Techno-Science Research Journal V14 (Special Issue) (2026) P 90-102

Content list available at ITC

Techno-Science Research Journal

Journal Homepage: http://techno-srj.itc.edu.kh/

Techno-Science Research Journal

A Recall-Oriented Learning Analytics Framework for Early Identification of At-Risk Students in Blended Learning Environments Toward Sustainable Education

Chentra Pen[1*], Sokkhey Phauk[1], Sothea Has[1], Dona Valy[2]

[1] *Department of Applied Mathematics and Statistics, Institute of Technology of Cambodia, Russian Federation Blvd., P.O. Box 86, Phnom Penh, Cambodia.*

[2] *Department of Information and Communication Engineering, Institute of Technology of Cambodia, Russian Federation Blvd., P.O. Box 86, Phnom Penh, Cambodia*

Received: 06 April 2026; Revised: 07 July 2026; Accepted: 14 July 2026; Available online: 31 July 2026

Abstract: *Enhancing students' academic outcomes in blended learning (BL) environments requires timely intervention for at-risk students, making early detection a critical task for educators and researchers. However, this task remains challenging due to imbalanced educational data and limited model sensitivity. This study investigated the application of machine learning (ML) for the early identification of at-risk students in BL environments, aiming to improve educational quality in alignment with Sustainable Development Goal 4 (SDG 4). A dataset of 342 students, collected from a BL environment at the Institute of Technology of Cambodia (ITC), was used to model academic performance as a three-class classification problem: At-risk, Good, and Excellent. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was employed, and multiple models including Random Forest, Gradient Boosting, and ensemble methods were evaluated using stratified cross-validation. The best multi-class performance was achieved by Gradient Boosting with SMOTE, yielding an accuracy of 75.43% and a macro F1-score of 68.74%. Nevertheless, this model detected at-risk students only 41.67% of the time. To resolve this limitation, a binary classification approach with a focus on recall was proposed. By reframing the problem as an early-warning task and lowering the classification threshold from 0.50 to 0.10, recall improve dramatically from 52.78% to 91.67%, allowing for highly sensitive detection of at-risk students. However, this improvement was at the cost of lower precision with more false positives. To overcome this trade-off, we further evaluated the F2-score and threshold analysis and identified a threshold of around 0.40 as a better trade-off between recall and precision in real educational intervention scenarios. These results indicate that threshold optimization is an essential technique for educational early-warning systems, where recall is more important than overall accuracy. The study offers a practical and scalable framework for improving student monitoring, enabling timely intervention, and enhancing learning outcomes in blended learning environments.*

Keywords: Academic Performance Prediction; Blended Learning; Early Warning Systems; Learning analytics; Machine Learning

1. INTRODUCTION

1.1. Background

BL has evolved from a supplementary method into a dominant paradigm for teaching and learning in higher education (HE). This model effectively integrates traditional face-to-face physical classroom with more flexible online platforms such as Moodle [1,2]. This transition has been accelerated not only by the COVID-19 pandemic [3], but also by broader global trends including digital transformation, economic pressures, and increasing costs associated with physical access to education. In many developing countries like Cambodia, rising living costs, including transportation and fuel expenses, create additional barriers for students to attend physical classes regularly. These challenges highlight the importance of a more flexible learning environment that reduces dependency on face-to-face classroom presence while maintaining educational quality at the same time. As a result, BL has emerged as a practical solution to support accessibility, continuity, and inclusiveness in education.

[*] Corresponding author: Chentra PEN
E-mail: pen.chentra@itc.edu.kh

1.2. Harnessing learning analytics for academic success

While BL is increasingly important in HE, Learning Management System (LMS) becomes indispensable part of the implementation. The system captures and stores large volumes of the student interaction data, including login frequency, learning behaviors, assessment performance etc. This information is significantly useful for educators to understand the students' learning, behaviors and their engagement by integrating learning analytics technology and framework into the system. Learning analytics enables institutions to transform these data into actionable insights to support teaching and learning processes and build an intelligent system [4].

One of the main purposes of the integrated learning analytics is to build an effective Early Warning Systems (EWS) intended to improve student retention and academic success. Recent research in educational data mining(EDM) indicates that ML models are able to predict student performance by utilizing both academic and behavioral data [5, 6]. Nevertheless, a lot of the current predictive approaches are mostly concerned with getting the most accurate results overall, which may be very misleading in educational environments. Since "at-risk" students are usually only a small percentage of a group, a model can be 90% accurate just by guessing that all students will pass. This means that it will not identify the 10% who really need support. This class imbalance results in standard algorithms fail to work in the same way that they should in the areas where they are most needed. Recent studies highlight the ethical and practical demand for EWS that prioritize Recall over Accuracy [7]. In a demanding setting such as an EWS, the cost of a "False Negative" (missing a student who does not succeed) is far higher than the cost of a "False Positive" (offering extra help to a student who probably passed anyway). In Cambodia's developing education system, there is a critical need for local systems that integrate these recall-oriented methods in order to help with effective decision-making [8,9].

1.3. Significance of the study

This study contributes an in-depth technical framework for EDM that prioritizes highly sensitive detection over raw accuracy, ensuring the systematic identification of at-risk students within imbalanced BL datasets. Specifically, the framework addresses the accuracy paradox by placing greater weight on the cost of missing at-risk students than on the cost of false alarms; it demonstrates that adjusting the decision boundary beyond the default 0.50 threshold yields a flexible and robust detection system that institutions with varying levels of capability can adopt; and it produces a more stable representation of student risk by combining LMS behavioral logs with academic metrics through SMOTE-based balancing and ensemble architectures.

2. RELATED WORK

2.1. Learning Analytics and Student Performance Prediction

To understand and predict student performance in digital learning environments, EDM and learning analytics have been frequently employed. The earlier work of Romero& Ventura [5] offered a thorough analysis of data mining techniques in education, highlighting their potential to enhance learning outcomes. Additionally, Macfadyen and Dawson [10] provided empirical evidence that students' activity data from LMS platform can be utilized for predicting academic success, highlighting the significance of students' behavioral engagement data as primary features for any predictive modeling tasks.

