Wiley
Applied Computational Intelligence and Soft Computing
Volume 2024, Article ID 4795606, 15 pages
https://doi.org/10.1155/2024/4795606

WILEY

Research Article

Increment of Academic Performance Prediction of At-Risk Student by Dealing With Data Imbalance Problem

Nguyen Giap Cu,[1] **Thi Lich Nghiem**,[1] **Thi Hoai Ngo**,[1] **Manh Tuong Lam Nguyen**,[2] **and Hong Quan Phung**[2]

[1]*Faculty of Economic Information System and E-Commerce, Thuongmai University, Hanoi, Vietnam*
[2]*Faculty of Applied Sciences, International School-Vietnam National University, Hanoi, Vietnam*

Correspondence should be addressed to Thi Lich Nghiem; lichnt72@tmu.edu.vn

Received 25 January 2024; Revised 3 September 2024; Accepted 27 September 2024

Academic Editor: Abidhan Bardhan

Studies on automatically predicting student learning outcomes often focus on developing and optimizing machine learning algorithms that fit the data captured from different education systems. This approach has a fatal weakness when it is used for disadvantaged groups, such as those with academic warnings or who have dropped out, because these groups are often much smaller than other common groups in number. The imbalanced data that have class distribution skew create a big challenge to training good classification models. The significant approach to tackle this challenge is applying oversampling methods to increase the number of minor classes; however, generating good new samples from the existing instances of a minor class is still a hard issue and requires new investigation. This study presents two new methods of handling data imbalance based on the original algorithms SMOTE and adaptive synthetic sampling (ADASYN), called Improved SMOTE (I_SMOTE) and Improved ADASYN (I_ADASYN). These modifications involve a new selecting fit candidate method based on a new similarity measurement and a roulette wheel selection to generate synthetic data samples. The aim is to rebalance data and therefore improve the prediction accuracy of minor groups. The proposal methods were designed and applied to education datasets, and they were tested on public datasets and a dataset collected from a Vietnamese university for evaluation. The experimental results on learning datasets showed the high potential of novel algorithms, I_SMOTE and I_ADASYN, for student academic performance problems in general and at-risk student groups especially. Empirical results proved that the recall, precision, and F1-score of the minority class of I_SMOTE and I_ADASYN are strongly better than the original balancing algorithms. Besides, the I_SMOTE and I_ADASYN also improve relatively by 6.6% and 8.0% of the ROC area compared to the original SMOTE and ADASYN, respectively.

Keywords: academic student performance prediction; imbalanced data; Improved ADASYN; Improved SMOTE

1. Introduction

Recent changes in the education models of universities, especially the wide implementation of digital transformation as a mandatory requirement in the education model, have led to a new drive toward developing and applying prediction models for student academic performance [1]. Applying machine learning techniques in data mining is an effective approach to discovering and utilizing new valuable information by uncovering latent patterns in educational data [2, 3]. This is particularly significant in student academic performance prediction (SAPP). Obviously, SAPP is essential for all stakeholders. For example, based on the results of data mining techniques, students can choose appropriate courses, schedules, and study programs. Simultaneously, it also aids teachers to adjust their teaching methods, learning materials, and curricula to better meet the needs and learning goals of each student. For educational institutions, the application of predictive models helps managers review, update training programs, and take appropriate measures to monitor the students' performance at risk of underachievement or dropping out [4–6].

In the literature, numerous studies have focused on predicting student learning outcomes to improve student learning efficiency and to achieve various goals such as reducing training costs and improving the quality of educational management [6–8]. However, researchers often encounter a significant challenge when analyzing educational data: imbalance data. This occurs when the percentage of students with very low qualifications or those who drop out is much less than the rate of typical students. The presence of unbalanced data can exert a substantial influence on the construction of prediction models, potentially resulting in biased predictions. This problem can significantly affect the classification models' performance [9–12]. Nonetheless, in education systems, predicting such minority classes is a primary focus because the students of minority classes such as at-risk students really need support from education stakeholders.

In fact, almost all classification algorithms frequently achieve outstanding accuracy with the majority class; however, this is not the case with the minority class [10]. Although the overall accuracy might be relatively high, minority class components are often disregarded or misclassified more frequently than majority samples. Several techniques have been proposed to improve classification accuracy with imbalanced data, especially for enhancing the classification quality of minority classes [12, 13]. In these techniques, data processing is considered a critical stage to be improved before applying classifiers.

This study aims to rebalance the data and enhance the classification performance, especially the accuracy of prediction for minority classes and target to the educational dataset. To achieve it, two new balancing algorithms are proposed by improving two well-known rebalancing algorithms SMOTE and adaptive synthetic sampling (ADASYN). In the stage of generating a new instance, the random choice mechanism is replaced by a statistical probability-based method, a roulette wheel mechanism, to better performance.

Particularly, rebalancing algorithms SMOTE and ADASYN select a set of candidate instances close to a minority class instance based on the K-nearest neighbors algorithm using a distance measure, such as Euclidian distance. A subset of these candidates is then chosen randomly to generate new instances. In this study, to select better instances, the roulette wheel mechanism requires a fitness measurement that is used as the statistical probability of selection. This fitness measure extends the distance between instances in the datasets by modifying the distance with the weight of data attributes. The proposed approach guarantees that more fit instances have a higher selection opportunity. It is much better than the random choice mechanism.

The proposed algorithms' performance was analyzed on four real datasets collected from India, Uwezo, Oman, and Vietnam. While the datasets from India, Uwezo, and Oman are publicly available, the Vietnamese dataset was created from the data extracted directly from a university and survey data. We compare the performance of our proposed model by evaluating the results of common classifiers on the data after balancing and the original data and compare to the original SMOTE [14] and ADASYN [15] algorithms.

The main contributions of this paper are summarized as follows:

- Improving two prominent data balancing algorithms, SMOTE and ADASYN, using an improved distance estimation method and a method for selecting paired objects through roulette rotation
- Evaluating the predictive performance of the proposed model before and after balancing and comparing it with the original data balancing methods on public datasets as well as on one dataset collected in this study for Vietnamese student

The rest of our paper is structured as follows: Section 2 provides a brief review of relevant works on predicting students' performance and data balancing methods. Section 3 introduces our methodology, while Section 4 describes the experimental results based on four real datasets, including one private dataset from Vietnam. Finally, the conclusions and future works are summarized in Section 5.

