.53 ± 3.61	70.73 ± 4.27	68.05 ± 5.46	74.31 ± 4.14
F1-Score (Weighted)	61.98 ± 6.18	70.89 ± 3.42	73.09 ± 3.44	73.87 ± 5.15	73.99 ± 3.98



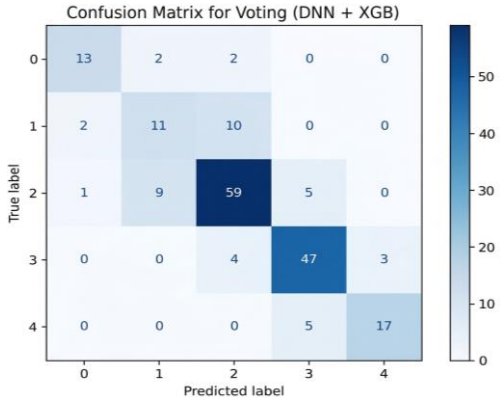


Fig. 3: Confusion matrix of each model in Dataset 1

The hybrid ensemble shows moderate variability, whereas KNN exhibited the highest standard deviation. The high training accuracy observed in some baseline models, particularly KNN, raises concerns about potential overfitting. The models tend to overfit the training data and fail to generalize effectively to unseen data. This issue may be due to the use of SMOTE on the training set, which creates highly similar synthetic samples that simplify the learning process. In contrast, the hybrid ensemble demonstrates a more balanced performance between training and testing results. Although the hybrid ensemble shows a modest improvement over the top baseline model, it consistently outperforms across multiple evaluation metrics in both hold-out and stratified cross validation evaluations, which are crucial for accurately identifying at-risk students.

For Dataset 2, Table 8 presents the overall model performance using the hold-out method. The proposed hybrid ensemble model achieved the highest accuracy of 74.13% and macro F1-score of 81.53%, along with moderate standard deviation. This performance surpasses all baseline models, including XGBoost (72.64% accuracy, 80.54% F1-score) and SVM (73.63% accuracy, 77.49%). As illustrated in Fig. 4, the hybrid ensemble model exhibits low error rates for the at-risk grade categories (F and D). Despite the small support for Grade C (Class 2, 13 samples), the ensemble has shown its dual-architecture strengths to correctly classify these instances, which often pose a challenge for traditional learners.

Moreover, both macro and weighted metrics are presented in Table 8 to provide a comprehensive evaluation. Macro metrics treat all classes equally and are more suitable for imbalanced datasets, while weighted metrics reflect overall performance based on class distribution. The results show that macro recall values (83–85%) are higher than accuracy (68–74), which is consistent with the imbalanced multi-class nature of Dataset 2. The accuracy may not fully reflect model performance, as it is influenced by the distribution of majority classes. This observation is supported by Sujon et al. (2025) that accuracy may be less reliable under class imbalance, whereas recall is more critical when false negatives are costly, and the F1-score provides a more balanced evaluation. Therefore, the higher macro recall and F1-score observed in this study indicate that the proposed model is effective in identifying at-risk students.

Similarly, the findings remain consistent when using 10-fold stratified cross validation (see Table 9).

Table 8. Overall model performance on Dataset 2 (Hold-out method)

Model	KNN	SVM	XGB	DNN	Hybrid Ensemble
Train Accuracy	99.13 ± 0.11	87.35 ± 0.61	88.77 ± 0.23	82.69 ± 0.64	87.67 ± 0.28
Validation Accuracy	64.00 ± 3.26	69.00 ± 3.71	72.00 ± 3.14	70.00 ± 4.63	71.00 ± 3.07
Test Accuracy	68.16 ± 4.62	73.63 ± 2.32	72.64 ± 2.75	72.14 ± 3.16	74.13 ± 2.86
Precision (Macro)	67.91 ± 5.02	73.22 ± 2.77	79.11 ± 4.20	68.78 ± 2.66	79.67 ± 4.08
Precision (Weighted)	70.14 ± 3.98	74.80 ± 2.43	73.80 ± 2.55	74.85 ± 2.75	76.11 ± 2.45
Recall (Macro)	83.53 ± 5.69	84.28 ± 2.37	82.72 ± 4.96	85.15 ± 2.57	85.03 ± 4.15
Recall (Weighted)	68.16 ± 4.62	73.63 ± 2.32	72.64 ± 2.75	72.14 ± 3.16	74.13 ± 2.86
F1-Score (Macro)	73.16 ± 5.14	77.49 ± 2.59	80.54 ± 4.81	73.83 ± 2.63	81.53 ± 4.12
F1-Score (Weighted)	68.47 ± 4.56	73.93 ± 2.39	72.97 ± 2.72	72.78 ± 3.24	74.58 ± 2.98