For classification tasks involving complex student data, for example detecting at-risk students where data is highly imbalance and may contain non-linear relationship, more robust ML algorithms provide strong performance. Empirically, ensemble methods such as Random Forest(RF) [11] and Gradient Boosting(GB) [12] have become the gold standard for complex dataset. Unlike single-tree models, these "forest" and "boosting" methods are capable of capturing non-linear interactions between diverse features. However, the effectiveness of these models is frequently weakened by the previously mentioned class imbalance. For instance, when it comes to finding students who are at risk, where the students' group usually makes up a minority proportion, often results in these models to fall apart. To mitigate this, the Synthetic Minority Over-sampling Technique (SMOTE) was developed as an advanced way to generate more data [13]. Instead of just copying existing "at-risk" records, SMOTE generates new, synthetic examples in the feature space. This makes the model learn a more nuanced decision boundary for the minority class, which significantly enhances its ability to detect "at-risk" patterns in actual data.

The setting of the "decision threshold" is as significant as the selection of algorithm. To avoid at-risk students from being neglected, Aldowah [7] highlighted the need for predictive models that focus on recall priority. However, traditional ML models use a default probability threshold of 0.50, which is not always the best choice for early detection. The system becomes more sensitive when the threshold is lowered, which means it may detect more students who are at risk even when the probability of failure is lower. Several studies have investigated learning analytics and digital transformation in Cambodian higher education, integrating advanced machine learning techniques, such as ensemble

learning and hybrid model optimization[8,9]; however, the use recall-oriented optimization strategies, specifically threshold tuning, remains limited. This study is intended to solve that gap by offering a technique that can be both technically effective and practical oriented aimed at ensuring no student is overlooked.

Table 1. Summary of Related Studies on Student Performance Prediction

Study	Data Source	Model / Approach	Imbalance Handling	Threshold Optimization	Research Focus	Research Gap
Romero & Ventura (2010) [5]	Educational datasets	Survey / Review	Not addressed	Not addressed	Educational Data Mining (EDM) overview	No predictive modelling framework
Macfadyen & Dawson (2010) [10]	LMS (Moodle)	Logistic Regression	Not addressed	Not addressed	Student engagement analytics	No treatment of class imbalance
Breiman (2001) [11]	General ML datasets	Random Forest	Not addressed	Not addressed	Ensemble learning methodology	Not specific to educational context
Chawla et al. (2002) [13]	General ML datasets	SMOTE	Applied	Not addressed	Class imbalance mitigation	No application to LMS or education domain
Aldowah et al. (2021) [7]	LMS + ML systems	Machine learning models	Partially addressed	Limited	Early warning systems	Lack of threshold optimization
Dhawan [14] (2020)	Online learning platforms	Conceptual study	Not addressed	Not addressed	Blended learning	No predictive modelling component
Phauk Sokkhey et al. (2020) [8]	Local educational context	Learning analytics	Limited	Not addressed	Cambodian education	Limited integration of advanced ML techniques
This study (proposed)	LMS+ academic Records (ITC,342)	RF, GB, soft-Voting Ensemble	SMOTE (in fold)	Applied (threshold= 0.10)	Recall-oriented EWS	Addresses all Three gaps.

2.2. Research Gap

In this study, the term "research gap" refers to a specific, unmet methodological needs in the existing literature, namely the absence of recall-oriented threshold optimization in early-warning models (EWM) built on imbalanced LMS data. In educational datasets from learning management systems (LMS), class imbalance continues to be a major issue, particularly in BL environments where at-risk students are a small minority and skew model evaluations in favor of the majority classes. The over-reliance on overall accuracy or weighted metrics, which hiding poor minority-class performance and limit early intervention for weak learners, is still identified by recent studies in 2025, while multimodal analytics integration is not sufficiently explored [16]. The specific gaps motivating are:

- Most studies prioritize accuracy over recall for at-risk detection.
- Few studies explore decision-threshold optimization as an explicit lever for recall.
- There is insufficient integration of academic and behavioral LMS in a single model.

To address these gaps, this study proposes a recall-oriented early warning framework for student performance prediction. While existing approaches employ ML and occasionally address class imbalance, they often fail to prioritize the detection of at-risk students. The proposed method integrates SMOTE with ensemble models (RF and

GB) and incorporates threshold tuning to enhance recall. Using real LMS and academic data from a BL context in Cambodia, this study provides a practical and context-aware solution for improving the early identification of at-risk students.

2.3. Research Objectives

This study aims to develop a technically rigorous early warning system (EWS) for BL environments in HE, addressing key limitations of conventional classification approaches through the following objectives:

Objective 1: Develop a learning analytics framework that integrates LMS behavioral data with academic performance data for comprehensive student modeling.

Objective 2: Enhance the detection of at-risk students through threshold optimization within a recall-oriented classification framework.

Objective 3: Evaluate model performance using stratified cross-validation and SMOTE to address class imbalance and ensure robust generalization.

Objective 4: Support data-driven early intervention strategies to improve student outcomes in blended learning environments.

3. METHODOLOGY

3.1 Dataset and Feature Description

The dataset consists of 342 student records collected from a blended learning environment at the Institute of Technology of Cambodia. The dataset integrates academic, behavioral (LMS-based), and contextual features, enabling a comprehensive representation of student performance and engagement.

The target variable is derived from the final semester score and formulated as a three-class classification problem and a binary at-risk detection task[5,8,17,18]. The grade boundaries were defined base on the institutional grading practice used at the Institute of Cambodia, where a score below 65 indicates a student requiring academic support. Accordingly:

- **At risk**: score < 65, having 36 students.
- **Good**: 65<= score < 80, containing 175 students
- **Excellent**: score >=80, consisting 131 students.

This categorization supports the objective of early-warning analysis by separating students who require academic support from those with satisfactory or strong performance. The same band structure is consistent with grade-band formulations adopted in prior educational data-mining studies on student performance[5,8,17,18].