2. Related Work

Utilizing the machine learning technique is popular, and among them, applying machine learning for predicting student academic performance is a significant yet challenging task in the educational data mining (EDM) [3, 11]. This challenge is amplified because of the need to deal with imbalanced data as the oversight of minority classes by classifiers results in erroneous classifications. However, accurately predicting the academic outcomes of minority groups is a crucial requirement in EDM practically. Therefore, to address this challenge, numerous researchers have exploited various data mining and machine learning techniques to forecast students' academic performance [11] with the focus point on minority classes. Most of these studies have concentrated on the extraction or selection of student features and classification models while often ignoring the effect of imbalance in student data [12].

In the literature, to solve these class imbalance learning problems, several researchers have developed various solutions, primarily categorized into three approaches: the data-level approach, the algorithm-level approach, and the hybrid-level approach.

2.1. The Data-Level Approach. The data-level approach pertains to the data manipulation activities during the preprocessing stage to balance the data of different classes used in the training process. This strategy involves resampling the original dataset by either oversampling the minority class or undersampling the majority class until the classes are approximately equal [13, 16].

One common undersampling technique is random undersampling (RUS), which reduces the majority class size to create a balanced class distribution. However, RUS may lead to the loss of important information essential for model construction [8]. To overcome this limitation, alternative

sampling methods such as Tomek links [17] and neighborhood cleaning (NCL) [18] have been proposed.

Conversely, oversampling the minority class is a prevalent practice in addressing class imbalance. The most straightforward oversampling algorithm duplicates random instances from the minority class, effectively increasing its size [10]. While this method is simple, it can cause classifiers to overfit and result in longer training times for large datasets. To mitigate these issues, more advanced oversampling algorithms such as SMOTE [14], Borderline SMOTE [11], and Safe-level SMOTE [12] have been developed to improve the outcomes. Researchers have widely utilized the SMOTE family to solve imbalance dataset problems. They have then built models using various classification and deep learning algorithms, such as artificial neural networks (ANNs), K-nearest neighbor (KNN), support vector machines (SVMs), naïve Bayes (NB), random forest (RF), decision tree, J48, and sequential minimal optimization (SMO) [19–21].

For instance, a recent study introduced a novel method called automachine learning with using SMOTE to rebalance data before predicting students' performance [22]. This approach demonstrated superior performance compared to other methods, including ANN, SVM, NB, KNN, and logistic regression (LR). Another study focused on classifiers such as decision tree, neural networks, and balanced bagging, using resampling methods such as SMOTE and ADASYN to predict dropout rates [23].

In a different research effort, a dataset from a public 4-year university was employed to develop predictive models based on various machine learning algorithms, including deep ANN, decision trees, RF, gradient boosting, LR, SVM, and KNN to predict student's academic performance. Various resampling techniques, such as SMOTE, random oversampling (ROS), ADASYN, and SMOTE-ENN, were used to address the imbalanced dataset problem and enhance models' performance [24]. The results indicated that the best outcomes were achieved by training the predictive model using deep neural networks (DNNs) on balanced datasets created via SMOTE oversampling. In addition, research [22] applied three methods to rebalance data, namely, ROS, RUS, and SMOTE, before using four machine learning approaches, including LR, decision trees, neural networks, and SVM. The research illustrated that the combination of SVM and SMOTE proved to be the most effective approach for predicting at-risk students and reducing student dropout rates.

Many studies have also combined SMOTE with feature selection methods to reduce data dimensions, such as wrapper and filter [25], phi correlation coefficients, SPO, and information collection [26]. In another research effort [27], the ADASYN method was used to rebalance the dataset. This research aimed to enhance higher education academic performance prediction using an ANN model and compared it with traditional models, such as LR, decision tree, RF, gradient boosted trees, fast large margin, and generalized linear model. The results demonstrated that ANN performed effectively in solving classification problems, even when dealing with imbalanced data.

2.2. The Algorithm-Level Approach. In addition to research that primarily focuses on data-level approaches, several studies have proposed models based on algorithm-level solutions.

For example, one research [28] proposed a novel approach to predict academic performance involving a two-step process using a feature-weighted SVM and an ANN. The feature-weighted SVM calculates the importance of different features based on information gain ratios and performs coarse-grained binary classification (such as pass, P1, or fail, P0). Subsequently, the detailed score levels are divided from D to A+, and ANN learning is employed for fine-grained, multiclass training of the P1 and P0 classes separately.

Another study [29] focuses on preventing student dropout by exploring various classification techniques on a real dataset that is both multivariate and sequential. To address the challenge of imbalanced data, two strategies were investigated: one involving a weighted loss function and the other utilizing synthetic data generated through a combination of the variational autoencoder and ADASYN technique. The proposed models, employing LSTM and 1D convolutional neural networks (CNNs), aim to leverage the sequential nature of the dataset. The results indicate that the LSTM architecture without a weighted loss function on the preprocessed original dataset achieved the highest sensitivity, while the LSTM with a weighted loss function demonstrated superior specificity and area under the curve.

2.3. The Hybrid Approach. According to Refs. [10, 13, 16], combined methods integrate both data-level and algorithm-level approaches to address the imbalanced data and improve the classification performance. These hybrid methods capitalize on the advantages of both approaches, effectively leveraging their strengths while mitigating their individual weaknesses to achieve higher efficiency. These studies also indicate that algorithm-level approaches involve activities related to classifiers and aim to address imbalanced data issues without modifying the training set. These approaches involve developing novel classification algorithms or enhancing existing algorithms to tackle the bias caused by imbalanced data.

Many studies showed that hybrid algorithms combining both undersampling and oversampling techniques, such as SMOTE and one-sided selection (OSS) [30], as well as SMOTE combined with spread subsampling [31], can be effective. For instance, one research [31] addressed data imbalance in forecasting undergraduate students' success. This research employed supervised learning techniques to construct a postgraduate student graduation time prediction model (PS_GTPM), utilizing the RF ensemble method and ADASYN technique to address data imbalance issues. The results showed that PS_GPTM, when applied to balanced data, outperformed other ensemble models, including bagging and boosting. Another study [21] explored various resampling methods, including both single (SMOTE, Borderline SMOTE) and combined approaches (SMOTE with Tomek links). Classification algorithms used included LR,

KNN, CART, RF, SVM, and stacking. The results showed that the SMOTE–Tomek method performed best with the RF classification.