Table 9: Overall model performance on Dataset 2 (10-fold stratified cross validation)

Model	KNN	SVM	XGB	DNN	Hybrid Ensemble
Train Accuracy	96.53 ± 0.14	81.98 ± 0.80	85.31 ± 0.56	81.50 ± 1.29	84.41 ± 0.52
Test Accuracy	61.53 ± 4.56	68.34 ± 3.09	69.63 ± 4.54	69.34 ± 4.37	73.03 ± 3.84
Precision (Macro)	61.68 ± 4.49	68.04 ± 4.59	73.91 ± 4.89	68.64 ± 3.39	75.41 ± 3.67
Precision (Weighted)	65.42 ± 4.50	69.97 ± 4.61	69.93 ± 4.84	70.32 ± 3.91	70.27 ± 4.47
Recall (Macro)	63.39 ± 5.10	78.96 ± 5.12	75.52 ± 4.55	79.36 ± 3.32	79.71 ± 4.16
Recall (Weighted)	61.53 ± 4.56	68.34 ± 3.09	69.63 ± 4.54	69.34 ± 4.37	73.03 ± 3.84
F1-Score (Macro)	62.52 ± 5.66	73.09 ± 5.15	74.71 ± 3.77	73.61 ± 3.20	77.50 ± 3.22
F1-Score (Weighted)	63.42 ± 4.62	69.15 ± 4.63	69.78 ± 5.29	69.83 ± 4.87	71.62 ± 4.42

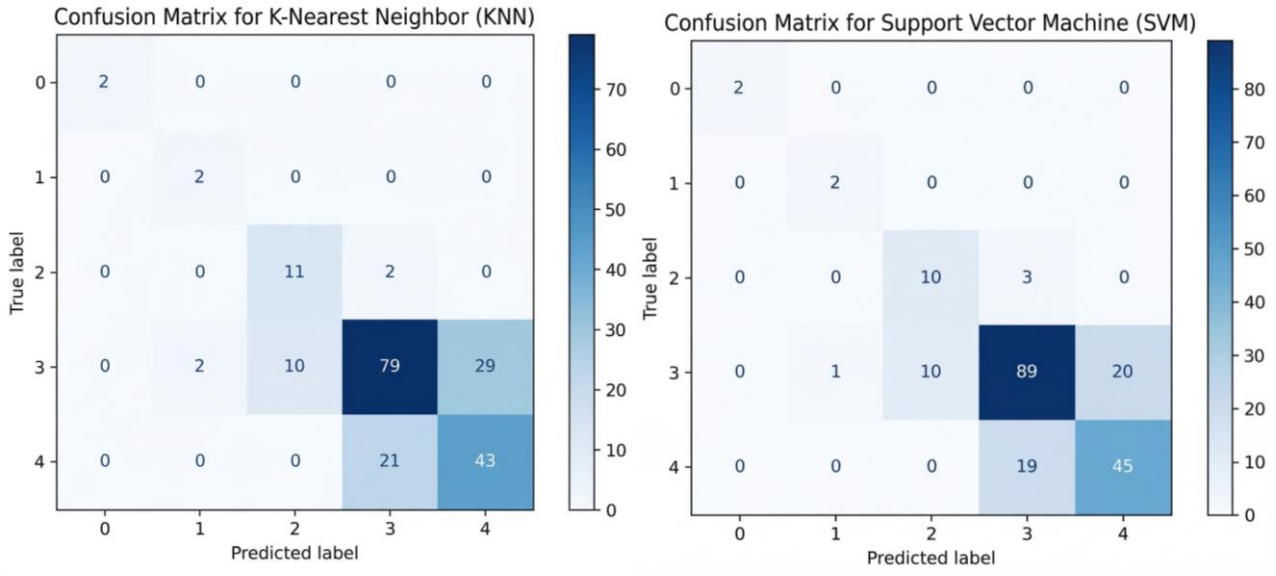
The proposed hybrid ensemble continues to outperform baseline models with the highest accuracy of 73.03% and macro F1-score of 77.50%, while KNN has the lowest scores (accuracy: 63.53%, macro F1-score: 62.52%). The hybrid ensemble shows moderate variability, whereas KNN shows the highest variability and exhibits potential overfitting. Overall, the proposed hybrid ensemble shows better generalization and confirms its effectiveness across hold-out and 10-fold stratified cross validation.

