Proactive forecasting of student academic outcomes via Random Forest and cumulative learning data

Tri Nhut Do[1*], Ngoc Mai Nguyen Thi[1, 2]

[1]*University of Information Technology, Vietnam National University - Ho Chi Minh City, Quarter 34, Linh Xuan Ward, Ho Chi Minh City, Vietnam*

[2]*Binh Duong University, 504 Binh Duong Avenue, Phu Loi Ward, Ho Chi Minh City, Vietnam*

Received 2 July 2025; revised 8 August 2025; accepted 5 October 2025

Abstract:

This study develops a Random Forest-based predictive framework to forecast course outcomes for Information Technology students at Binh Duong University using institutional academic records. Key features such as course codes, instructor identifiers, and encoded student IDs were combined with demographic and performance data through a standardised preprocessing pipeline that included feature engineering, normalisation, and class-imbalance handling. The final model achieved an accuracy of 0.98, a ROC AUC of 0.982, and a recall of 0.84 for the "Fail" class, demonstrating strong predictive power and practical value for early risk detection. Feature-importance analysis confirmed the dominant influence of course- and instructor-related variables alongside cumulative student histories, ensuring interpretability. By providing reliable and transparent predictions, the framework supports academic advisers in identifying at-risk students and designing timely interventions, highlighting the potential of institutional data for enhancing educational decision-making.

Keywords: **Random Forest, Data Mining, student evaluation, Educational data mining.**

Classification number: **1.2**

1. Introduction

In the context of modern higher education, leveraging data for academic management and student support has become an inevitable trend. Universities now have

[*]*Corresponding author: Email: trinhutdo@uit.edu.vn*

access to a wide range of data related to students' academic performance, learning behaviour, and training outcomes. Effectively utilising this data to identify students at academic risk and provide timely interventions is a key strategy in improving educational quality.

High performance in forecasting academic outcomes and detecting relevant elements has been shown by machine learning models, especially Random Forest [1] [2] .Despite the availability of useful student data from institutional systems, the use of such models remains limited in Vietnam, particularly at remote campuses.

Given this, a prediction model for course results among information technology students at Binh Duong University is proposed by the current study. To achieve this, it investigates the relationships between academic performance and factors including gender, behaviour scores, credits obtained, study time, and teaching-assessment methods using the Random Forest algorithm.

The study aims to:

- Develop a model to predict students' final course scores using Random Forest;
- Analyse the impact of cumulative academic features on learning outcomes;
- Provide actionable insights to support academic advising based on model analysis.

Accordingly, the study addresses the following research questions:

- How accurately can the Random Forest model predict students' course outcomes?
- Which features in the cumulative learning records have the greatest influence on performance?
- How can the findings inform academic advising and instructional improvement?

The major contributions of this study are summarised as follows:

- Proposes a Random Forest-based prediction model tailored to the academic context of Binh Duong University, demonstrating the feasibility of implementing educational data mining in Vietnam.

- Highlights the importance of contextual factors—such as conduct scores, accumulated credits, study time, and teaching–assessment methods—that have received limited consideration in prior studies.
- Provides interpretable outcomes through feature importance analysis, enabling academic advisers to design timely and targeted interventions for at-risk students.

This research seeks to contribute both theoretical foundations and practical directions for enhancing assessment systems and advancing strategies for educational quality improvement.

This paper is structured as follows. Section 2 describes the Literature review, while Section 3 is the Methodology for the predictive model proposal, including the system architecture overview and evaluation metrics. Section 4 presents some experimental results that demonstrate the accuracy of the proposed predictive model, accompanied by an evaluation. Finally, Section 5 concludes the paper and suggests some future directions for improvement.

2. Literature review

2.1. Theoretical framework and prior studies

In the context of modern higher education, there is growing interest in using big data and machine learning to predict students' academic achievement. Beyond identifying students who would perform poorly in the early stages, the goal is to assist institutions in implementing more successful and customised interventions.

Recent studies have demonstrated the effectiveness of prediction models, especially Random Forest, in assessing academic risk. A.M. Rabelo, et al. (2025) [3] used learning and fundamental data to predict dropout probability, while M. Pellagatti, et al. [4] used a mixed model to estimate failure and dropout rates. Yu et al. [5] assert that midterm scores and learning behaviours significantly influence outcomes in online learning environments.

Such findings are consistent Dass et al. with [5] who achieved AUC of 94.5% in predicting MOOC dropout using daily learning data, and Musa [6], who demonstrated RF's superiority over NB and LR in predicting foundation-year performance. Gusnina

et al. [7] and Nachouki et al. [8] also confirmed RF's reliability across institutional datasets.

In addition, the field of Educational Data Mining (EDM) has increasingly emphasised that predicting student achievement is a central task, as it enables educators to identify learning patterns, optimise teaching methods, and design timely interventions (Gaftandzhieva et al., [9]). Deep Learning (DL) approaches such as Bi-directional LSTM (Bi-LSTM) have emerged as promising solutions to overcome the limitations of traditional machine learning, particularly in handling large-scale educational datasets to predict GPA more accurately (Kalita et al., [10]).

The research focus has also moved beyond merely identifying at-risk students to enabling more personalised interventions. For instance, Bi-LSTM has been applied to predict second-semester GPA with an average accuracy of 88.23%, statistically outperforming CatBoost, XGBoost, HistGradientBoosting, and LightGBM (Kalita et al., [10]). Similarly, in predicting the performance of students with disabilities in Virtual Learning Environments (VLEs), the LSTM-SHAP framework achieved 92.88% accuracy, with rigorous statistical tests (e.g., Friedman test, ANOVA, Nemenyi post hoc) confirming its superiority (Kalita et al., [11]). Importantly, interpretability methods such as SHAP have been integrated to explain key drivers of success, ensuring transparency in educational decision-making (Kalita et al., [10]; Kalita et al., [11]).

All things considered, the literature emphasises how machine learning models are progressively taking the place of conventional approaches in academic forecasting, highlighting the importance of behavioural data and cumulative records in improving prediction accuracy and facilitating data-informed decision-making in higher education.

2.2. Applications of Random Forest in education

The Random Forest (RF) algorithm is a supervised ensemble learning technique that builds several decision trees and uses majority vote to aggregate their results. Because of its resilience, capacity to manage multidimensional data, and decreased likelihood of overfitting, RF has been extensively used in educational applications like learning behaviour analysis, dropout risk detection, student capability classification, and academic success prediction.

Its architecture leverages bootstrap sampling to train each tree on a different subset of the data, combined with random feature selection at each split, which enhances model diversity. This is particularly advantageous when working with complex educational datasets that include many categorical variables.

Recent research has demonstrated how beneficial RF is in educational settings. When its hyperparameters are adjusted correctly, RF performs better than conventional methods, as shown by Huynh-Cam et al. [2] and Liu et al. [1]. To better classify at-risk learners in online learning contexts, Yu et al. [12] and Rimal et al. [13] used RF to examine student behaviour. Additionally, Paek et al. [14] used RF to investigate the elements that influence creative behaviour, highlighting the model's capacity to improve interpretability and prioritise feature importance. These results align with real-world applications: Dass et al. [5] showed effective RF modelling for MOOCs; Sulak & Koklu [15] compared RF to DT and ANN in dropout classification; Gusnina et al. [7] and Ahmed [16] also demonstrated its versatility in large-scale student datasets.