Additionally, a research [32] proposed the hybrid model to enhance the efficiency of student performance prediction. This hybrid model combined the gray wolf optimization (GWO) for feature selection, SMOTE for data balancing, and random forest for data classification. Experimental analysis showed that the proposed hybrid model achieved significantly improved accuracy, reaching a 98.8% accuracy rate compared to the previous approach, which achieved 93.07% accuracy (RFBTRF-GWO + KNN).

Other researchers have also ventured into hybrid-level, including combining ROS with AdaBoost (ADB) [20] and integrating SMOTE with ensemble methods such as bagging and boosting [33]. In Ref. [29], the issue of predicting student dropouts was approached by generating new samples through a combination of the variational autoencoder and ADASYN techniques. Two models, long short-term memory (LSTM) and 1D CNN, were proposed for this purpose, leveraging the multivariate sequential properties of the data. The results demonstrated that LSTM, particularly when used with a weighted loss function on balanced data, delivered superior outcomes.

However, commonly used data balancing methods such as SMOTE or ADASYN employ random selection in the process of generating new samples for the minority class. This study focuses on enhancing the data balancing process by using selection methods that prioritize elements with a higher likelihood of being chosen based on their fitness. As a result, the data balancing outcome is expected to be more effective.

3. Methodology

3.1. The Proposed Model. Numerous research studies focused on imbalanced datasets generally highlight that the unequal distribution of data among classes tends to introduce a bias in classifier performance, leading to errors skewed toward the majority class. For SAPPs, this bias significantly affects the effectiveness of predicting students' academic performance, particularly in cases where the system aims to identify underperforming individuals within imbalanced datasets. To address this challenge and uphold prediction accuracy for the majority class while improving prediction effectiveness for the minority class, we use a method called "minority class oversampling" with novel improvement for generating new minority samples. The proposals involve applying specialized algorithms designed for handling imbalanced data to rebalance the distribution among classes. Additionally, machine learning algorithms play a crucial role in SAPP and common machine learning algorithms are trained and tested with both public standard datasets and regional-specific datasets rebalanced by our algorithms to evaluate proposals' performance.

In general, research methodology has four stages, as illustrated in Figure 1 with the main improvement occurring in Stage 2, which involves processing imbalanced data using two novel proposals.

3.1.1. Stage 1: Data Preprocessing. It is one of the most crucial steps in SAPP and classification problems, in general, as it converts the raw data into a suitable format to solve errors in the dataset collected from the real world and to achieve better results. The stages carried out in this process are data cleaning, data transformation, and data normalization, which are briefly described below:

- Data cleaning

  During the process of data cleaning, outliers are removed from the dataset and missing values are replaced with the average for numerical data and mode for categorical data. We utilized the KNN algorithm to fill in missing values based on the mean or median of neighboring data points.

- Data transformation

  We transformed our data using both ordinal and one-hot encoding techniques. Ordinal encoding was applied to categorical attributes with a defined order such as grade levels, while one-hot encoding was used for those without any orders. Additionally, we incorporated a frequency encoding method as part of our data transformation process.

- Data normalization

  We used the min–max normalization technique to standardize data features within a specific range. This technique scales the values of features in the range [0, 1].

3.1.2. Stage 2: The Handling of Imbalanced Data. Student performance datasets are observed to be highly imbalanced datasets, and the distribution of the class label of students based on their performance was not equal. To address this issue, we utilized various resampling methods such as SMOTE, ADASYN, and our own improved versions of these approaches. Then, the dataset is divided into two parts: 70% for training and 30% for model testing. The two algorithms for handling imbalanced data are applied to ensure a balanced ratio between the majority and minority classes. This strategic approach aimed to mitigate the impact of class imbalance, fostering improved model training and evaluation accuracy.

3.1.3. Stage 3: Model Implementation. No algorithm can perform well on all aspects (accuracy, learning speed, support for multiclass, handling discrete/binary, continuous attributes, etc.). Each machine learning algorithm has its own strengths and weaknesses and is suitable for specific deployment fields [26]. Therefore, the research used algorithms trained from different techniques to address this problem, including J48, SMO, multilayer perceptron (MLP), NNge, ADB, and LR. These models were chosen because they can provide promising results in predicting student dropout rates.