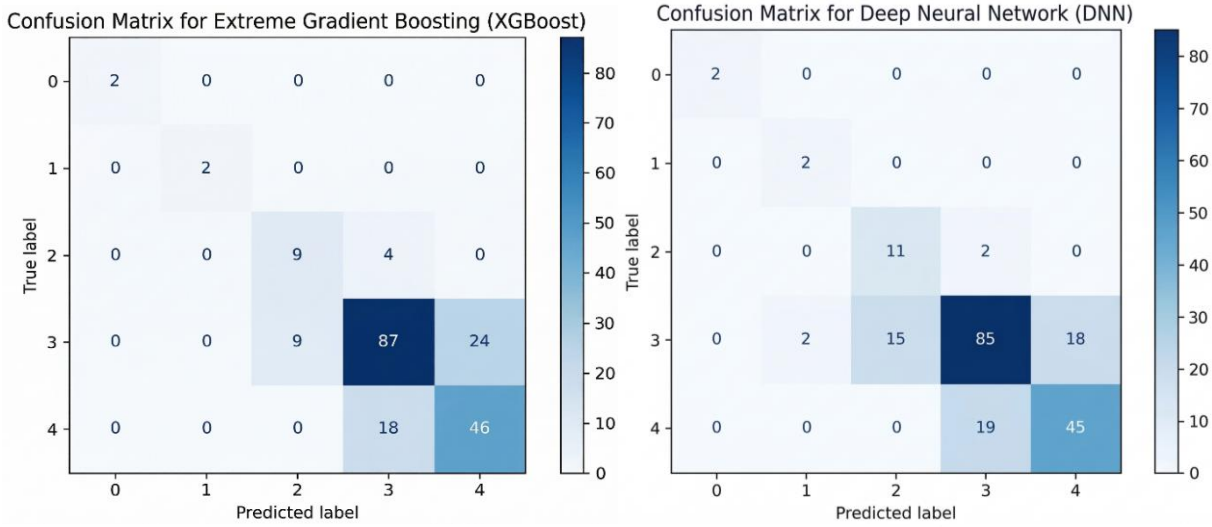
Furthermore, a t-test was used to assess the statistical significance of differences between the hybrid ensemble and the baseline models for Dataset 1 and Dataset 2 at the 0.05 significance level (refer to Table 10). The results indicate statistically significant differences between the hybrid ensemble and all baseline models (KNN, SVM, XGB, and DNN). These findings show that the hybrid ensemble achieves significantly better performance.

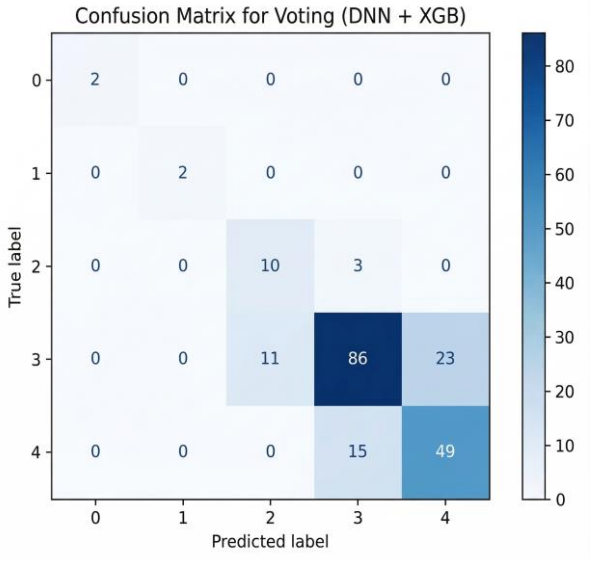
While the primary objective of this study is the early identification of at-risk students (Class 0: Grade F and Class 1: Grade D), it is equally important to evaluate the model's performance across middle-tier grades (Class 3: Grade B and Class 2: Grade C). Analysis of the confusion matrices (Figs. 3-4) reveals that the hybrid ensemble also shows discriminative ability in these categories, though specific challenges remain. Misclassifications in the middle tiers predominantly occur between adjacent classes, for example, actual Class 2 (Grade C) students being predicted as Class 3 (Grade B), or actual Class 3 students predicted as Class 4 (Grade A). This is a common phenomenon in educational data mining, as the behavioural and academic feature profiles (e.g., assignment scores and attendance) of Grade C and Grade B students often exhibit significant overlap, creating a highly complex decision boundary. Despite this inherent similarity, the proposed hybrid ensemble's soft voting mechanism leverages the combined strengths of XGBoost and DNN to mitigate these boundary errors. Consequently, the hybrid model correctly classifies most middle-tier instances and produces noticeably fewer false positives in the B and C categories compared to the standalone baseline models.