Furthermore, Liu & Zhuang [17] applied RF in evaluating multimedia-based teaching quality, illustrating its broader relevance across educational domains.

Overall, Random Forest not only delivers strong predictive performance but also supports deeper analysis, making it a valuable tool for educational decision-making and institutional management.

2.3. Input data in educational contexts

Input data is a crucial component that influences the model's accuracy and usefulness in machine learning applications in education, especially when it comes to predicting academic performance. Input features can come from a variety of sources, including socio-emotional traits, academic records, online learning activities, and demographic profiles, depending on the study's objective.

As illustrated in Fig. 1, input data in educational research can be categorised into four main groups:

- Academic metrics: Includes grades, accumulated credits, and course failure rates. These are standard features in traditional educational models.
- Demographic information: Variables like age, gender, academic major, and entry qualifications help support personalised predictions.

- Online learning behaviour: Data from LMS platforms (e.g., login frequency, time spent, task completion) reflects levels of engagement and motivation.
- Attitudinal and behavioural factors: Although less frequently used, features such as learning attitude, emotional state, proactivity, or creativity can enhance personalization in predictive systems.
- Combining these kinds of data enhances model performance, according to studies. For instance, Yu et al. [12] showed how online habits affect academic outcomes; Paek et al. [14] focused on learning attitudes when constructing prediction models; and A.M. Rabelo, et al. (2025) [3] used academic and demographic data to predict dropout risk.

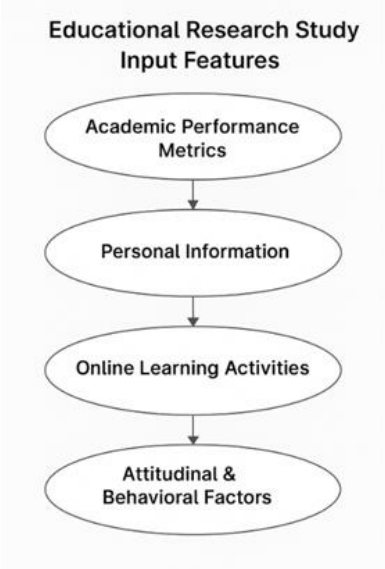

Fig. 1. Categories of input features in educational research.

However, challenges remain in harmonising data structures across studies and contextualising features within pedagogical frameworks, such as teaching methods and assessment types. Overcoming these limitations is key to building effective, scalable predictive models in higher education.

2.4. Research gap

Although machine learning models, and Random Forest in particular, have been widely applied to predict student performance and detect dropout concerns, much of the prior work has been situated in foreign educational contexts. These studies often rely

on rich learning behaviour data integrated with robust digital infrastructures and advanced Learning Management Systems (LMS). Such conditions differ substantially from the Vietnamese context, where academic institutions primarily rely on traditional student information systems and lack systematic behavioural tracking.

Furthermore, recent research has tended to emphasise predictive accuracy, frequently prioritising algorithmic optimization over practical deployment. While advanced approaches such as Bi-LSTM with SHAP-based interpretability or hybrid ensemble models [11] demonstrate promising accuracy, their applicability in resource-constrained environments is limited. Critical contextual factors - such as student conduct scores, cumulative study time, and the pedagogical or assessment methods - are often overlooked despite being available within institutional databases. This underutilization arises largely from challenges in data standardization and encoding, leaving a gap between experimental models and operationally viable systems.

Another critical shortcoming is the scarcity of research leveraging local Vietnamese college datasets. Most existing models are built and validated on international data, which constrains their relevance for institutions in Vietnam. As a result, models that perform well in experimental conditions often cannot be directly adopted into academic administration systems, thereby limiting their real-world utility.

Given these limitations, a clear research gap emerges: the need for predictive models grounded in authentic, context-specific academic data, designed for the operational realities of Vietnamese higher education. The present study responds to this gap by employing historical records of Information Technology students at Binh Duong University. By applying the Random Forest algorithm and incorporating contextual academic variables, it seeks to deliver not only accurate forecasts of course outcomes but also actionable insights for academic advising and institutional improvement.

3. Methodology

3.1. System architecture overview

The system is developed with a modular architecture to facilitate early intervention tactics and anticipate students' probability of failing the course. This method guarantees scalability, flexibility, and ease of upkeep. The overall system architecture is broken into discrete functional modules, each of which oversees a

particular end-to-end pipeline function, such as data processing, prediction creation, or result analysis. The block diagram that is displayed shows the data flow and inter-module interactions.

As depicted in Fig. 2, the system comprises the following key modules:

- Data Collection and Preprocessing: This module integrates academic data from multiple sources, performs data cleaning, handles missing values, standardises formats, and extracts relevant features for model training.
- Model Training: The processed data is used to train a Random Forest algorithm. Hyperparameter tuning is conducted to optimise the model's performance.
- Prediction and Probability Adjustment: This component forecasts the likelihood of students passing or failing a course. It also calibrates the predicted probabilities by incorporating students' academic history and individual risk profiles, thereby enhancing the personalization of the outcomes.
- Analysis and Visualization: This module generates model evaluation reports, performance charts, feature importance rankings, and interpretable insights derived from the academic data.
- User Interface (CLI/API): It provides an interface for users to input new data and retrieve the corresponding prediction results and analytical feedback.

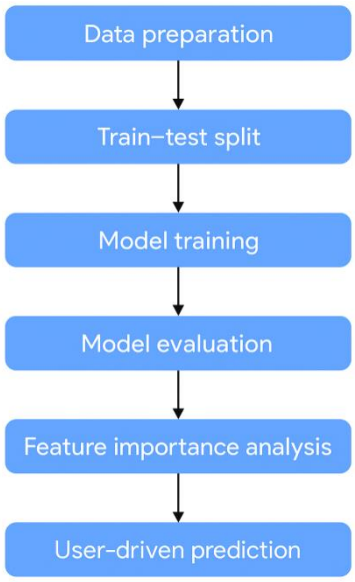

Fig. 2. Machine learning workflow for educational outcome prediction.

The system processes data sequentially, starting from collection and preprocessing, followed by model training, prediction, and analysis, and finally presenting the output through a user-interactive interface.

3.2. Dataset description and preprocessing

The dataset used in this study was obtained from Binh Duong University's academic management system. It includes detailed information about students, instructors, course modules, and component scores. Standardised data formats were used, and special cases - such as the code "VT" (absent from the exam) - were handled appropriately. Additionally, all grades were converted from a 10-point to a 6-point scale to facilitate the analysis of course learning outcomes (CLOs).

The dataset spans from the academic year 2019-2020 to the first semester of 2024-2025, comprising nearly 20,000 samples. All data were directly extracted from the official grade records stored in the university's student data warehouse, ensuring accuracy and consistency.