3.1.4. Stage 4: Evaluation. To evaluate the effectiveness of the classification model, four common metrics were used,

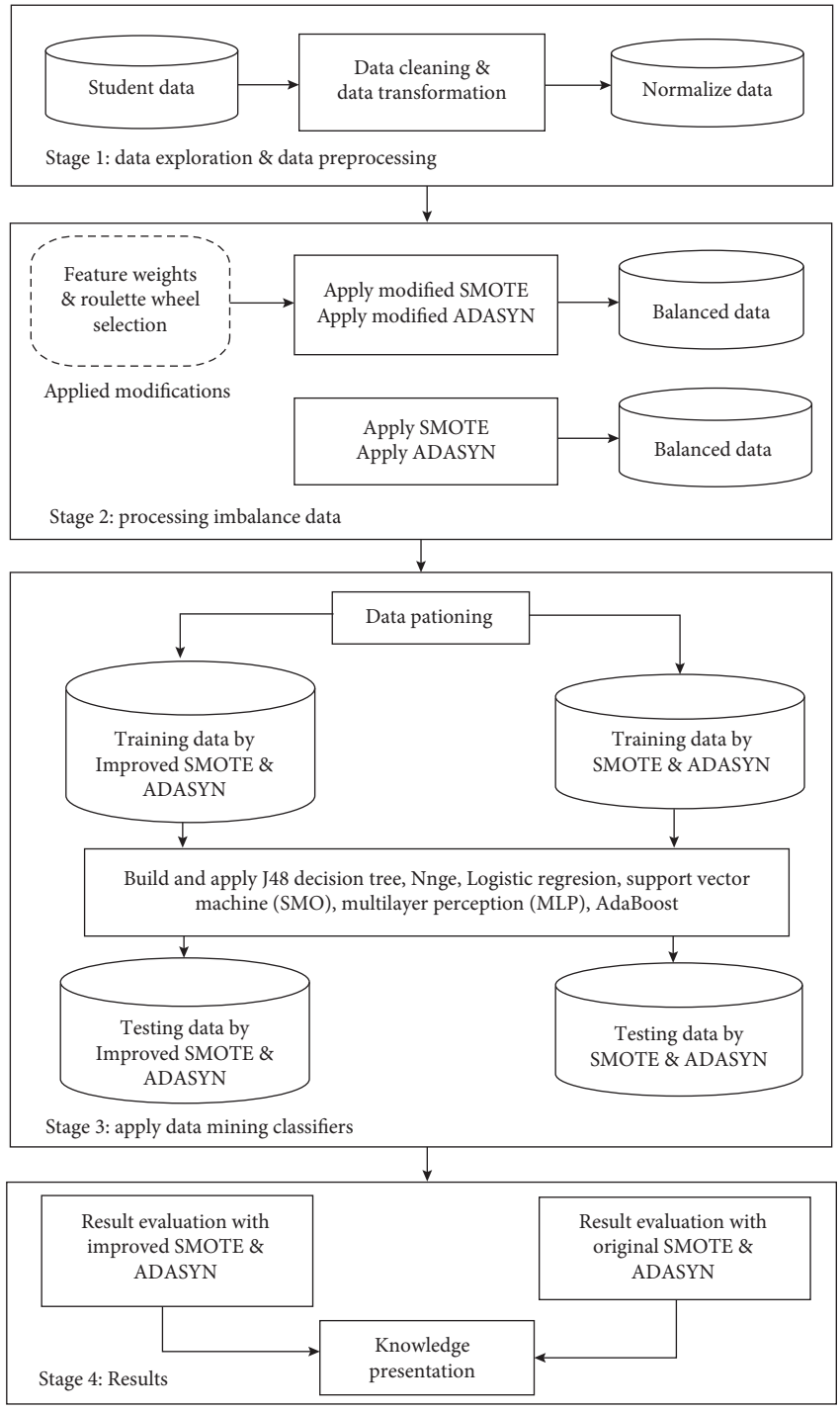

Figure 1: The proposed model.

including recall, precision, F1-score, and true-positive (TP) rate.

3.2. The Handling of Imbalanced Data Techniques. In this study, we applied two techniques for processing imbalanced data, including SMOTE and ADASYN. These algorithms are not only utilized but also enhanced to elevate the quality of the dataset. Subsequently, to further improve the predictive efficiency, we applied ADB to the model. With these algorithms, we proposed two improvements, including enhancing the distance calculation step and improving the mechanism for randomly selecting samples from the K-nearest neighbors to generate new points.

- Regarding the improvement of the distance calculation step, we replaced the Euclidean distance with the Manhattan distance to increase the diversity of synthetic samples. This transformation is valuable in avoiding inaccuracies caused by spherical distance

measurements, promoting flexibility in similar assessments between samples. Additionally, assigning weights to each attribute during distance calculation ensures a more nuanced consideration of the importance of individual dimensions, contributing to the generation of higher quality synthetic samples.

- The second improvement relates to the mechanism for selecting samples from the K-nearest neighbors to generate new samples. We replaced the original mechanism with the roulette wheel selection mechanism. Adopting this method in place of random selection optimizes the process by assigning probabilities based on the similarity between samples. This shift minimizes the chances of generating inaccurate samples, improving the overall accuracy of the data synthesis process. The probability evaluation, based on the similarity between samples, ensures that the synthesis of new samples relies not only on random similarity but also on the essential characteristics of the data, enhancing the efficacy of the SMOTE algorithm.

Our improvement not only focuses on the data synthesis process but also aims to assist the model in optimizing performance and generalizing well on imbalanced data, providing a comprehensive solution to this issue. These enhancements are designed to clarify and strengthen the theoretical foundation of the proposed SMOTE improvement.

3.2.1. Proposal Improvement of SMOTE. SMOTE, proposed by Chawla et al. [14], is an algorithm used to tackle the issue of data imbalance between classes in a dataset for the classification problems, especially when the minority class has significantly fewer samples than the majority class. It is utilized to augment samples in the minority class in imbalanced class problems. The operation of SMOTE involves taking each sample from the minority class and generating synthetic samples by combining them with the nearest minority neighbors. To create a synthetic sample, it calculates the difference between the feature vector of the currently selected sample and the nearest neighbor. By multiplying this difference with a random number in the range [0, 1] and adding it to the feature vector under consideration, a synthetic sample is generated [14].

3.2.1.1. Advantages and Disadvantages. SMOTE, as an improvement over the ROS method, is designed to address overfitting issues and improve the classification model's result on imbalanced datasets. Instead of simply copying minority samples, SMOTE proposes to create new synthetic samples by using interpolation between minority samples within the defined neighborhood [14]. SMOTE generates new samples based on the features of the nearest neighbors, aiming to enhance the model's generalization capabilities.

Although SMOTE is a useful method for augmenting samples in the minority class, it also has some drawbacks concerning overlapping data points when there is an intersection or separation between classes in the feature space [34]. Moreover, SMOTE's random selection approach of K-nearest neighbors may lead to generating samples that do not accurately reflect the characteristics of the minority class.

Creating a fitness measurement for sample's neighbors and using a roulette wheel selection in neighbor selection step of SMOTE can give an improvement for this Algorithm 1.

3.2.2. Proposal Improvement of ADASYN. ADASYN is an oversampling algorithm used to address the issue of class imbalance in machine learning. Using density contribution, ADASYN determines the level of enhancement and generates synthetic samples that simulate the distribution of the original data [15]. This algorithm creates synthetic samples based on the imbalance ratio of each sample in the minority class. Consequently, more samples are generated in areas with a sparser minority class ratio, while fewer samples are created in regions with a denser minority class ratio. ADASYN stands out as a flexible and automatic method widely used to improve the performance of machine learning algorithms when dealing with imbalanced data.

3.2.2.1. Advantages and Disadvantages. The use of $\hat{r}_i$ (density distribution) has helped the ADASYN algorithm generate more samples in regions with high class imbalance. The sampling mechanism of the ADASYN algorithm involves randomly selecting a point from the neighbors to create pairs and generate new samples. However, at points with high class imbalance, this means that within a neighborhood, samples from the majority class are more likely to be chosen, and the newly generated samples may no longer preserve the characteristics of the original minority class.