The at-risk minority class has only 36 samples (about 10.5% of the dataset), which constitutes a major class-imbalance problem (Fig. 1). With standard 5-fold cross-validation, each validation fold contains only about seven at-risk students, so estimates of recall-sensitive metrics can be unstable. This was mitigated by using stratified 5-fold cross-validation to preserve the class distribution in every fold and by applying SMOTE oversampling only within the training folds, so that no synthetic samples leak into validation. The evaluation metrics are reported as mean ± standard deviation across the validation folds, providing a more robust and statistically reliable assessment of model performance, particularly for recall-oriented early-warning tasks on imbalanced educational data.

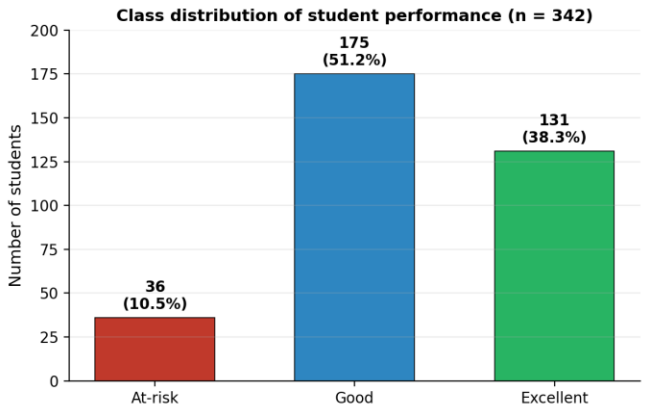

Fig. 1. Class distribution of student performance (n = 342).

Table 2 defines each feature used in the model, together with it type, measurement, and source. All variables were extracted either from institutional academic records and from Moodle LMS activity logs over the study semester.

Table 2. Description of Variables

Feature	Type	Description/Measurement	Source
Previous score	Numeric (0-100)	Previous academic score	Academic records
Midterm score	Numeric (0-100)	Mid-semester examination score	Academic records
Homework	Numeric (0–10)	homework score	Academic records

Number of logs	Integer count	Total LMS login events during the semester	Moodle LMS logs
Material views (Avg.)	Numeric	Average number of learning-material accesses per resource	Moodle LMS logs
Number of Absent	Integer count	Total recorded absences from face-to-face sessions	Attendance records
Sex	Categorical (M/F)	Student gender	Student background
Location	Categorical	Residential location category	Student background
Final grade (Target)	Categorical (3 classes)	At-risk / Good / Excellent (derived from final score)	Derived

It should be noted that the present dataset captures academic, behavioral and contextual dimensions but does not yet include indicators such as internet accessibility, device availability, or examination-integrity measures. These additional blended-learning factors are acknowledged as a limitation and are planned for inclusion in future data-collection rounds to further improve construct validity.

Let the dataset be denoted as

$$D = \{(x_i, y_i)\}_{i=1}^{n}, x_i \in \mathbb{R}^p, y_i \in Y$$

where $n$is the number of students, $p$is the number of features, $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ represents the feature vector of the i^{th}student, and y_i denotes the corresponding class label.

The target variable corresponds to student performance categories:

$$Y = \{\text{At-risk}, \text{Good}, \text{Excelllent}\}$$

The learning objective is to construct a predictive function:

$$f: \mathbb{R}^p \rightarrow Y,$$

such that

$$\hat{y}_i = f(x_i),$$

Where $\hat{y}_i$ is the predicted performance class of student i.

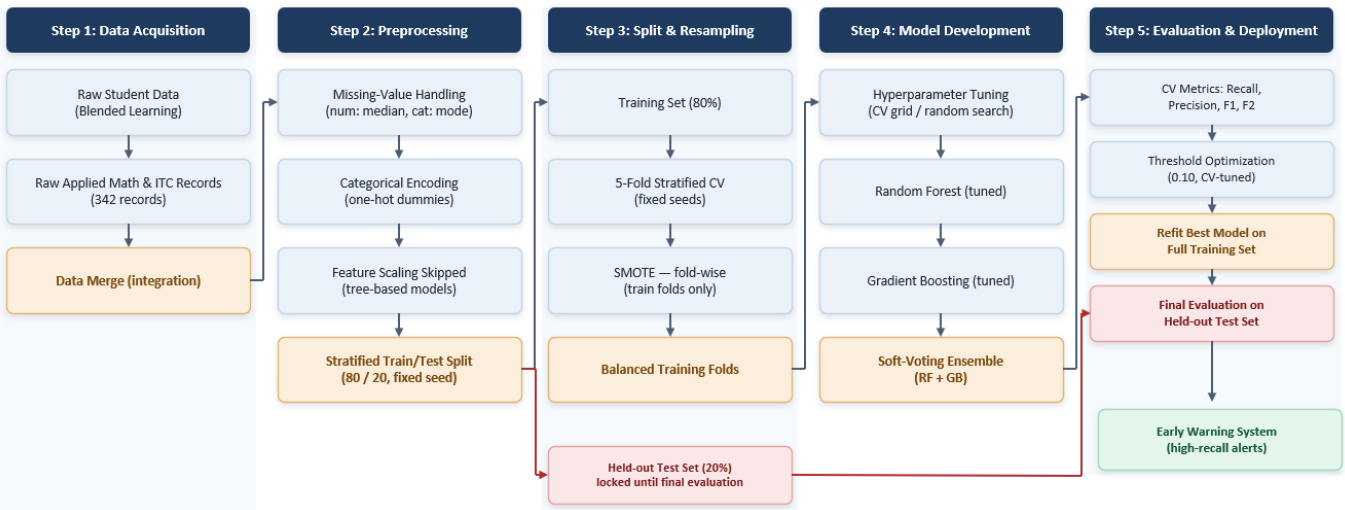

Fig. 2. Research workflow.

As illustrates, the two raw data sources each feed the data-merge step once, and the balanced training set feeds RF and GB, which act as parallel base learners combined by a soft-voting ensemble.

3.2 Data preprocessing

The data preparation is a central part of modelling pipeline. To ensure data quality and compatibility with ML models, the preprocessing stage in this study includes missing

value handling, categorical encoding, and consideration of feature scaling that the model required.