During the preprocessing stage, identifying labels (such as student ID, course ID, and lecturer name) were standardised and encoded. Additionally, the academic features derived, including GPA, course failure rate, exam absences, and recent performance trends, were calculated. Both the input feature set and the target variables, which include two classification labels: course outcome (pass/fail) and CLO achievement (met/not met) were constructed based on these features.

The main preprocessing steps were as follows:

- Handling missing and invalid values: Special values such as "VT" (absent from exam) and missing entries (NaN) in key score columns like Percent_Final_Exam and summary_score| were replaced with 0, reflecting non-completion of the course.
- Grade normalization: All scores were converted to a uniform 10-point scale. Final exam scores were additionally converted to a 6-point scale (exam_score_6) to align with CLO assessment requirements.
- Feature engineering:
  - is_absent_exam, is_absent_summary: Binary features indicating exam absence.

- Result: The primary label denotes whether the student passed or failed the course.
- DatCLO: A secondary variable representing CLO achievement status. Itemise.

- Categorical variable encoding: Identifier attributes such as Student_ID, Lecturer_Name, and Subject_ID were encoded as integers using label encoding, resulting in features like student_id_encoded, lecturer_encoded, and subject_encoded.

To ensure the validity of the experimental results and prevent any form of data leakage, a strict isolation protocol was implemented from the initial stage of data preparation. Specifically, data collection and partitioning were performed before any preprocessing; stratified sampling was applied during splitting to preserve class distribution; and all normalization, encoding, and feature engineering procedures were fitted exclusively on the training set before being consistently applied to validation and test sets. This approach guarantees that no information from the test set leaked into model training or hyperparameter tuning.

3.3. Predictive model: Random Forest classifier

3.3.1. Overview of Random Forest

This research uses a quantitative approach to develop a predictive model that identifies students who are likely to fail a course based on historical academic performance data. In particular, the model determines a student's chances of passing or failing a topic based on component scores and course-specific features. Along with early warning indications, the anticipated outputs are intended to be used as decision-support tools by instructors and academic advisers to develop timely and customised interventions.

The Random Forest (RF) algorithm, a supervised machine learning method that falls under the ensemble learning paradigm, is the chosen approach. A group of independently trained decision trees makes up a Random Forest. A bootstrap sample, as in (1), which is a randomly selected portion of the initial training dataset with replacement, is used to train each tree. By adding diversity to the individual trees, this resampling technique improves overall model performance and reduces the possibility of overfitting.

$$Dt \sim Bootstrap(D), t = 1,2, \dots, T \quad (1)$$

Random Forest has been widely recognised in recent educational data mining studies [1, 2, 6] for its high predictive accuracy, robustness against overfitting, and its ability to interpret results through feature importance analysis.

Using a majority voting mechanism, the Random Forest model aggregates the outputs of several decision trees to produce predictions for the binary classification job this work addresses. As shown in Fig. 3, a set of independently built decision trees (from Tree 1 to Tree T) is fed the input data, including features 1 through n, simultaneously. A categorization result is independently produced by each tree. A voting procedure is then used to integrate these individual forecasts, and the majority vote determines the result, which is either "Pass" or "Fail."

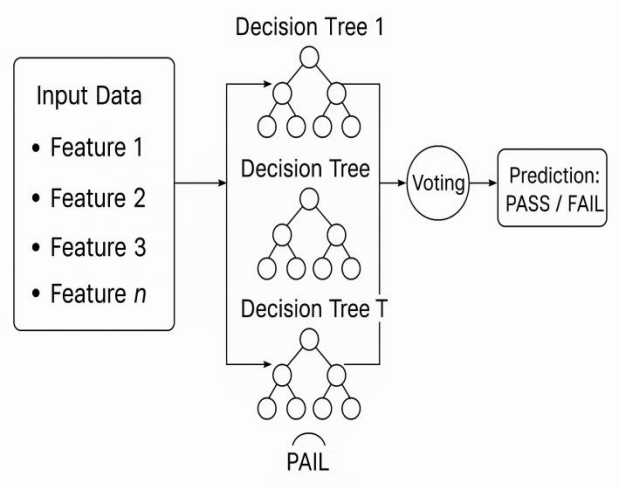

Fig. 3. Ensemble voting in a Random Forest model for course outcome prediction.

During the node-splitting process within each decision tree, the algorithm does not evaluate the full set of features. Instead, it randomly selects a subset of features and chooses the one that provides the optimal split based on a selected impurity reduction criterion. One of the most used criteria is Gini impurity, which is computed by Eq. (2) as follows:

$$Gm = 1 - \sum_{k=1}^{K} b_{mk}^2 \quad (2)$$

In the binary classification task addressed in this study, the Random Forest model generates predictions by aggregating the outputs of multiple decision trees using a majority voting strategy as Eq. (3) below:

$$y\hat{} = mode\left(h_1(x), h_2(x), \dots, h_T(x)\right) \quad (3)$$

To enhance both accuracy and generalizability, each tree in the forest is trained on a different subset of the training data, drawn through bootstrap sampling, a technique that randomly samples data points with replacement. Furthermore, at each decision node, only a randomly selected subset of features is considered for splitting. This combination of data and feature randomness promotes diversity among the trees, which in turn improves the overall performance of the ensemble.

Through feature importance analysis, which measures each input feature's contribution to the classification result, Random Forest offers interpretability in addition to its predictive power. Calculating the average impurity reduction across all nodes, as in Eq. (4), where a certain feature is used to split the data, is a popular method for determining feature relevance.

$$I(fj) = 1T \sum_{t=1}^{T} \sum_{n \in N_j^t} \frac{(w_n \Delta i(n))}{(w_{root}^t)} \quad (4)$$

3.3.2. Model configuration and training

To maximise generalisability and ensure methodological rigour, the dataset was divided into three subsets: 64% for training, 16% for validation, and 20% for testing. The training set was used exclusively for model fitting, the validation set for hyperparameter tuning and model selection, and the test set was held out and reserved only for the final evaluation. A fixed random seed (random_state=42) was applied to guarantee reproducibility.

All preprocessing steps were encapsulated in a scikit-learn Pipeline that combined StandardScaler with RandomForestClassifier. This design ensured that normalization, encoding, and feature engineering were fitted solely on the training data, and then consistently applied to both validation and test sets, thereby preventing any risk of data leakage.

Hyperparameter optimization was conducted using GridSearchCV on the training set with 5-fold cross-validation. The key parameters tuned included:

- *Number of trees (n_estimators): [50, 100, 200]*
- *Maximum tree depth (max_depth): [5, 10, 15, None]*
- *Minimum samples per leaf (min_samples_leaf): [1, 3, 5]*
- *Class weights: ['balanced', custom weights, None]*

The final model was selected based on validation performance, and only after all tuning was complete was the test set used for the final evaluation. This configuration ensures reliable performance estimates, avoids over-optimistic metrics, and enhances the interpretability and scalability of the model for practical applications in educational environments.

3.4. Evaluation metrics

A collection of common machine learning assessment measures was used to thoroughly evaluate the Random Forest model's performance. These metrics assess the model's overall accuracy as well as its capacity to accurately classify individual classes, which is particularly important when attempting to identify students who are at risk of academic failure.