Like SMOTE, creating a fitness measurement for sample's neighbors and using a roulette wheel selection in the neighbor selection step of SMOTE can improve for this Algorithm 2.

3.3. The Classification Techniques. To analyze and evaluate the effectiveness of the proposed data balancing methods on data before and after balancing will be the basis for evaluating the performance of the proposed algorithms, and some well-known machine learning models are used for the classification task mentioned in Stage 3. These machine learning models are briefly introduced as follows.

- *J48* is a family of decision tree algorithms which is a decision support system that uses tree-like graph decisions [35]. This algorithm, also known as C4.5, J48, represents an enhanced version of the ID3 (Iterative Dichotomiser 3) decision tree algorithm, proposed by Quinlan [36].
- *NNGHE* is a nearest neighbor–like algorithm based on generalized exemplars stored in memory. A nearest-neighbor learner uses the distance between a new example and a set of exemplars in memory to make a decision whether the new example belongs to a particular class [37].

Input:
D: Dataset $\{Y, X_u\}$, in which Y is the dependent variable, X_u is the independent variable ($u = 1, \ldots, n$).
m: number of samples; m_s: number of minority sample; m_l: number of majority samples
$\boldsymbol{d_{th}}$: preset threshold for maximum tolerated degree of imbalance ratio
$\boldsymbol{N}$%: percent of synthetic sample
Output:
$\boldsymbol{D'}$: Dataset $\{Y', X_u'\}$ is rebalancing dataset.
The algorithm process is as follows:
(1) Calculate the number of synthetic samples:
$G = (N/100) \times m$
The result G is rounded.
- If $G > 100$. Go to 2.
- If $G < 100$. Go to 5.
(2) Generating synthetic sample
(a) Calculate $\mathbf{cor_u}(\boldsymbol{Y}; \boldsymbol{X_u})$: The correlation coefficient of each independent variable with the dependent variable.
(b) Compute the weight for each attribute by normalizing $\mathrm{cor}_u(Y; X_u)$: $w_u(Y, X_u) = cor_u(Y, X_u)/\sum_{u=1}^{n} \mathrm{cor}_u(Y, X_u), u = 1, \ldots, n$
(c) For each example $x_i \in$ *minority class*, find K nearest using calculate the distance from $x_i \in$ to the remaining observed samples based on the Manhattan distance formula and the corresponding weight for each attribute.
$\forall x_i \in m_s, \forall x_j \in m$:
$D(x_i, x_j) = \sum_{u=1}^{n} |x_{iu} - x_{ju}| * w(Y, X_u)$
where $i = 1, \ldots, m_s, u = 1, \ldots, n, j = 1, \ldots, m$
//Input for applying the Roulette Wheel Selection Algorithm
- $\boldsymbol{f_i}$: fitness value for the i^{th} neighbor in K-nearest neighbors
- $\boldsymbol{p_i}$: **selection probability** for the i^{th} neighbor in K-nearest neighbors
(d) Fitness value for the i^{th} neighbor is defined as:
$f_i = (1/D(K_i)), i = 1, \ldots, K$
$D(K_i)$: distance from the i^{th} neighbor to example x_i, Calculate the probability for the i^{th} neighbor:
$p_i = f_i/\sum_{i=1}^{K} f_i, i = 1, \ldots, K$
Do the **Loop** from i to G:
(a) Select alternately each sample x_i belonging to the minority class.
(b) Randomly select neighbor x_{zi} from K nearest neighbors based on probability p_i using roulette wheel.
(c) Generate the synthetic sample based on formula:
$s_i = x_i + (x_{zi} - x_i) \times \lambda$
Where $(x_{zi} - x_i)$ is the difference vector in u dimensional spaces, and λ is a random number, $\lambda \in [0, 1]$.
(d) Adding s_i to D to create D′
END **Loop**
Return D′

Algorithm 1: Proposal improvement of SMOTE algorithm (I_SMOTE).

Input:
D: Dataset $\{Y, Xu\}$, in which Y is the dependent variable, Xu is the independent variable ($u = 1, \ldots, n$).
m: number of samples; m_s: number of minority sample; m_l: number of majority sample
$\boldsymbol{d_{th}}$: preset threshold for maximum tolerated degree of imbalance ratio
$\boldsymbol{\beta} \in [0; 1]$: the desired balance ratio, $\beta = 1$ mean a fully balanced dataset
Output:
$\boldsymbol{D'}$: Dataset $\{Y', X_u'\}$ is rebalancing dataset.
The algorithm process is as follow:
(1) Calculate imbalance ratio:
$d = m_s - m_l$
(2) if $d < \boldsymbol{d_{th}}$:
(a) Calculate total number of synthetic samples:
$G = (m_l - m_s) * \beta$
(b) Calculate $\mathbf{cor_u}(\boldsymbol{Y}; \boldsymbol{X_u})$: The correlation coefficient of each independent variable with the dependent variable
(c) Compute the weight for each attribute by normalizing $\mathrm{cor}_u(Y, X_u)$: $w_u(Y, X_u) = \mathrm{cor}_u(Y, X_u)/\sum_{u=1}^{n} \mathrm{cor}_u(Y, X_u), u = 1, \ldots, n$
(d) For each example $\boldsymbol{x_i} \in$ *minority class*, find K nearest using calculate the distance from x_i to the remaining observed samples based on the Manhattan distance formula and the corresponding weight of each attribute.

Algorithm 2: Continued.