Table 10: T-test statistical significance results (p-values) between the proposed hybrid ensemble and baseline models.

Dataset	KNN	SVM	XGB	DNN
Dataset 1 (Hold-out)	0.005	0.038	0.006	0.0003
Dataset 1 (10-fold CV)	0.008	0.024	0.045	0.019
Dataset 2 (Hold-out)	0.050	0.025	0.011	0.014
Dataset 2 (10-fold CV)	0.035	0.005	0.011	0.007







Proposed Hybrid Ensemble (XGBoost + DNN)

Fig. 4: Confusion matrix of each model in Dataset 2

Impact of SMOTE and Feature Engineering

The performance of the hybrid ensemble was driven by two critical interventions which are class balancing and feature correlation. Addressing the inherent grade skew through SMOTE was essential. As summarized in Table 11, without SMOTE technique, the hybrid ensemble's macro F1-score decreased by 8.45% on Dataset 1 and 11.38% on Dataset 2. While the absolute numerical increase is modest, the oversampling was essential for stabilizing the decision boundaries of the rare "at-risk" categories (Classes 0 and 1). To validate the feature engineering stage, an ablation study was conducted by removing the engineered interaction features (G1_G2 and Assignment1_Assignment2). The results confirmed that these interaction terms are highly critical for model performance. For Dataset 1, removing G1_G2 caused the macro F1-score to drop from 74.50% to 71.86%. Similarly, for Dataset 2, removing the assignment interaction features resulted in a drop from 81.53% to 80.34%. By mathematically capturing performance trends over consecutive assessments, these features allow the model to distinguish between consistent performers and students experiencing sudden academic declines.

Computational Time Analysis

Furthermore, a comparison of the average execution time for KNN, SVM, XGBoost, DNN, and the proposed hybrid ensemble model is shown in Fig. 5. The KNN model demonstrated the lowest average execution time among all models, appearing nearly

instantaneous on the scale. Meanwhile, the proposed hybrid ensemble model required the longest average execution time to complete its tasks for both datasets. Overall, the execution time of the proposed hybrid ensemble model was slightly higher than the standalone DNN. In contrast, traditional machine learning models like SVM and XGBoost require significantly less computational time.

The proposed hybrid ensemble has the highest execution time, which is expected due to the integration of multiple models. While this hybrid ensemble architecture requires more computational resources than other baseline models, the model retraining can be conducted offline.

In practical deployment within a Learning Management System (LMS), student performance prediction is computationally lightweight. It can be performed in real time, either on cloud-based platforms or on local university servers. The framework can be set to automatically update students' predicted grades whenever new scores, such as midterm exam or assignment data, become available. This ensures that early warnings for at-risk students will stay up to date throughout the semester. Thus, the increased computational cost is justified by its improved predictive performance in identifying at-risk students and remains acceptable for practical use.

Comparison With Prior Study

The hybrid ensemble's performance is highly competitive with recent literature (Table 12). On Dataset 2, the proposed model achieved a test accuracy of 74.13% and a macro F1-score of 81.53%. These results outperform the Random Forest model reported by Dervenis et al. (2022), which attained an accuracy of 67.6%, and the multi-classification framework proposed by Balachandar and Venkatesh (2025), which achieved an accuracy of 76.0% and an F1-score of 73.0%. This comparison indicates that the proposed ensemble, which balances deep learning and gradient boosting, exhibits generalization capability for predicting complex, multi-class educational categories.