3.2.1 Missing value Handling

To address incomplete observations, different imputation strategies were applied according to the type of variable. For numerical features, missing values were replaced by the median of the corresponding feature, since the median is robust to extreme values and outliers [19].

3.2.2 Categorical Encoding

One-hot encoding transforms categorical variables such as gender and location. For variables with multiple groups (including a location with four regions), one-hot encoding turns each category into its own binary (dummy) variable. A vector of binary values with only one category active represents each observation. This method does not add artificial ordinal relationships, which allows models work with categorical data well [19]. Suppose a categorical variable C has Kpossible categories $C \in \{c_1, c_2, \ldots, c_K\}$. One-hot encoding transforms Cinto a binary vector

$$z = (z_1, z_2, \ldots, z_K),$$

where

$$z_k = \begin{cases} 1, \text{if } C = c_k \\ 0, \text{otherwise} \end{cases} \quad \text{(Eq. 01)}$$

3.3 Model Development

To build the prediction of students perform and at-risk learners, the study applied three ML techniques such as RF, GB, and an ensemble voting model base on their strong performance on structured educational data and their capacity to detect irregular relationships between behavioral, academic, and attendance characteristics.

3.3.1 Random Forest

Several decision trees are trained on various subsets of the data and their predictions are combined in Random Forest (RF), an ensemble learning technique based on the bagging (bootstrap aggregation) principle [11].

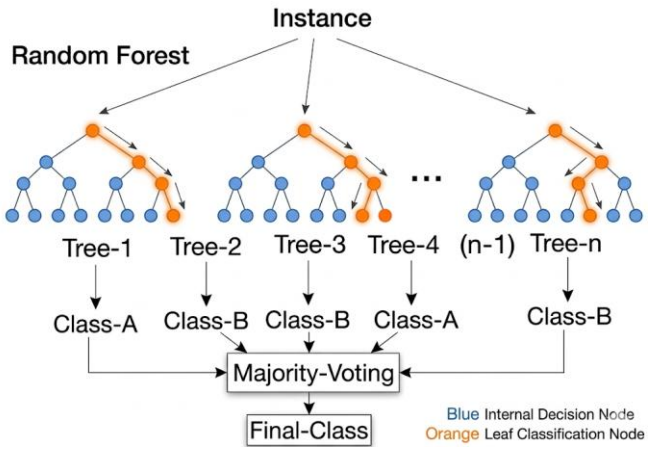

Fig. 3. Process of random forest

A random subset of features is utilized to build each tree, resulting in improved generalization and reduces correlation between trees. For classification tasks, a majority vote is employed for making the final prediction.

The strong points to consider are

- Ability to capture the nonlinear feature relationships
- Perform well to the noise and outliers
- Reduce risk of overfitting because of ensemble averaging
- Capability of provide feature importance measures.

These characteristics make Random Forest particularly suitable for educational datasets, in which student behavior and performance indicate complex relationships.

3.3.2 Gradient Boosting

Gradient Boosting (GB) builds models sequentially, with each new model seeking to correct the errors of its predecessors[12]. Unlike Random Forest, which constructs trees independently, GB uses iterative optimization to minimize a given loss function.

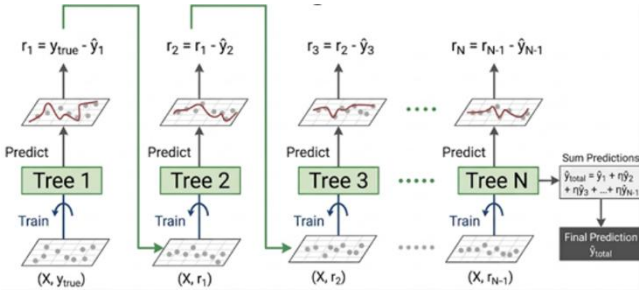

Fig. 4. Detail workflow of the sequential learning process in Gradient boosted tree.

At each iteration a new decision tree is fitted to the residual errors of the previous model; by focusing on hard-to-predict samples, the model gradually improves (Fig. 4).

Gradient Boosting (GB) offers high accuracy through sequential learning, the ability to model complex interactions, and flexibility in the choice of loss function. However, it is more sensitive to hyperparameters and requires careful tuning to avoid overfitting.

3.3.3 Ensemble voting Model

An ensemble voting model combines several base learners, here Random Forest and Gradient Boosting, to produce more accurate and reliable predictions. This study uses a soft-voting approach, in which the predicted class probabilities from each model are averaged to form the final prediction. This leverages the strengths of the individual models and lowers the variance associated with any single classifier[20](Fig.5).

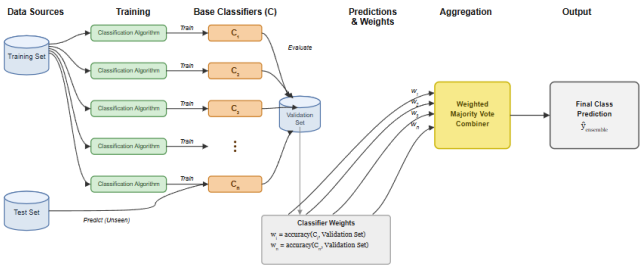

Fig. 5. Ensemble Voting Overview

In summary, by aggregating predictions across diverse base learners, the ensemble improves generalization, reduces model-specific bias and variance, and yields a more stable model that is less sensitive to noise, outliers, and variability in the data.

3.3.4 Model Selection Justification

Random Forest, Gradient Boosting, and the soft-voting ensemble were selected for their strong performance on structured educational data and their suitability for the challenges of this study, including heterogeneous features, class imbalance, and the need for interpretable and reliable predictions in early warning systems. Specifically, the selected models can handle mixed numerical and categorical variables, perform effectively in imbalanced settings when combined with balancing techniques, capture complex non-linear relationships among student behavior, engagement and performance, provide interpretable outputs such as feature importance, and improve generalization and robustness through ensemble learning.