$\forall x_i \in m_s, \forall x_j \in m$:
$D(x_i, x_j) = \sum_{u=1}^{n} |x_{iu} - x_{ju}| * w(Y, X_u)$
Where $i = 1, \ldots, m_s$, $u = 1, \ldots, n$, $j = 1, \ldots, m$.
(e) Calculate r_i the ratio:
(1) The ratio $\boldsymbol{r_i}$ is defined as:
$r_i = \Delta_i / K$, $i = 1, \ldots, K$
where Δ_i is the number of examples in the K-nearest neighbors of $\boldsymbol{x_i}$ that belong to the majority class, therefore $r_i \in [0, 1]$
2. Normalized r_i is defined as:
$\hat{r}_i = r_i / \sum_{i=1}^{m_s} r_i$
(f) Calculate the number of synthetic data examples that need to be generated for the each minority example $\boldsymbol{x_i}$:
$g_i = G * \hat{r}_i$
//Input for applying the Roulette Wheel Selection Algorithm
- $\boldsymbol{f_i}$: fitness value for the i^{th} neighbor in K-nearest neighbors
- $\boldsymbol{p_i}$: selection probability for the i^{th} neighbor in K-nearest neighbors
(g) Fitness value for the i^{th} neighbor is defined as:
$f_i = 1/D(K_i), i = 1, \ldots, K$
where $D\boldsymbol{K_i}$: distance from the i^{th} neighbor to example $\boldsymbol{x_i}$
(h) Calculate the probability for the i^{th} neighbor:
$p_i = f_i / \sum_{i=1}^{K} f_i, i = 1, \ldots, K$
(i) For each example $\boldsymbol{x_i}$, generate g_i synthetic data examples following these steps:
Do the **Loop** from 1 to g_i:
(1) Randomly select a data example x_{zi} from K nearest neighbors with p_i for data example $\boldsymbol{x_i}$:
(2) Generate synthetic data example:
$s_i = x_i + (x_{zi} - x_i) * \lambda$
Where $(x_{zi} - x_i)$ is the difference vector in u dimensional spaces, and λ is a random number, $\lambda \in [0, 1]$.
(3) Adding s_i to D to create D'
End **Loop**
Return D'

Algorithm 2: Proposal improvement of ADASYN algorithm (I_ADASYN).

- *MLP* is fundamentally constructed of many layers of linked nodes (neurons). The input layer, one or more hidden layers, and the output layer are the three fundamental layers. There are a set number of neurons in each layer, and each neuron in a layer is connected to every neuron in the layer below it. As part of the learning process, these connections' weights and biases are changed [26].
- *LR* models are used to study effects of predictor variables on categorical outcomes [38].
- *SMO* is an algorithm to solve the quadratic programming (QP) problem that arises during the training of SVM. SMO breaks this problem into a series of smallest possible subproblems, which are then solved analytically [39].
- *AdaBoost* is a classifier that generates a strong classifier by combining many weak classifiers. It is an algorithm that is widely used in various fields due to its ability to handle complex datasets and improve classification performance. The algorithm also efficiently handles complex datasets containing nonlinear relationships and multidimensional feature spaces [10].

4. Results and Discussion

4.1. Datasets. This study used four datasets consisting of three public datasets from India [40], Uwezo [41], and Oman [42], and the remaining dataset from a university in Vietnam. The general statistic information of these datasets is presented in Table 1. The first dataset was students' dropout rates from India collected in 2016. This dataset consisted of nearly 19,000 samples, of which nearly 5% of students dropped out and the rest 95% continued to study. The second dataset was children's data collected in Uwezo in 2017. The Uwezo dataset comprised 76,000 samples, of which 3.5% of children drop out, while the remaining 96.5% attended school or were underage. The collected dataset in Vietnam consisted of nearly 2% of students who had weak academic results, while the other 98% achieved average or better results. These datasets are highly imbalanced datasets.

The India dataset included variables such as student IDs; gender; caste; mathematics, English, and science marks; science and language teachers; guardians, Internet access; school IDs; total students; total toilets; and establishment year. The Uwezo dataset contained 45 variables, providing information about the areas where the children reside, assessing the quality of healthcare, and details about their education, including the type of education, school type, and the completion status of their courses.

The third dataset, from the higher educational institution in the Sultanate of Oman, comprises five modules of data from 2017 to 2021, grouped into three categories: student academic information, student activities, and student video interactions. The student academic information data are extracted from the student information system, while student activity data are obtained from Moodle and student

Table 1: Experiment dataset.

Dataset	Size	Variables	Majority class	Minority class
India	18,762	17	17,870	892
Uwezo	44,359	45	41,783	2576
Oman	326	21	264	62
Vietnam	524	44	513	11

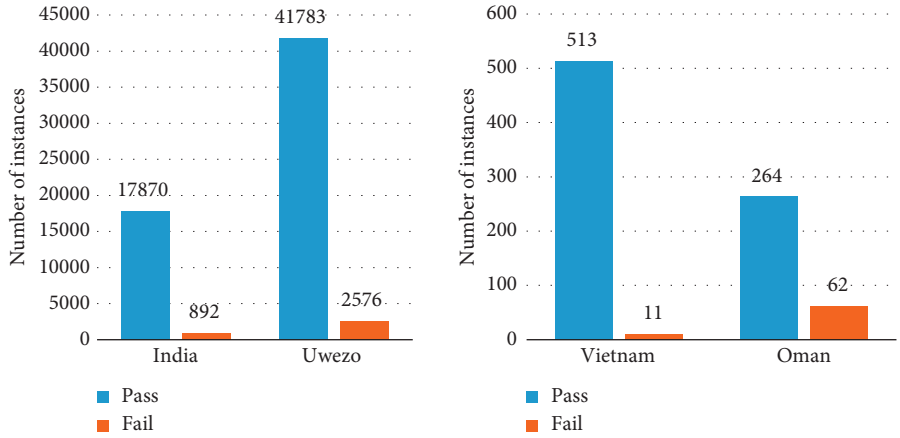

Figure 2: Imbalanced dataset.

video interaction from the eDify mobile application. We utilized a final dataset that consisted of 326 instances and 21 features.

The remaining dataset was collected from student information at Thuongmai University, Vietnam. We gathered a dataset containing 524 samples with 44 variables, including information about students before they started college, such as their admission area and university entrance score. It also included details about their first- and second-year study results, categorized by subject types such as general, supplementary, and specialized subjects.

After completing the data preprocessing, those four datasets were highly imbalanced, as presented in Figure 2.

These datasets provide a comprehensive array of information, ranging from socioeconomic factors to academic performance indicators, enabling a thorough exploration of the factors influencing student outcomes in diverse educational contexts.

4.2. Results. This study aims to improve classification performance in the minority class; however, the overall performance needs to be at the same level as original data. Table 2 shows the performance of all algorithms for both minority and majority classes.

The results show that the prediction accuracy of the classification models does not change significantly when applying data balancing algorithms.