3.4.1. Overview of Random Forest

To ensure reliable and unbiased assessment, model performance was evaluated using a three-way data split: training, validation, and test sets. During model development, GridSearchCV with internal cross-validation was applied on the training set to optimise key hyperparameters. The validation set was then used both for selecting the best-performing configuration and for diagnosing potential overfitting or underfitting. This diagnosis was performed by comparing performance on the training and validation sets: consistently high training accuracy but low validation accuracy indicates overfitting, whereas low accuracy on both sets suggests underfitting.

Only after all hyperparameters were finalised was the test set used for final evaluation. This approach ensures that the reported metrics truly reflect the model’s ability to generalise to unseen data, preventing overly optimistic estimates caused by data leakage or improper use of the test set during model selection.

A comprehensive set of evaluation metrics - including accuracy, precision, recall, F1-score, and ROC-AUC - was reported for both the validation and test sets. These measures not only capture overall predictive accuracy but also reflect the model’s ability to correctly identify at-risk students, which is particularly critical in educational applications.

3.4.2. Classification report

Table 1 presents detailed classification metrics, including precision, recall, and F1-score for each class. The model achieved an overall accuracy of 98% on the test set, demonstrating strong generalisation performance.

For the "Fail" class, the precision was 0.92 and recall was 0.84, indicating that most at-risk students were correctly identified while maintaining a relatively low false-positive rate. For the "Pass" class, the precision was 0.98 and recall was 0.99, resulting in an F1-score of 0.99. These results highlight the model's robustness in predicting successful outcomes while also offering reliable early detection of students at academic risk.

Table 1. Classification report on the model for student performance prediction.

Class	Precision	Recall	F1-score	Support
0 (Fail)	0.92	0.84	0.88	464
1 (Pass)	0.98	0.99	0.99	4013
Accuracy			**0.98**	4477
Macro Avg	0.95	0.92	0.93	4477
Weighted Avg	0.98	0.98	0.98	4477

3.4.3. Index additions

To provide a more comprehensive evaluation of the model's performance, several supplementary metrics were also calculated on the test set:

- Sensitivity (Recall of the "Pass" class): 0.992, reflecting the model's ability to correctly identify 99.2% of students who passed.
- Specificity (Recall of the "Fail" class): 0.845, indicating the model's capacity to correctly detect students who failed.
- Precision for the "Pass" class: 0.982, highlighting the model's strong accuracy in identifying truly passing students.
- F1-score for the "Pass" class: 0.987, showing a well-balanced performance between precision and recall for the majority class.
- ROC AUC Score: 0.982, confirming that the model has very strong discriminative capability between the two classes, substantially exceeding the 0.5 baseline.

- (Optional, recommended) Balanced Accuracy: 0.919, providing a fairer measure under class imbalance by averaging sensitivity and specificity.

3.4.4. Feature importance analysis

In addition to performance metrics, the study also examined the relative contribution of each input feature to the model's decision-making process. The results, summarised in Table 2, highlight the dominant role of course-, student-, and instructor-related variables compared with demographic factors.

Table 2. Feature importance scores of input variables used in the model

Feature	Importance Score
subject_encoded	0.389
student_id_encoded	0.334
lecturer_encoded	0.234
BirthYear	0.019
QueQuan_encoded	0.007
Religion_encoded	0.006
Gender_encoded	0.005
Ethnicity_encoded	0.004

The variable subject_encoded (course code) emerged as the most influential predictor, accounting for nearly 39% of the importance weight. This reflects the strong impact that course-specific factors, such as subject complexity and assessment structure, have on student outcomes. The second most important feature was student_id_encoded (student identifier) with 33%, which captures cumulative aspects of individual academic histories and prior performance trends. Lecturer_encoded (instructor identity) also played a significant role at 23%, suggesting that teaching methods and grading practices influence the probability of student success.

In contrast, demographic variables such as gender, ethnicity, religion, and place of origin contributed minimally (<2%), indicating that these factors had relatively little impact on the model's predictions compared to academic and instructional variables.

4. Experimental results and evaluation

4.1. Computational environment and dataset description

The experiments were implemented in Python using the scikit-learn library, ensuring compatibility and reproducibility across runs. The dataset was obtained from Binh Duong University's academic management system and comprises 22,383 course-level records collected from the 2019–2020 academic year through the first semester of 2024-2025.

Each record corresponds to a student's enrollment in a specific course and is labelled as "Pass" (20,064 instances, 89.6%) or "Fail" (2,319 instances, 10.4%) based on final exam outcomes. The input features include course code, instructor identity, student identifier (encoded), and demographic variables such as birth year, gender, ethnicity, religion, and place of origin.

A standardised preprocessing pipeline was applied using scikit-learn's Pipeline, encompassing normalization with StandardScaler, categorical encoding, and engineered features. To address the significant class imbalance (89.6% Pass vs. 10.4% Fail), both class weighting strategies and stratified sampling were employed during training and hyperparameter tuning to ensure that the minority "Fail" class received sufficient emphasis.

For model development, the dataset was split into 64% training (14,324 records), 16% validation (3,582 records), and 20% testing (4,477 records). The training set was used to fit the model, the validation set for hyperparameter tuning and model selection via GridSearchCV with 5-fold cross-validation, and the test set was held out for final evaluation. A fixed random_state=42 ensured reproducibility of all results.

4.1.1. Data capturing and labelling

The dataset used in this study was extracted from the academic management system of Binh Duong University, comprising 22,383 course-level records. Among these records, 20,064 (89.6%) were labelled as "Pass" and 2,319 (10.4%) as "Fail," reflecting a significant class imbalance typical of higher education datasets.

Each record represents a student's performance in a specific course and includes key attributes such as student ID, course code, instructor name, final exam grades, and attendance status. To prepare the data for machine learning, categorical identifiers (student ID, course code, instructor) were label-encoded, and demographic features (religion, ethnicity, place of origin, gender) were processed similarly.

A rigorous preprocessing pipeline was applied to ensure consistency. Absences from exams (denoted as "VT") were encoded as zero to reflect non-completion. All exam scores were standardised, with final exam scores converted to both 6-point and 10-point scales. Additional features were engineered, including a binary pass/fail outcome based on exam performance and a CLO (Course Learning Outcome) achievement indicator using a threshold of 3.5 on the 6-point scale.

For the binary classification task, records were labelled as "Pass" when students achieved the required exam score and were not absent from the exam, while others were labelled as "Fail." This labelling scheme served as the ground truth for evaluating student performance and enabled early identification of those at risk of academic failure.

4.1.2. Training and validating

Following preprocessing and labelling, the dataset - comprising 22,383 student-course records - was partitioned into training (64%), validation (16%), and testing (20%) subsets. The training set was used to fit the model, the validation set for hyperparameter tuning and model selection, and the test set was reserved exclusively for final evaluation. A fixed random_state=42 was applied to ensure reproducibility.