3.4 Handling Class Imbalance

With only 36 of 342 students (10.5%) categorized as at-risk, the dataset is highly imbalanced, which can bias models toward the majority class and weaken detection of the minority class. The Synthetic Minority Over-sampling Technique (SMOTE) was used to address this issue [13]. SMOTE generates synthetic minority-class samples by interpolating along the line segments connecting a given minority instance to its nearest neighbors in the feature space. This improves class balance without simply replicating existing samples, reducing overfitting and improving generalization.

Crucially, SMOTE was applied only to the training data within each cross-validation fold, so that synthetic samples never influence the validation set. This prevents data leakage and yields unbiased performance estimates.

3.5 Evaluation Approach

This study uses Stratified K-Fold Cross-Validation with k = 5 to ensure reliable and objective evaluation[1]. To ensure responsibility, a fixed random seed was used for the fold splits, and, given the small number of at-risk cases (approximately 7 per fold), every experiment was repeated across multiple random seeds; the mean and standard deviation of each metric are reported (Results section 4.1)

The following evaluation metrics are used:

- Accuracy: measures the overall percentage of cases that are correctly classified.

$$Accuracy = \frac{TP+TN}{\mathrm{TP+TN+FP+FN}} \quad \text{(Eq. 02)}$$

- Precision: measures the percentage of expected positive cases that turn out to be positive.

$$Precision = \frac{TP}{\mathrm{TP+FP}} \quad \text{(Eq. 03)}$$

- Recall: measures the percentage of actual positive cases that are accurately identified.

$$Recall = \frac{TP}{\mathrm{TP+FN}} \quad \text{(Eq. 04)}$$

- F1-score: harmonic mean of precision and recall

$$F1 - score = 2\frac{(Precision)(Recall)}{\mathrm{Precision+Recall}} \quad \text{(Eq. 05)}$$

Because failing to identify at-risk students (false negatives) is more costly than false alarms, this study additionally reports the F2-score (Eq. 6), which weights recall more heavily than precision (β = 2) and is therefore a more

honest summary metric for an early-warning system than recall alone:

- $F2 - score = 5\frac{(Precision)(Recall)}{4 \text{ x Precision+Recall}}$ (Eq. 06)

3.6 Threshold Optimization

Most classification models assign labels using a default decision threshold of 0.50; a sample is predicted positive (at-risk) when $P(y = 1|x) \geq 0.50$, as in Eq. 07:

$$\hat{y} = \begin{cases} 1, & if\ P(y = 1|x) \geq 0.5 \\ 0, & otherwise \end{cases} \quad \text{(Eq. 07)}$$

In situations in which the objective is to achieve the the majority recall, nevertheless, this threshold may not be the best one.

In the study, the decision threshold r is adjusted to enhance identify the at-risk learner. It is defined as

$$\hat{y} = \begin{cases} 1, & if\ P(y = 1|x) \geq r \\ 0, & otherwise \end{cases} \quad \text{(Eq. 08)}$$

The ideal threshold is selected by maximizing recall while maintaining an acceptable level of precision after a range of threshold values is evaluated.

Concretely, the threshold was selected as the lowest value at which recall is maximized without precision collapsing below an institutionally acceptable level; the recall-weighted F2-score (Eq. 6) is used as the tie-breaking selection criterion, since it explicitly rewards recall while still penalizing excessive false positives.

This approach is aligned with the objective of early warning systems, where minimizing false negatives is critical.

4. RESULTS AND DISCUSSION

4.1 Multi-Class Classification Performance

Stratified cross-validation was used to summarize the performance of the proposed models for multi-class classification (At-risk, Good, Excellent). The best results were obtained by the tuned Gradient Boosting model with SMOTE (Table 3).

Table 3. Performance metrics of the multi-class classification model

Metrics	Mean ± SD across folds (%)
Accuracy	75.43 ± 5.52
Precision	70.30 ± 10.53
Recall	68.64 ± 6.27
F1-Score	68.74 ± 7.54

Each metric is reported as the mean ± standard deviation across the five stratified cross-validation folds, providing a statistically reliable assessment of model performance under class imbalance.

The performance differences between the candidate models were small. To test whether the difference between Gradient Boosting and the soft-voting ensemble is statistically meaningful, McNemar's test was applied to the pooled out-of-fold predictions and a paired t-test to the per-fold macro-F1 scores. Neither test found a significant difference (McNemar p = 1.00; paired t-test t = 0.79, p = 0.47; Wilcoxon p = 0.63), confirming that the two models perform comparably and that the prediction task is intrinsically challenging rather than that one model is decisively superior. Gradient Boosting is retained as the primary model on the basis of its slightly higher mean macro-F1 (0.6874) and lower variance, with the ensemble offering equivalent performance.

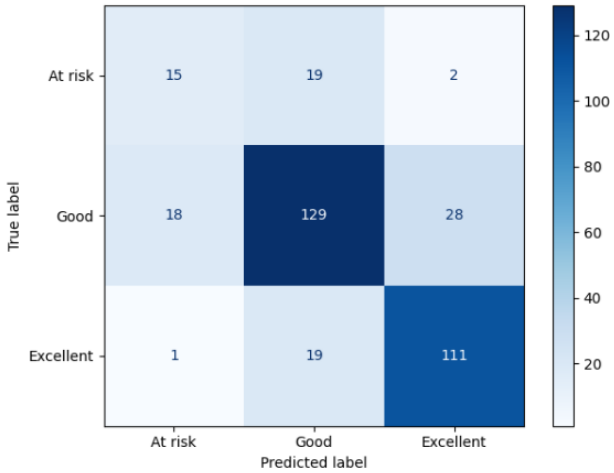

Fig. 6. Confusion matrix of classification model.

Figure 6 presents the confusion matrix of the multi-class classification model. The results indicate that the model performs well in classifying students in the "Good" and "Excellent" categories, with a high number of correctly predicted instances. However, the "At-risk" class remains difficult to identify, achieving a recall of only 41.67% (15 out of 36 cases). This indicates that more than half of the at-risk students are misclassified, highlighting the limitation of the

model in detecting the minority class and motivating the need for further improvement through threshold optimization.