For each dataset and classification model, the best result among original dataset and balancing algorithms is set in bold type. In some cases, data balancing algorithms can improve the overall forecast performance by a small margin, but in other cases, the forecast results in terms of both the majority and minority classes will be decreased, but the degree of decrease is not large. Besides, the effect of rebalancing methods is different with different classifiers. In general, it means that users can apply balancing algorithm in the data preprocess phase without worrying of loss overall accuracy.

Note that the most important prediction on education datasets is minority class prediction. The results of evaluating the classification ability of the classification algorithms for the minority classes are shown in the following tables below.

Table 3 shows performance metrics for machine learning algorithms across four datasets, with a clear variation in results.

In Table 3, for the India dataset, J48 and NNge algorithms exhibit exceptional performance with perfect recall, precision, and F1-scores of 100%. In contrast, the Uwezo dataset poses a challenge for these algorithms, with significantly lower F1-scores, the highest being 41.7% for J48. The Vietnam dataset reveals a disparity in algorithm performance, with MLP achieving an F1-score of 75%, suggesting a better fit for these particular data. Meanwhile, the Oman dataset presents a varied scenario; however, J48 stands out with an F1-score of 56%. These figures highlight the significant impact dataset characteristics have on the predictive power of machine learning algorithms.

Table 4 illustrates the impact of SMOTE and ADASYN, two class balancing techniques, on the performance of machine learning algorithms across four datasets. In the India dataset, J48 and NNge algorithms particularly stand out, with J48 hitting a perfect F1-score of 100% with ADASYN and NNge close behind at 98.4%. For the Uwezo dataset, while improvements are evident, the algorithms still

Table 2: Classification accuracy.

Dataset	J48	NNge	LR	MLP	SMO	ADB
India	**100.0**	**100.0**	**95.4**	97.3	95.4	95.4
India_SMOTE	99.9	**100.0**	93.0	**100.0**	95.4	89.3
India_ADASYN	**100.0**	99.9	91.6	99.8	95.0	94.3
India_I_MOTE	**99.9**	**99.5**	79.4	99.6	95.0	**95.9**
India_I_ADASYN	99.9	98.9	82.9	98.5	**98.9**	92.9
Uwezo	94.8	**93.4**	**94.4**	**93.9**	**94.4**	**94.4**
Uwezo_SMOTE	94.9	90.8	93.7	91.0	**94.4**	93.4
Uwezo_ADASYN	94.5	88.7	93.9	91.4	**94.4**	93.4
Uwezo_I_MOTE	**96.8**	89.9	81.5	85.5	82.1	85.1
Uwezo_I_ADASYN	96.6	89.1	79.9	83.5	80.4	83.8
Oman	88.8	81.6	**82.7**	83.7	**84.7**	**89.8**
Oman_SMOTE	**90.8**	83.7	80.6	81.6	**84.7**	87.8
Oman_ADASYN	83.7	84.7	81.6	82.7	80.6	88.8
Oman_I_MOTE	81.6	**86.7**	75.5	80.6	79.6	**89.8**
Oman_I_ADASYN	85.7	83.7	78.6	**84.7**	82.7	82.7
Vietnam	97.8	97.8	**96.7**	98.9	98.3	98.3
Vietnam_SMOTE	97.2	**98.9**	95.0	**99.4**	**98.9**	97.8
Vietnam_ADASYN	97.2	**98.9**	**96.7**	**99.4**	**98.9**	**98.9**
Vietnam_I_MOTE	98.3	88.3	92.1	93.3	89.7	98.3
Vietnam_I_ADASYN	**98.8**	91.7	94.7	93.1	90.9	98.8

Note: Bold values present the rebalanced dataset that give best accuracy for each algorithm.

Table 3: Performance of machine learning algorithms across original datasets.

Dataset	Algorithms	Recall (%)	Precision (%)	F1-score (%)	ROC area (%)	TP rate (%)	FP rate (%)
India	J48	100.0	100.0	100.0	100.0	100.0	0.0
	NNge	100.0	100.0	100.0	100.0	100.0	0.0
	MLP	41.1	100.0	58.2	62.3	41.1	0.0
	LR	0.0	—-	—-	71.8	0.0	0.0
	SMO	0.0	—-	—-	50.0	0.0	0.0
	ADB	0.0	—-	—-	78.8	0.0	0.0
Uwezo	J48	33.6	55.1	41.7	79.8	33.6	1.6
	NNge	19.9	34.3	25.2	58.8	19.9	2.2
	MLP	12.9	36.5	19.0	71.5	12.9	1.3
	LR	0.9	43.8	1.9	72.2	0.9	0.1
	SMO	0.0	—-	—-	50.0	0.0	0.0
	ADB	0.0	—-	—-	77.6	0.0	0.0.
Oman	J48	41.2	87.5	56.0	75.1	41.2	1.2
	NNge	0.0	—-	—-	49.4	0.0	1.2
	MLP	23.5	57.1	33.3	60.3	23.5	3.7
	LR	17.6	50.0	26.1	54.9	17.6	3.7
	SMO	11.8	100.0	21.1	55.9	11.8	0.0
	ADB	41.2	100.0	58.3	69.2	41.2	0.0
Vietnam	J48	0.0	—-	—-	51.7	0.0	0.0.
	NNge	25.0	50.0	33.3	62.2	25.0	0.6
	MLP	75.0	75.0	75.0	98.0	75.0	0.6
	LR	50.0	33.3	40.0	76.0	50.0	2.3
	SMO	25.0	100.0	40.0	62.5	25.0	0.0
	ADB	75.0	60.0	66.7	98.6	75.0	1.1

struggle, with the highest F1-score being 50.6% for J48 with SMOTE. The Vietnam dataset shows remarkable improvement using ADASYN, with NNge, MLP, and ADB algorithms all achieving an F1-score of 85.7%, 85.7%, and 80%, respectively. Finally, the Oman dataset results are mixed, but ADB with ADASYN again shines with an F1-score of 56%. These numbers highlight the effectiveness of class balancing methods in enhancing algorithmic predictions where there is data imbalance.

Table 5 summarizes results after applying the proposed I_SMOTE and I_ADASYN algorithms to various datasets. Overall, the I_SMOTE and I_ADASYN improve relatively by 6.6% and 8.0% of ROC area compared to the original SMOTE and ADASYN. It is easy to see that the recall and

Table 4: Experimental results using the original SMOTE and ADASYN algorithms.