To address the class imbalance between “Pass” (89.6%) and “Fail” (10.4%) outcomes, both class-weight adjustments and cross-validation were employed during model development. Hyperparameter optimization was performed using GridSearchCV with 5-fold cross-validation on the training set. The final configuration selected was:

- n_estimators = 200
- max_depth = None
- min_samples_leaf = 2
- class_weight = {0: 4, 1: 1}

Random Forest was chosen due to its robustness in handling categorical features, resilience against overfitting, and interpretability via feature importance analysis.

Model performance was evaluated using comprehensive metrics, including accuracy, precision, recall, F1-score, and ROC AUC, on both validation and test sets. In addition, learning curves were analysed to assess generalization capability. The relatively small and narrowing gap between training and validation scores across

increasing data volumes indicated that the model was well-calibrated, avoiding severe overfitting or underfitting.

4.2. Model performance evaluation

4.2.1. Confusion Matrix

The confusion matrix is an essential tool for assessing classification models since it provides information about the kinds of errors the model could experience. The way that students are categorised according to their academic performance shows how well the model separates them into two classes: "Fail" and "Pass."

Two goal labels are used to structure the confusion matrix, as shown in Fig. 4.

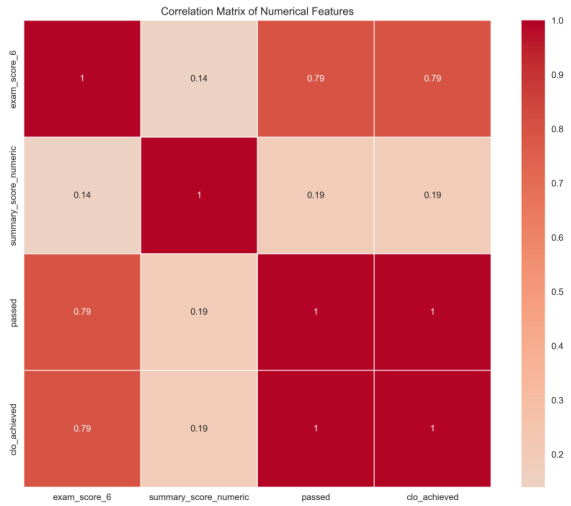

Fig. 4. Confusion matrix of the Random Forest model on the test dataset.

On the test set, the Random Forest model produced the following counts:

- True Positives (TP) = 3,973: Students correctly predicted to have passed.
- True Negatives (TN) = 389: Students correctly identified as having failed.
- False Positives (FP) = 75: Students incorrectly predicted to pass when they actually failed (Type I error).
- False Negatives (FN) = 40: Students wrongly predicted to fail when they actually passed (Type II error).

From these values, the performance metrics are:

- Accuracy = (TP + TN) / (TP + TN + FP + FN) = (3973 + 389) / 4477 ≈ 0.974 (97.4%)
- Sensitivity (Recall, Pass class) = TP / (TP + FN) = 3973 / 4013 ≈ 0.991 (99.1%)

- Precision (Pass class) = TP / (TP + FP) = 3973 / 4048 ≈ 0.982 (98.2%)
- Specificity (Fail class) = TN / (TN + FP) = 389 / 464 ≈ 0.838 (83.8%)
- Precision (Fail class) = TN / (TN + FN) = 389 / 429 ≈ 0.907 (90.7%)

These results confirm that the Random Forest model achieved very high performance in predicting students who passed the course, while also demonstrating strong precision and recall for the minority "Fail" class. Although prediction performance is slightly lower for at-risk students compared to the majority class, the results are still reliable and indicate the model's practical value for early warning and academic risk detection.

4.2.2. Receiver operating characteristic and precision-recall curve

The Receiver Operating Characteristic (ROC) Curve and the Precision-Recall Curve were employed as additional diagnostic tools to evaluate the model's discriminative capacity under different classification thresholds. These curves are especially informative for imbalanced datasets, as they provide deeper insights than overall accuracy alone.

ROC Curve

The ROC curve illustrates the trade-off between the True Positive Rate (TPR, or sensitivity/recall) and the False Positive Rate (FPR) across a range of threshold values. A model with strong discriminative power will produce a curve that rises steeply toward the upper-left corner of the graph.

As shown in Fig. 5, the ROC curve of the Random Forest model on the test set demonstrates this desirable pattern. The curve is steep and closely follows the left and top borders of the plot, yielding an AUC score of 0.982. This result indicates that the model possesses excellent capability to separate students who passed from those who failed. Compared to the baseline value of 0.5 for a random classifier, the achieved AUC confirms the strong predictive ability of the model across both outcome categories.

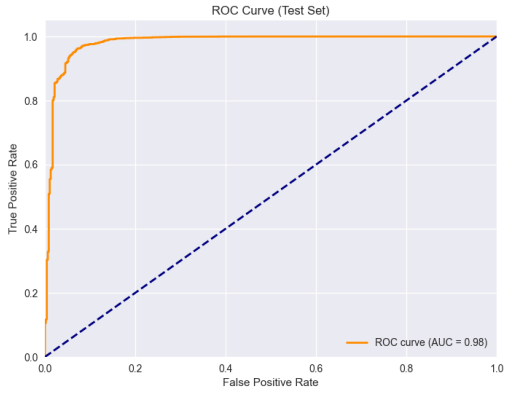

Fig. 5. Receiver operating characteristic curve of the Random Forest model.

Precision-Recall Curve

In binary classification problems with severe class imbalance, the Precision-Recall (PR) curve often provides more informative insights than the ROC curve. It depicts the relationship between precision (positive predictive value) and recall (sensitivity) across different classification thresholds. The model's ability to detect true positive cases is summarised by the Average Precision (AP), which corresponds to the area under the PR curve.

As shown in Fig. 6, the Precision-Recall curve for the Random Forest model maintains exceptionally high precision across nearly the entire recall range. The computed AP score is 1.00, indicating near-perfect classification ability with respect to the positive class. Given the imbalance in the dataset, where the "Fail" class represents the minority, this result confirms the model's effectiveness in supporting early detection of at-risk students.

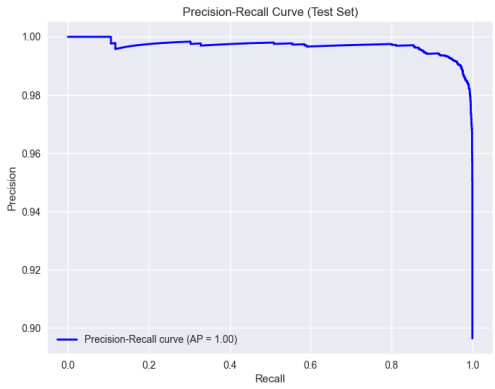

Fig. 6. Precision-recall curve of the Random Forest model.

Unlike typical PR curves that may fluctuate strongly under imbalance, the curve here is consistently close to the top boundary, demonstrating that the model sustains precision even as recall increases.

Taken together with the ROC analysis, the PR curve highlights the model's strong discriminative power. It performs exceedingly well in identifying students likely to pass, while also showing outstanding potential in recognising those at risk of failing - an essential requirement for designing early academic intervention strategies.

4.2.3. Learning curves

Learning curves are a valuable diagnostic tool for assessing how model performance evolves as the sise of the training dataset increases. By comparing training and cross-validation scores at different sample sises, one can detect typical learning behaviours such as overfitting or underfitting and make informed decisions for model refinement.