4.2 Feature Importance Analysis

Figure 7 ranks the most influential variables under two complementary methods. Mean Decrease Impurity (MDI; Fig. 7a) identifies previous score as the dominant predictor, followed by number of absences, homework score, and total login activity, whereas demographic features such as gender and location contribute comparatively little. Because MDI is known to be biased toward continuous, high-cardinality features, permutation importance was additionally computed as a robustness check (Fig. 7b), scored with macro-F1 over 30 repeats.

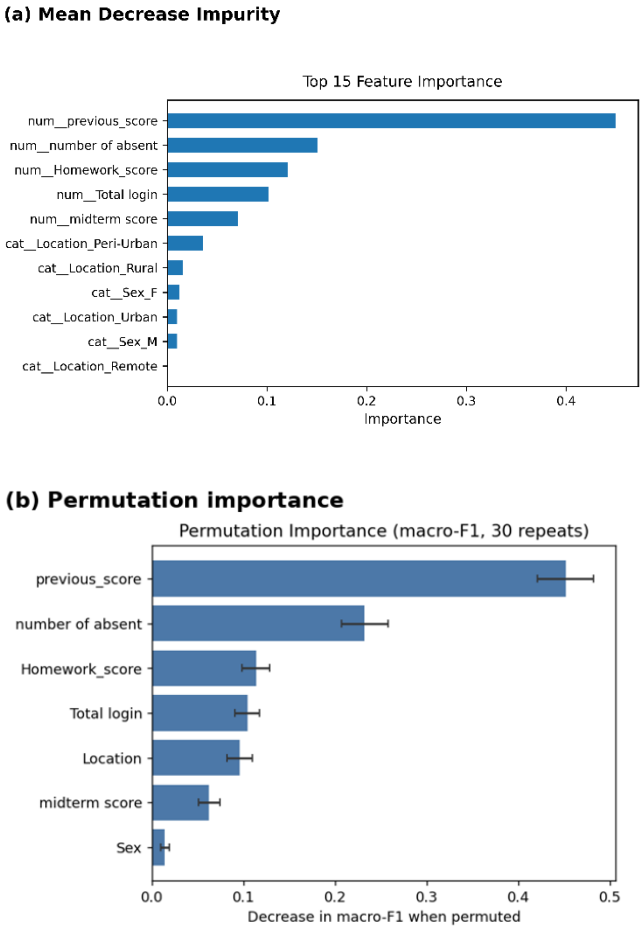

Fig. 7. Feature importance under two methods: (a) Mean Decrease Impurity (MDI); (b) permutation importance (macro-F1, 30 repeats).

The two methods are in close agreement since they identify the same four dominant predictors in the case previous score (0.45), number of absences (0.23), homework score (0.11) and total login (0.10) are in the same order, with only minor reordering among the lower-ranked variables (midterm score and location). This agreement confirms that the importance ranking is robust rather than an artefact of MDI's bias. Prior academic performance is by a wide margin the strongest predictor of at-risk status, followed by attendance (number of absences) and coursework engagement, indicating that both prior achievement and behavioral engagement contribute to the model.

4.3 Binary Classification (Early Warning System)

The problem is redefined as a binary classification task that improves the detection of at-risk students. The model achieves an accuracy of 86.55% using the default threshold of 0.50, however the recall for at-risk students stands at 52.78%, indicating insufficient level of sensitivity.

4.4 Performance with Default Threshold (0.50)

Table 4. Performance metrics of the binary classification model using the default threshold (0.50), reported as mean ± standard deviation across stratified cross-validation folds.

Metrics	Values (%)
Accuracy	86.55± 2.78
Precision	39.58± 6.02
Recall	52.78± 9.07
F1-Score	45.24± 4.98
F2-score	49.48± 6.48

With an overall accuracy of 86.55%, the model performs well overall. Nonetheless, the recall for the at-risk class is comparatively low (52.78%), indicating that almost half of the at-risk students are not accurately identified. The relatively higher standard deviation observed for recall (± 9.07%) reflects the instability expected when only about seven at-risk students fall in each validation fold. The recall-weighted F2-score of 49.48% confirms that, despite the high accuracy, the model is inadequate as an early-warning detector at this threshold.

Fig 8 above illustrates the 17 false negatives, in which students who are at risk are mistakenly identified as non-risk. This draws attention to a crucial drawback of the default threshold: early intervention systems grow less effective if at-risk students are not identified.

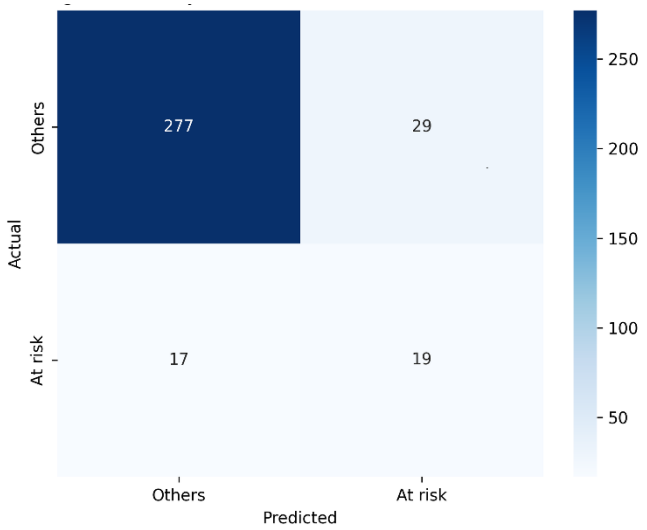

Fig. 8. Binary confusion matrix (threshold 0.5).

4.5 Threshold Optimization

The optimal threshold was determined through a threshold-tuning analysis using recall, precision, F1-score and the F2-score. Since early-warning systems prioritize minimizing false negatives, the F2-score was adopted as the primary selection criterion because it places greater emphasis on recall. Threshold values were evaluated across the interval [0, 1], allowing comparison between highly sensitive and more balanced operational settings.

Figure 9 shows the trade-off between recall and precision across different classification thresholds. Lower thresholds substantially increase recall while reducing precision because of additional false positives. The highest recall (≈ 91.67%) was obtained at the threshold of 0.10, suitable for highly sensitive early-warning scenarios whose main goal is to minimize false negatives; however, it also yields the lowest precision, and therefore more false alarms.