Table 11: SMOTE and Interaction Terms Ablation Study Results using Hold-Out Method.

Dataset	Model Configuration	Macro F1-Score (%)	Difference (%)
Dataset 1	Proposed Ensemble (G1_G2, SMOTE)	74.50	-
	Without SMOTE	66.05	-8.45
	Without G1_G2 Interaction	71.86	-2.64
Dataset 2	Proposed Ensemble (G1_G2, Assignment Interaction)	81.53	-
	Without SMOTE	70.15	-11.38%
	Without Assignment Interaction	80.34	-1.19

Table 12: Proposed Model Comparison with Previous Multiclassification Study

Paper	Metrics			
	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Dervenis et al., 2022	67.60	67.60	67.20	66.60
Balachandar and Venkatesh, 2025	76.00	79.00	81.00	73.00
Hybrid Ensemble (refer Dataset 2 at Table 8)	74.13	79.67	85.03	81.53

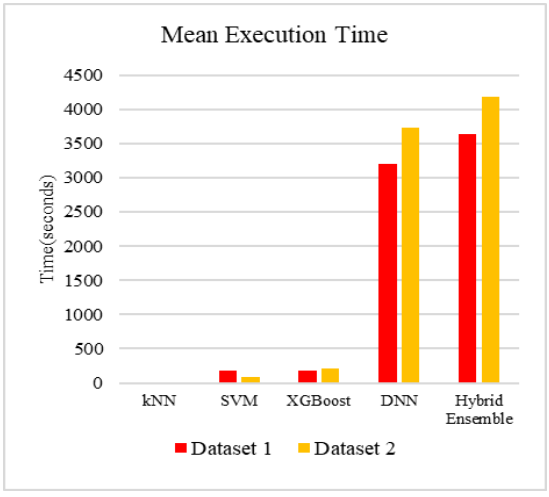


Fig. 5: Comparison of average execution time (seconds) for the hybrid ensemble and baseline models

Threats to Validity

Several factors may potentially compromise the validity of the evaluation results. In particular, internal validity threats, such as model overfitting, are explicitly observed in baseline models like KNN, which achieved training accuracies exceeding 99% but significantly lower test performance. This substantial generalization gap was a primary motivator for moving toward the proposed hybrid ensemble, which incorporates structural regularization and Early Stopping to ensure a more performance on unseen data. Furthermore, although high precision in identifying minority classes is supported by SMOTE, the synthetic instances generated may not fully capture the subtle behavioural patterns of real students.

For external validity, the framework was evaluated using two different datasets: Dataset 1 (n = 950) and Dataset 2 (n = 985), which can be considered moderate in

size. However, due to the multi-class (A–F) and imbalanced nature of the data, the effective sample size per class is reduced, particularly for minority classes. In addition, Dataset 2 consists of data from three cohorts, which may not be sufficient to support claims regarding the generalizability of the proposed approach. To enhance the generalization of the proposed framework, additional validation using datasets from other universities is recommended.

Conclusion

This study presented a hybrid ensemble XGBoost model using a weighted soft-voting strategy for multi-class student performance prediction and at-risk student identification. Five models (KNN, SVM, XGBoost, DNN, hybrid ensemble) were evaluated on both datasets. The methodology integrated class imbalance handling via SMOTE, feature engineering through academic interaction terms, and a dual-optimization approach using Optuna and Hyperband. On Dataset 1, the hybrid ensemble achieved a peak accuracy of 77.37% and a Macro F1-score of 74.50%. On Dataset 2, the model attained an accuracy of 74.13% and a Macro F1-score of 81.53%, consistently outperforming standalone baseline models.

Future research direction includes the application of generative AI-based data augmentation, such as Generative Adversarial Networks (GANs), to further refine synthetic samples beyond the traditional SMOTE approach. Exploring advanced machine learning techniques, such as attention-based models, could optimize feature weighting and improve accuracy. Additionally, employing explainable AI (XAI) methods, like SHAP or LIME, can improve model interpretability and support educators in understanding prediction outcomes and provide a more effective early intervention. Expanding the dataset to include diverse courses or larger student cohorts may also improve the robustness of multi-class student performance prediction.