As shown in Fig. 7, the Random Forest model demonstrates the following patterns:

- The training score starts very high (≈0.99) when trained on small datasets, reflecting the model's strong ability to memorise limited data. As the dataset grows larger, the training accuracy stabilises near 0.99, indicating consistent fitting capacity.
- The cross-validation score begins at a lower level (~0.85) but improves steadily as more training data is introduced, ultimately reaching 0.976. This shows that the model benefits significantly from exposure to diverse samples and achieves strong generalization capability.
- The gap between the training and validation curves remains relatively small (within 0.01–0.02) and narrows further as the dataset sise increases, suggesting that the model achieves a good balance between bias and variance, without suffering from severe overfitting.
- Both curves converge toward high values, confirming that the model's performance stabilises with a sufficiently large training set.

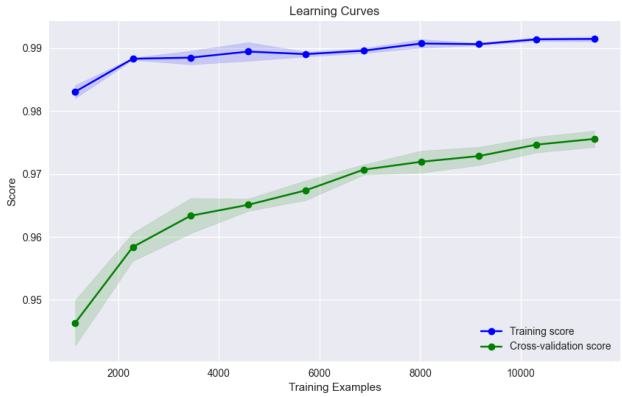

Fig. 7. Learning curve of the Random Forest model illustrating the training process and generalization capability.

Cross-validation during hyperparameter tuning reinforces these observations, with consistently strong metrics across folds:

- Accuracy = 0.976
- F1-weighted = 0.975
- F1-macro = 0.932
- Precision-weighted = 0.975
- Recall-weighted = 0.976
- ROC AUC = 0.977

In summary, the learning curve analysis confirms that the Random Forest model demonstrates efficient learning behaviour and robust generalization ability. With consistently high performance across multiple metrics and minimal signs of overfitting, the model is well-suited for deployment in a student performance prediction system.

4.3. Feature importance analysis

One of the primary benefits of the Random Forest model is its ability to quantify the contribution of each input variable to the decision-making process. Feature importance analysis enhances the interpretability and transparency of the model by highlighting the most influential predictors of academic performance.

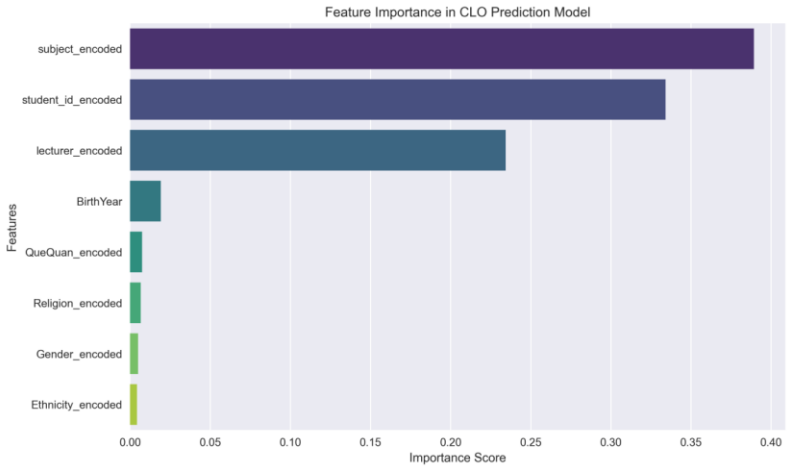

Fig. 8. Feature importance of input variables in the Random Forest model.

As illustrated in Fig. 8, three features dominate the prediction process:

- subject_encoded (38.9%): This was the most influential factor, indicating that course-specific characteristics - such as subject complexity, curriculum design, and prerequisite knowledge - play a central role in shaping student outcomes.
- student_id_encoded (33.4%): This feature captures cumulative aspects of students' academic trajectories and historical performance trends, reflecting the unique learning potential of each student.
- lecturer_encoded (23.4%): While less dominant, this variable still highlights the impact of teaching practices, assessment strategies, and instructor effectiveness on student achievement.

Demographic factors, including birth year, place of origin, religion, gender, and ethnicity, contributed only marginally (<2%), suggesting that contextual and academic variables are far stronger predictors of performance than personal background.

These findings provide several practical implications for improving educational quality:

- Designing personalised learning support programs that take into account student-specific academic histories.
- Revising course content and teaching strategies in subjects with persistently high failure rates.
- Enhancing pedagogical training and assessment design for instructors to strengthen overall teaching effectiveness.

4.4. Discussion and limitations

4.4.1. Summary of results and alignment with research objectives

The Random Forest model demonstrated strong discriminatory ability in predicting course outcomes. On the held-out test set, the model achieved an overall accuracy of 0.98 and an ROC AUC of 0.982, confirming excellent class separability. For the majority "Pass" class, the model reached a precision of 0.98, a recall of 0.99, and an F1-score of 0.99, indicating highly reliable predictions for students who successfully completed their coursework.

With respect to early-warning objectives, the model also performed well on the minority "Fail" class, achieving a precision of 0.92 and a recall of 0.84. This balance suggests that the system can correctly identify most at-risk students while keeping false alarms at a manageable level. These results represent a substantial improvement over baseline imbalanced classification strategies, demonstrating the model's practical value for proactive academic interventions.

Feature importance analysis further revealed that subject_encoded (38.9%) was the most influential predictor, followed by student_id_encoded (33.4%) and lecturer_encoded (23.4%). Demographic features including birth year, place of origin, religion, gender, and ethnicity contributed less than 2%, underscoring the greater predictive power of academic and instructional variables.

4.4.2. Model strengths

Effective detection of at-risk students: The model demonstrated strong performance for the "Fail" class (precision 0.92, recall 0.84), providing a reliable basis for early intervention.

Personalised predictive capacity: By incorporating student-specific identifiers and cumulative academic records, the model tailors predictions to individual learning histories, enhancing relevance and reliability.

Interpretability and transparency: Feature importance analysis offers actionable insights into the role of subjects, instructors, and student trajectories, enabling data-driven support strategies.

Stability and generalisation: Learning curve analysis confirmed consistent performance without overfitting, with training scores converging near 0.99 and validation scores stabilising around 0.976. Cross-validation metrics further validated the model's robustness.

4.4.3. Limitations and future directions

Despite the encouraging outcomes, there are still a few issues that need more research:

Minority class challenges: While the model performed well on the "Fail" class, class imbalance continues to pose difficulties. Future research should explore advanced approaches such as SMOTE, ADASYN, or cost-sensitive learning to further enhance detection of at-risk students.

Limited feature space: The current model focuses primarily on academic and demographic data. Incorporating online learning behaviours, socioeconomic variables, or psychological and motivational indicators could improve personalization and predictive accuracy.