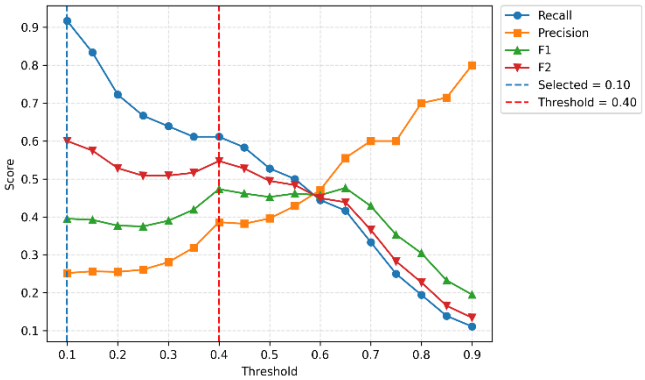

Fig. 9. Threshold-tuning analysis showing the trade-off among recall, precision, F1-score and F2-score for recall-oriented at-risk detection; the selected operating point is 0.10.

Figure 10 presents the corresponding precision–recall curve, which makes the trade-off explicit: precision declines as recall rises. The curve supports the choice of a lower threshold (0.10), where recall is maximized and most at-risk students are identified.

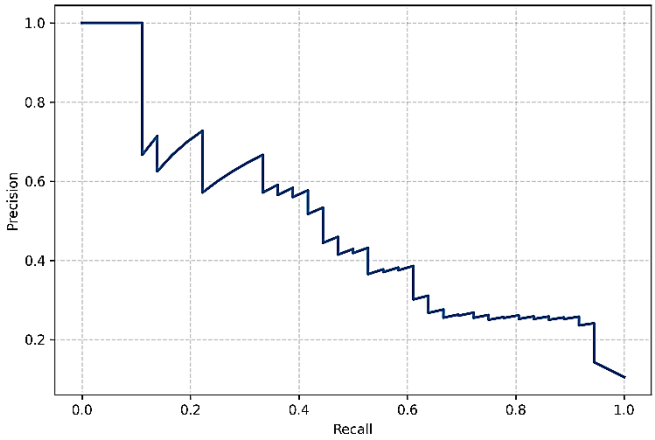

Fig. 10. Precision–recall curve for the binary early-warning model

Table 5 illustrates the impact of threshold tuning on recall-oriented at-risk detection. Lowering the decision threshold from 0.50 to 0.10 substantially increased recall from 52.78% to 91.66%, significantly improving the detection of at-risk students. This threshold also achieved the highest F2-score (60.00%), indicating strong performance under recall-oriented evaluation criteria; however, the increase in recall reduced precision because of additional false positives. In contrast, a threshold of 0.40 provided a more balanced trade-off between recall and precision, achieving the highest F1-score (47.31%) while maintaining markedly improved recall (61.11%) relative to the default threshold. The 0.40 operating point is therefore recommended for routine intervention, while 0.10 remains preferable when the institutional priority is to miss as few at-risk students as possible.

Table 5. Performance comparison across different decision thresholds (%).

Metric	Threshold 0.50	Threshold 0.10	Threshold 0.4
Accuracy	86.55	70.47	85.67

Precision	39.58	25.19	38.60
Recall	52.78	91.66	61.11
F1-score	45.24	39.52	47.31
F2-score	49.48	60.00	54.72

At threshold 0.10 the model recovers 33 of 36 at-risk students (TP = 33, FN = 3), missing only three, with FP = 98 and TN = 208; at the recommended balanced threshold of 0.40 it recovers 22 of 36 at-risk students (TP = 22, FN = 14, FP = 35, TN = 271). The complete confusion matrix at the highly sensitive operating point (0.10) is shown in Fig. 11.

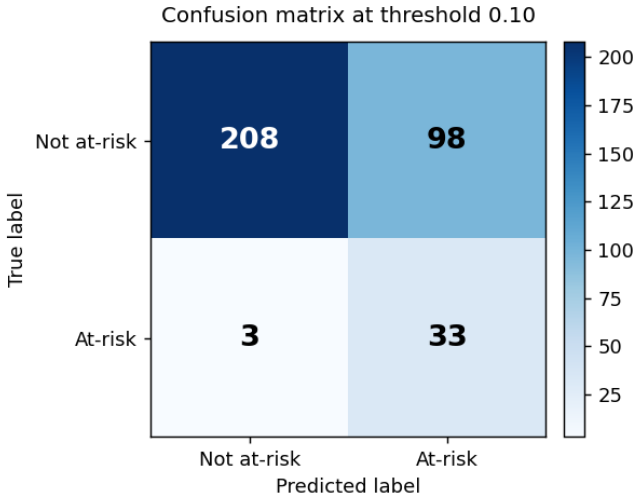

Fig. 11. Binary confusion matrix at the optimised threshold 0.10 (TP = 33, FN = 3, FP = 98, TN = 208).

4.6 Comparison with Prior Approaches

To position the proposed early-warning system against the techniques reviewed in Section 2, Table 6 contrasts the present results with representative prior studies. Earlier Cambodian-context models by Sokkhey et al. and Sokkhey & Okazaki[8,9] and the LMS-based early-warning literature [7 ,10] report competitive overall accuracy but either do not isolate minority-class recall or do not apply threshold optimization. In contrast, the proposed framework explicitly trades a portion of accuracy for a large gain in at-risk recall (from 52.78% to 91.67%), which is the operationally relevant quantity for intervention. This comparison shows that the contribution of this work is not a higher headline accuracy but a substantially more sensitive at-risk detector built on the same class of ensemble models.