Acknowledgment

The authors acknowledge UNIMAS Scholarship of Teaching and Learning Research Grant (UNI/F08/SoTL-RG/85796/2023) and Universiti Malaysia Sarawak for supporting this project.

Funding Information

The authors have not received any financial support or funding to report.

Authors Contributions

All authors contributed equally to this study.

Ethics

The use of predictive models in higher education requires responsible and fair data use. Although the proposed framework provides early detection, it should be treated as a supporting tool for early identification rather than a conclusive assessment. Therefore, predictions aim to support student academic success while avoiding unfair labelling or categorization

References

Adil, K., Messaoudi, F., Ahmed, A., & Youness, M. (2023). Machine Learning and Deep Learning based Students' Grades Prediction. *Operations Research Forum*, *4*(4), 1–21. https://doi.org/10.21203/rs.3.rs-3192793/v1

Ahmed, R., Fahad, N., Miah, S. U., Hossen, Jakir, & Bhattacharjee, K. (2026). HyOPTEnsemble: custom-weighted soft voting hyperparameter optimization ensemble model, explainable-AI for predicting mental state among university students. *Discover Artificial Intelligence*, *6*(1), 136. https://doi.org/10.1007/s44163-025-00708-9

Al-Salih, O. G., Guangjian, D., Al-Mudhafar, W. J., & Wood, D. A. (2026). Using extreme gradient boosting with Optuna hyperparameter tuning for efficient lost circulation prediction. *Energy Geoscience*, *7*(2), 100540. https://doi.org/10.1016/j.engeos.2026.100540

Ayienda, R., Rimiru, R., & Cheruiyot, W. (2021). Predicting Students Academic Performance using a Hybrid of Machine Learning Algorithms. *IEEE AFRICON*, 1–6. https://doi.org/10.1109/africon51333.2021.9571012

Bujang, A., S. D., Selamat, A., Krejcar, O., Mohamed, F., Cheng, L. K., Chiu, P. C., & Fujita, H. (2023). Imbalanced Classification Methods for Student Grade Prediction: A Systematic Literature Review. *IEEE Access*, *11*, 1970–1989. https://doi.org/10.1109/access.2022.3225404

Balachandar, V., & Venkatesh, K. (2025). A multi-dimensional student performance prediction model (MSPP): An advanced framework for accurate academic classification and analysis. *MethodsX*, *14*, 103148. https://doi.org/10.1016/j.mex.2024.103148

Bujang, S. D. A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., Fujita, H., & Ghani, N. A. (2021). Multiclass Prediction Model for Student Grade Prediction Using Machine Learning. *IEEE*

Access, *9*, 95608–95621. https://doi.org/10.1109/access.2021.3093563

Chiu, C., P., Selamat, A., Krejcar, O., Kuok Kuok, K., Herrera Viedma, E., & Fenza, G. (2021). Imputation of Rainfall Data Using the Sine Cosine Function Fitting Neural Network. *International Journal of Interactive Multimedia and Artificial Intelligence*, *6*(7), 39–48. https://doi.org/10.9781/ijimai.2021.08.013

Dervenis, C., Kyriatzis, V., Stoufis, S., & Fitsilis, P. (2022). Predicting Students' Performance Using Machine Learning Algorithms. *Proceedings of the 6th International Conference on Algorithms, Computing and Systems*, 1–7. https://doi.org/10.1145/3564982.3564990

Dhar, J. (2021). Multistage Ensemble Learning Model With Weighted Voting and Genetic Algorithm Optimization Strategy for Detecting Chronic Obstructive Pulmonary Disease. *IEEE Access*, *9*, 48640–48657. https://doi.org/10.1109/access.2021.3067949

Efendi, A., Fitri, I., & Nurcahyo, G. W. (2026). Development of a machine learning model with optuna and ensemble learning to improve performance on multiple datasets. *Indonesian Journal of Electrical Engineering and Computer Science*, *41*(1), 375–386. https://doi.org/10.11591/ijeecs.v41.i1.pp375-386

Farissi, A., Mohamed Dahlan, H., & Samsuryadi. (2020). Genetic Algorithm Based Feature Selection With Ensemble Methods For Student Academic Performance Prediction. *Journal of Physics: Conference Series*, *1500*(1), 012110. https://doi.org/10.1088/1742-6596/1500/1/012110