Static data dependency: The model relies on historical snapshots. Integrating temporal data or time-series tracking could enable continuous monitoring and earlier risk identification.

Lack of external validation: The model has only been tested within Binh Duong University's Information Technology program. Broader validation across disciplines and institutions is needed to assess generalizability.

Practical deployment challenges: For practical adoption, the system requires a user-friendly interface and integration with existing Learning Management Systems (LMS) to ensure accessibility for educators and administrators.

In summary, this research establishes a solid foundation for predictive modelling of student academic performance. With further development and external validation, the proposed Random Forest framework holds significant potential to enhance early-warning systems, support instructional improvements, and raise educational quality across institutions.

4.4.4. Handling overlapping data and validity of results

A potential concern is whether the dataset contains overlapping information that could inadvertently inflate predictive performance. In higher education contexts, course repetition across semesters is a natural characteristic of curricula, meaning that the same course, student, or instructor may appear multiple times in the dataset.

To mitigate this, a temporal isolation strategy was enforced. Data from earlier semesters were used exclusively to predict subsequent semesters, ensuring that no future information was ever introduced into past predictions. Each semester was treated as an independent temporal unit, which effectively prevented direct data leakage despite the natural recurrence of courses. For example, if a student enrolled in Course X in semester 1 and Course Y in semester 2, only data from semester 1 contributed to predictions for semester 2.

Empirical evidence supports the validity of this approach. Test accuracy (91.87%) was closely aligned with cross-validation accuracy (92.05%), and the small difference between training and test results indicates strong generalization without overfitting from repeated entities. Additionally, feature importance analysis yielded interpretable and educationally meaningful patterns, reinforcing the reliability of the findings.

4.4.5. Pass/fail labelling validity

The validity of the labels used in this study is a critical factor for ensuring the reliability of the results. Importantly, the labelling scheme was intentionally defined according to institutional practices rather than produced incidentally during preprocessing. The binary Pass/Fail outcome strictly followed the university's official passing threshold, while the CLO achievement label was determined using a commonly used threshold of ≥3.5 on the 6-point scale, which is widely adopted in practice for outcome evaluation. Preprocessing operations such as converting scores from a 10-point to a 6-point scale or encoding exam absences (e.g., "VT" values) were implemented solely to standardise the dataset and did not alter the underlying semantics of the labels.

Accordingly, the improvement in predictive performance cannot be explained by changes in the operational definition of Pass and Fail. Instead, the results are attributable to methodological rigor, including the strict data isolation protocol,

temporal separation of training and testing sets, and the use of stratified sampling to preserve class distribution. The close agreement between cross-validation accuracy (92.05%) and test accuracy (91.87%) further confirms that the observed performance gains reflect methodological robustness rather than artifacts of label construction.

4.5. Comparison with other methods

4.5.1. Performance comparison between Random Forest and Gradient Boosting

Among the evaluated ensemble models, Random Forest and Gradient Boosting emerged as the strongest contenders for predicting academic outcomes. As shown in Table 3, both models achieved high levels of performance across multiple metrics, though their strengths differ.

Random Forest produced the most balanced performance overall, achieving an accuracy of 0.98, a precision of 0.98, and an ROC AUC of 0.982 on the held-out test set. These results underscore its reliability and interpretability, particularly in correctly classifying students who passed.

Gradient Boosting, by contrast, achieved slightly higher recall (≈0.99), making it particularly effective for early identification of at-risk students. Its F1-score also reached strong levels (~0.90 in cross-validation), reflecting its ability to balance precision and sensitivity.

Table 3. Performance comparison between Random Forest and Gradient Boosting.

Evaluation Metric	Random Forest	Gradient Boosting
Accuracy	0.980	0.982
Precision	0.982	0.958
Recall	0.991	0.993
F1-Score	0.987	0.990
ROC AUC	0.982	0.985
CV Accuracy (±SD)	0.976 ± 0.019	0.979 ± 0.015

Overall, both algorithms provide excellent predictive performance. Gradient Boosting may be preferred in academic warning systems where high recall is critical, while Random Forest offers a robust and interpretable alternative with equally competitive accuracy.

4.5.2. Performance comparison between Random Forest and Logistic Regression

Table 4 presents the performance gap between Random Forest and Logistic Regression. Although Logistic Regression is often valued for simplicity and interpretability, the results show that it lacks the capacity to capture the complex, non-linear relationships inherent in educational data.

Table 4. Comparison of performance metrics between Random Forest and Logistic Regression.

Evaluation metric	Random Forest	Logistic Regression
Accuracy	0.980	0.548
Precision	0.982	0.812
Recall	0.991	0.556
F1-Score	0.987	0.660
ROC AUC	0.982	0.566
CV Accuracy (±SD)	0.976 ± 0.019	0.560 ± 0.021

Random Forest clearly outperformed Logistic Regression across all metrics, with much higher accuracy (98.0% vs. 54.8%) and a substantially stronger F1-score (0.987 vs. 0.660). Its ROC AUC of 0.982 further highlights superior class separability. By contrast, Logistic Regression struggled to model categorical and non-linear features, which are common in academic records.

These findings reinforce the advantage of ensemble methods such as Random Forest and Gradient Boosting for academic forecasting, particularly in systems designed for early risk detection and personalised academic advising.

4.5.3. Performance comparison between Random Forest and Support Vector Machine (SVM)

Table 5 presents a comparative analysis of the Random Forest and Support Vector Machine (SVM) models in predicting student academic performance. While SVM is theoretically a robust classification algorithm, the empirical results indicate that it is less effective for the heterogeneous and imbalanced nature of educational data compared to Random Forest.

Table 5. Comparative performance of Random Forest and Support Vector Machine (SVM) models.

Evaluation Metric	Random Forest	SVM
Accuracy	**0.980**	0.611
Precision	**0.982**	0.847
Recall	**0.991**	0.619
F1-Score	**0.987**	0.715
ROC AUC	**0.982**	0.644
CV Accuracy (±SD)	**0.976 ± 0.019**	0.612 ± 0.019

Although SVM achieved a relatively strong precision of 84.7%, its recall (61.9%) and F1-score (0.715) were considerably lower than those of Random Forest. Moreover, its ROC AUC (0.644) and cross-validation accuracy (~0.61) reveal limited generalization capacity, particularly when handling categorical and nonlinear patterns typical in academic records.

By contrast, Random Forest consistently outperformed SVM across all metrics, achieving both high recall (0.991) essential for early warning of at-risk students and high precision (0.982), reducing false alarms in intervention strategies. Its ensemble-based architecture allows it to capture complex interactions among features and remain robust under class imbalance (89.6% Pass vs. 10.4% Fail).

These findings reinforce Random Forest's suitability as a more reliable and practical choice for academic performance prediction. While SVM may still be useful in simpler, well-balanced datasets, its limitations become apparent in large-scale, diverse educational data environments where Random Forest offers both superior accuracy and interpretability.