Table 6. Comparison of the proposed EWS with representative prior work

Study	**Imbalance handling**	**Threshold Tuning**	**At-risk Recall focus**	**Reported outcome**
Macfadyen & Dawson [10]	No	No	No	Engagement-based prediction
Aldowah et al. [7]	Partial	Limited	Yes (conceptual)	Recall emphasised, not optimised
Sokkhey et al. [8,9]	Limited	No	No	High accuracy, Cambodian context
Proposed EWS	**SMOTE (in-fold)**	**Yes (0.10)**	**Yes (primary)**	**Recall from 52.78%- to 91.67%**

4.7 Key Finding

The experiments yield two principal findings: (i) the minority “at-risk” class remains difficult to detect under standard multi-class modelling, and (ii) threshold optimization substantially improves recall. Gradient Boosting performs best overall, yet in the multi-class setting its ability to identify at-risk students is limited. Reframing the task as binary classification with threshold tuning raised recall from 52.78% to 91.67%, confirming that threshold adjustment is a highly effective early-detection technique.

In educational settings, where missing an at-risk student carries serious consequences, the resulting trade-off—lower

precision and overall accuracy in exchange for much higher recall—is acceptable, because timely intervention depends on early detection. Concretely, the highly sensitive threshold (0.10) flags 98 false positives alongside the 33 detected at-risk students, whereas the balanced threshold (0.40) reduces false positives to 35 while still recovering 22 at-risk students; institutions can therefore choose the operating point that best matches their advising capacity against the cost of missed at-risk students. The study also underscores the importance of data quality, since removing duplicate records produced more accurate performance estimates. Future work will incorporate multimodal data, such as behavioral and visual-engagement signals, to further improve prediction.

4. CONCLUSION

This study proposed a recall-oriented learning analytics framework for the early identification of at-risk students in blended learning environments. While multi-class classification achieved reasonable overall performance, it showed limited effectiveness in detecting minority at-risk students. To address this, the problem was reformulated into a binary classification task with threshold optimization.

The results demonstrate that lowering the classification threshold significantly improves recall, enabling more effective detection of at-risk students. Although this introduces a trade-off with precision and accuracy, it is appropriate for early warning systems where minimizing false negatives is critical. Feature importance analysis further highlights that both academic performance and behavioral engagement are key predictors of student outcomes.

This study contributes to Sustainable Development Goal 4 (SDG 4) by supporting inclusive and equitable education through early identification and targeted intervention. Concretely, the recommended operating point translates into a manageable list of flagged students (for example, 22 true at-risk detections against 35 false positives at threshold 0.40) that can be matched to available advising capacity, making the SDG link actionable rather than purely aspirational. The proposed framework provides a practical and scalable approach for improving student monitoring and academic success in blended learning environments.

ACKNOWLEDGMENTS

The authors acknowledge the financial support from the Higher Education Improvement Project (HEIP2). We also thank the Graduate School and the Institute of Technology of Cambodia (ITC) for their academic and institutional support.

In addition, we sincerely appreciate the Department of Applied Mathematics and Statistics (AMS) for providing the dataset and essential research support.

REFERENCES

[1] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *Proc. 14th Int. Joint Conf. Artificial Intelligence (IJCAI)*, 1995, pp. 1137–1145.

[2] B. Anthony, J. Adzhar, and K. Awanis, *Blended Learning Adoption and Implementation in Higher Education : A Theoretical and Systematic Review*, no. 0123456789. Springer Netherlands, 2020. doi: 10.1007/s10758-020-09477-z.

[3] S. Pokhrel and R. Chhetri, “A Literature Review on Impact of COVID-19 Pandemic on Teaching and Learning,” 2021, doi: 10.1177/2347631120983481.

[4] G. Siemens and R. S. J. Baker, “Learning Analytics and Educational Data Mining : Towards Communication and Collaboration,” 2010.

[5] C. Romero and S. Ventura, “Educational Data Mining: A Review of the State of the Art,” *IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev.*, vol. 40, no. 6, pp. 601–618, 2010.

[6] J. López-zambrano *et al.*, “Early Prediction of Student Learning Performance Through Data Mining : A Systematic Review,” vol. 33, no. 3, pp. 456–465, 2021, doi: 10.7334/psicothema2021.62.

[7] H. Aldowah, A. Baashar, et al., “Systematic Review of Research on Educational Data Mining and Learning Analytics,” *Int. J. Emerg. Technol. Learn.*, vol. 16, no. 24, pp. 128–145, 2021.

[8] P. Sokkhey, S. Navy, L. Tong, and T. Okazaki, “Multi-models of educational data mining for predicting student performance in mathematics: A case study on high schools in cambodia,” *IEIE Trans. Smart Process. Comput.*, vol. 9, no. 3, pp. 217–229, 2020, doi: 10.5573/IEIESPC.2020.9.3.185.

[9] P. Sokkhey and T. Okazaki, “Hybrid machine learning algorithms for predicting academic performance,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 1, pp. 32–41, 2020, doi: 10.14569/ijacsa.2020.0110104.

[10] S. D. Leah P. Macfadyen, “Mining LMS data to

develop an ‘early warning system’ for educators: A proof of concept,” *Comput. Educ.*, vol. 54, no. 2, pp. 588–599, 2010.

[11] L. Breiman, “Random Forest,” *Mach. Learn.*, vol. 45(1), no. 1, pp. 5–32, 2001.

[12] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001.

[13] W. P. K. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, “SMOTE: Synthetic Minority Over-sampling Technique,” *Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[14] S. Dhawan, “Online Learning : A Panacea in the Time of COVID-19 Crisis,” 2020, doi: 10.1177/0047239520934018.

[15] P. Sokkhey and T. Okazaki, “Developing web-based support systems for predicting poor-performing students using educational data mining techniques,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 7, pp. 23–32, 2020, doi: 10.14569/IJACSA.2020.0110704.

[16] Salim Rezvani, Xizhao Wang, “A broad review on class imbalance learning techniques,” *Appl. Soft Comput.*, vol. 143, no. 1, p. 110415, 2023.

[17] P. Sokkhey and T. Okazaki, “Comparative Study of Prediction Models for High School Student Performance in Mathematics,” vol. 8, no. 5, pp. 394–404, 2019.

[18] N. Alangari and R. Alturki, “Predicting Students Final GPA using 15 Classification Algorithms,” vol. 23, no. 3, pp. 238–249, 2020.

[19] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 2nd ed.* 2019.

[20] T. G. Dietterich, *Ensemble methods in machine learning*. 2000.