Hakkal, S., & Lahcen, A. A. (2024). XGBoost to Enhance Learner Performance Prediction. *Computers and Education: Artificial Intelligence*, *7*, 100254. https://doi.org/10.1016/j.caeai.2024.100254

Hasanshahi, M., Mehdizadeh, A., Mahmoudi, T., Ostovan, V. R., Nowroozzadeh, M. H., & Parsaei, H. (2026). An ensemble machine learning classifier for Parkinson's disease diagnosis using optical coherence tomography angiography. *Scientific Reports*, *16*(1), 7297. https://doi.org/10.1038/s41598-026-38407-9

Katuwal, KC. R., & Yang, S. (2026). A Hybrid Ensemble Approach for Early-Stage Diabetes Detection. *Healthcare Technology Letters*, *13*(1), e70060. https://doi.org/10.1049/htl2.70060

Kumar, D., Pawar, P. P., Addula, S. R., Meesala, M. K., Oni, O., Cheema, Q. N., Haq, A. U., & Sajja, G. S. (2025). AI-Powered Security for IoT Ecosystems: A Hybrid Deep Learning Approach to Anomaly Detection. *Journal of Cybersecurity and Privacy*, *5*(4), 90. https://doi.org/10.3390/jcp5040090

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2018). Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *Journal of Machine Learning Research*, *18*(185), 1–52.

Lim, T.-W., Khor, K.-C., & Ng, K.-H. (2019). Dimensionality reduction for predicting student performance in unbalanced data sets. *International Journal of Advanced Soft Computing and Its Applications*, *11*(2), 76–86.

Myrick, D. (2024). High School Student Performance & Demographics [Dataset]. Kaggle. *Kaggle*.

Nabil, A., Seyam, M., & Abou-Elfetouh, A. (2021). Prediction of Students' Academic Performance Based on Courses' Grades Using Deep Neural Networks. *IEEE Access*, *9*(2), 140731–140746. https://doi.org/10.1109/ACCESS.2021.3119596

Pujianto, U., Agung Prasetyo, W., & Rakhmat Taufani, A. (2020). Students Academic Performance Prediction with k-Nearest Neighbor and C4.5 on SMOTE-balanced data. *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 348–353. https://doi.org/10.1109/isriti51436.2020.9315439

Quinn, R. J., & Gray, G. (2019). Prediction of student academic performance using Moodle data from a Further Education setting. *Irish Journal of Technology Enhanced Learning*, *5*(1), 57. https://doi.org/10.22554/ijtel.v5i1.57

Sangodiah, A., Subramaniam, C. R. S. P. R., Beleya, P., & Muniandy, M. (2015). Minimizing student attrition in higher learning institutions in Malaysia using support vector machine. *Journal of Theoretical and Applied Information Technology*, *71*(3), 377–385.

Sarker, S., Paul, M. K., Thasin, S. T. H., & Hasan, Md. A. M. (2024). Analyzing students' academic performance using educational data mining. *Computers and Education: Artificial Intelligence*, *7*, 100263. https://doi.org/10.1016/j.caeai.2024.100263

Sujon, K. M., Hassan, R., Choi, K., & Samad, M. A. (2025). Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models. *Journal of Big Data*, *12*(1), 268. https://doi.org/10.1186/s40537-025-01313-4

Vaishnavi, P., Prathima, C., Rakesh, V., Sujala, P., Nitin, P. A., & Yadav, D. R. K. (2023). Employing the SMOTE Technique, a Machine Learning Model for Predicting Student Grades. *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 349–354. https://doi.org/10.1109/iciccs56967.2023.10142516

Villar, A., & de Andrade, C. R. V. (2024). Supervised

machine learning algorithms for predicting student dropout and academic success: a comparative study. *Discover Artificial Intelligence*, *4*(1), 2. https://doi.org/10.1007/s44163-023-00079-z

Wen, X., & Juan, H. (2023). Early Prediction of Students' Performance Using a Deep Neural Network Based on Online Learning Activity Sequence. *Applied Sciences*, *13*(15), 8933. https://doi.org/10.3390/app13158933

Yan, K. (2019). Student performance prediction using XGBoost method from a macro perspective. *Int. J. Adv. Soft Comput. Appl*, *11*(2), 76–86. https://doi.org/10.1109/CDS52072.2021.00084