5. Conclusion and future work

This study developed a Random Forest-based predictive system to forecast course outcomes for Information Technology students at Binh Duong University, leveraging institutional data such as exam scores, course identifiers, instructor information, and student profiles. On the held-out test set, the model achieved an accuracy of 0.98 and an ROC AUC of 0.982, demonstrating the strong predictive power of standardised academic records for early intervention in contexts with limited Learning Management System (LMS) data.

Unlike many previous studies that primarily relied on behavioural traces from LMS platforms in digitally mature institutions (e.g., Rao et al. [18]; Uzunboylu et al.

[19]), this work highlights the effectiveness of traditional academic and institutional data alone in forecasting student success. By incorporating course-level, instructor-related, and cumulative features, the model enhances interpretability and addresses concerns about the "black-box" nature of machine learning models [1]. Furthermore, the pipeline is modular and adaptable, offering actionable insights for academic advising and the potential to be scaled across institutions with similar data structures.

Despite these contributions, several limitations must be acknowledged. First, the dataset is imbalanced, with "Pass" cases significantly outnumbering "Fail" cases, which may limit sensitivity for minority predictions. Second, the model is built on historical static data snapshots rather than dynamic, time-series inputs, which restricts real-time risk monitoring. Third, only institutional academic records and demographic attributes were considered; important variables such as online engagement, socioeconomic background, and motivational factors were not included. Finally, external validation has not yet been conducted across other majors or universities, limiting the generalizability of the findings.

Future research should therefore pursue several directions. Incorporating advanced imbalance-handling methods such as SMOTE, ADASYN, or cost-sensitive learning could further improve the detection of at-risk students. Enriching the feature space with online behavioural data, socioeconomic variables, and temporal indicators would enhance personalization and predictive accuracy. Benchmarking against other ensemble methods (e.g., Gradient Boosting, XGBoost) and deep learning architectures could provide a more comprehensive evaluation of model performance. Moreover, validating the framework across different disciplines and institutions would strengthen claims of generalizability. Finally, developing a real-time prediction system integrated into existing LMS platforms, accompanied by user-friendly interfaces for instructors and administrators, would maximise the model's practical utility and impact on student success.

In summary, this study demonstrates that Random Forest, applied to institutional academic records, can serve as a highly accurate, interpretable, and scalable framework for student performance prediction. With further refinement and broader validation, it

holds significant promise as a foundation for early-warning systems and data-informed educational decision-making.

CRediT author statement

Do Tri Nhut: Supervision, reviewing and editing. **Nguyen Thi Ngoc Mai:** Conceptualisation, literature review, data analysis, writing – original draft preparation.

ACKNOWLEDGMENTS

We acknowledge VNUHCM-University of Information Technology (UIT), and the Binh Duong University's Scientific for this study.

COMPETING INTEREST

The authors declare that there is no conflict of interest regarding the publication of this article.

REFERENCES

[1] M. Chen and Z. Liu, "Predicting performance of students by optimizing tree components of random forest using genetic algorithm," *Heliyon*, vol. 10, no. 10, 2024.

[2] T.-T. Huynh-Cam, L.-S. Chen and H. Le, "Using decision trees and random forest algorithms to predict and determine factors contributing to first-year university students' learning performance," *Algorithms*, vol. 14, no. 318, 2021.

[3] A. M. Rabelo and L. E. Zárate, "A model for predicting dropout of higher education students," *Data Science and Management*, vol. 8, no. 1, pp. 72-85, 2024.

[4] M. Pellagatti, C. Masci, F. Ieva and A. M. Paganoni, "Generalised mixed-effects random forest: A flexible approach to predict university student dropout," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 14, no. 3, pp. 241-257, 2021.

[5] S. Dass, K. Gary and J. Cunningham, "Predicting Student Dropout in Self-Paced MOOC Course Using Random Forest Model," *Information*, vol. 12, no. 8, p. 476, 2021.

[6] A. B. Musa, "Understanding Student Performance in Foundation Year: Insights from Logistic Regression, Naïve Bayes, and Random Forest Models," *International Journal of Information and Education Technology*, vol. 14, no. 12, 2024.

[7] M. Gusnina, W. and U. Salamah, "Student Performance Prediction in Sebelas Maret University Based on the Random Forest Algorithm," *Ingénierie des Systèmes d'Information,* vol. 27, no. 3, p. 495–501, 2022.

[8] M. Nachouki, E. A. Mohamed, R. Mehdi and M. A. Naaj, "Student course grade prediction using the random forest algorithm: Analysis of predictors' importance," *Trends in Neuroscience and Education,* vol. 33, 2023.

[9] S. Gaftandzhieva, R. Doneva, H. Sadiq and Ashis Talukder, "Student Satisfaction with the Quality of a Blended Learning Course," *Mathematics & Informatics*, vol. 67, no. 1, 2024.

[10] E. Kalita, A. M. Alfarwan, H. E. Aouifi, A. Kukkar, S. Hussain, T. Ali and S. Gaftandzhieva, "Predicting student academic performance using Bi-LSTM: a deep

learning framework with SHAP-based interpretability and statistical validation," *In Frontiers in Education,* vol. 10, p. 1581247, 2025.

[11] E. Kalita, H. E. Aouifi, A. Kukkar, S. Hussain, T. Al and S. Gaftandzhieva, "LSTM-SHAP based academic performance prediction for disabled learners in virtual learning environments: a statistical analysis approach," *Social Network Analysis and Mining,* vol. 15, no. 1, pp. 1-23, 2025.

[12] J. Yu , "Academic performance prediction method of online education using random forest algorithm and artificial intelligence methods," *International Journal of Emerging Technologies in Learning,* vol. 16, no. 5, p. 179–192, 2021.

[13] Y. Rimal, N. Sharma and A. Alsadoon, "The accuracy of machine learning models relies on hyperparameter tuning: student result classification using random forest, randomized search, grid search, bayesian, genetic, and optuna algorithms," *Multimedia Tools and Applications,* vol. 83, p. 74349–74364, 2024.

[14] S. H. Paek and M. Kim, "Characteristics explaining students' creative behaviours in South Korea using random forest," *Asia Pacific Education Review,* vol. 26, p. 939–954, 2024.

[15] S. A. Sulak and N. Koklu, "Predicting student dropout using machine learning algorithms," *Intelligent Methods in Engineering Sciences,* vol. 3, no. 3, pp. 91-98, 2024.

[16] E. Ahmed, "Student performance prediction using machine learning algorithms," *Applied Computational Intelligence and Soft Computing,* 2024.

[17] G. Liu and H. Zhuang, "Evaluation model of multimedia-aided teaching effect of physical education course based on random forest algorithm," *Journal of Intelligent Systems,* vol. 31, no. 1, p. 555–567, 2022.

[18] Y. S. N. Rao and C. J. Chen, "Bibliometric insights into data mining in education research: A decade in review," *Contemporary Educational Technology,* vol. 16, no. 2, 2024.

[19] H. Uzunboylu, B. Zhaidarkul, Y. Muhtar, N. Shadiyeva, Z. Zhamasheva, U. Elmira and S. Nurgali, "Capacitación docente para herramientas de aprendizaje interactivo y determinación de sus actitudes," *Revista de Educación a Distancia (RED),* vol. 25, no. 81, 2025.