Dataset	Algorithms	SMOTE						ADASYN					
		Recall (%)	Precision (%)	F1-score (%)	ROC area (%)	TP rate (%)	FP rate (%)	Recall (%)	Precision (%)	F1-score (%)	ROC area (%)	TP rate (%)	FP rate (%)
India	J48	100.0	98.1	99.0	100.0	100.0	0.1	100.0	100.0	100.0	100.0	100.0	0.0
	NNge	100.0	100.0	100.0	100.0	100.0	0.0	96.1	100.0	98.4	98.4	96.9	0.0
	MLP	100.0	100.0	100.0	100.0	100.0	0.0	96.5	100.0	98.2	96.7	96.5	0.0
	LR	5.8	9.0	7.1	72.7	5.8	2.8	11.6	10.9	11.3	70.1	11.6	4.5
	SMO	0	—	—	50.0	0.0	0.0	7.4	30.6	11.9	53.3	7.4	0.8
	ADB	44.6	20.1	27.7	78.8	44.6	8.5	9.7	22.1	13.5	77.1	9.7	1.6
Uwezo	J48	47.0	54.8	50.6	76.2	47.0	2.3	43.8	50.5	47.0	77.3	43.8	2.5
	NNge	25.7	21.8	23.6	60.2	25.7	5.4	34.4	20.0	25.3	63.2	34.4	8.1
	MLP	27.7	23.6	25.5	70.3	27.7	5.3	24.5	23.8	24.1	69.2	24.5	4.6
	LR	11.5	30.9	16.8	68.5	11.5	1.5	8.1	30.8	12.8	67.6	8.1	1.1
	SMO	0	—	—	50.0	0.0	0.0	0	—	—	50.0	0.0	0.0
	ADB	12.7	28.8	17.7	73.2	12.7	1.8	12.7	28.8	17.7	73.0	12.7	1.8
Oman	J48	52.9	90.0	66.7	81.7	52.9	1.2	52.9	52.9	52.9	74.2	52.9	9.9
	NNge	5.9	100.0	11.1	52.9	5.9	0.0	11.8	100.0	21.1	55.9	11.8	0.0
	MLP	29.4	45.5	35.7	71.8	29.4	7.4	35.3	50.0	41.4	58.8	35.3	7.4
	LR	23.5	40.0	29.6	55.3	23.5	7.4	23.5	44.4	30.8	54.8	23.5	6.2
	SMO	41.2	58.3	48.3	67.5	41.2	6.2	17.6	37.5	24.0	55.7	17.6	6.2
	ADB	47.1	72.7	57.1	69.7	47.1	3.7	41.2	87.5	56.0	64.3	41.2	1.2
Vietnam	J48	75.0	42.9	54.5	86.4	75.0	2.3	75.0	42.9	54.5	86.4	75.0	2.3
	NNge	100.0	66.7	80.0	99.4	100.0	1.1	100.0	66.7	80.0	99.4	100.0	1.1
	MLP	75.0	100.0	85.7	98.7	75.0	0.0	75.0	100.0	85.7	98.7	75.0	0.0
	LR	50.0	22.2	30.8	79.7	50.0	4.0	25.0	25.0	25.0	87.6	25.0	1.7
	SMO	50.0	100.0	66.7	75.0	50.0	0.0	50.0	100.0	66.7	75.0	50.0	0.0
	ADB	75.0	50.0	60.0	99.2	75.0	1.7	100.0	66.7	80.0	99.6	100.0	1.1

Table 5: Experimental results using the I_SMOTE and I_ADASYN algorithms.

Dataset	Algorithms	I_SMOTE						I_ADASYN					
		Recall (%)	Precision (%)	F1-score (%)	ROC area (%)	TP rate (%)	FP rate (%)	Recall (%)	Precision (%)	F1-score (%)	ROC area (%)	TP rate (%)	FP rate (%)
India	J48	100	100	100	100	100	0	100	99.3	99.7	99.9	100	0.2
	NNge	100	100	100	100	100	0	99.7	95.3	97.4	99.2	99.7	1.2
	MLP	98.8	100	99.4	100	98.8	0	98.2	94.5	96.3	98.4	98.2	1.4
	LR	5.9	39.6	10.2	51.8	5.9	2.2	34.3	64.3	44.7	64.7	34.3	4.8
	SMO	81.9	92.4	86.8	90.1	81.9	1.7	99.7	95.3	97.4	99.2	99.7	1.2
	ADB	87.0	91.9	89.4	92.5	87.0	1.9	73.8	88.9	80.6	85.7	73.8	2.3
Uwezo	J48	97.1	88.4	92.5	96.9	97.1	3.2	97.2	87.7	92.2	96.8	97.2	3.6
	NNge	84.5	70.6	77.0	87.9	84.5	8.8	87.0	68.9	76.9	88.3	87.0	10.3
	MLP	41.7	74.8	53.5	69.1	41.7	3.5	32.5	73.8	45.1	64.7	32.5	3.0
	LR	14.0	68.8	23.3	56.2	14.0	1.6	8.0	63.1	14.1	53.4	8.0	1.2
	SMO	17.2	71.7	27.8	57.8	17.2	1.7	11.0	68.4	19.0	54.8	11.0	1.3
	ADB	33.2	80.8	47.1	65.6	33.2	2.0	28.2	82.2	41.9	63.3	28.2	1.6
Oman	J48	52.9	47.4	50	77.5	52.9	12.3	58.8	58.8	58.8	80.5	58.8	8.6
	NNge	47.1	88.9	51.5	72.9	47.1	1.2	47.1	53.3	50	69.2	47.1	8.6
	MLP	52.9	45	48.6	71.8	52.9	13.6	41.2	58.3	48.3	58.8	41.2	6.2
	LR	41.2	33.3	36.8	54.8	41.2	17.3	29.4	35.7	32.3	55.9	29.4	11.1
	SMO	41.2	41.2	41.2	64.4	41.2	12.3	29.4	50	37	61.6	29.4	6.2
	ADB	47.1	88.9	61.5	66.4	47.1	1.2	41.2	63.6	50	71.5	41.2	4.9
Vietnam	J48	100	92.3	96.0	99.0	100.00	2.1	100	94.4	97.1	99.3	100	1.5
	NNge	98.8	63.4	77.2	92.3	98.8	14.3	98.8	70.9	82.6	94.3	98.8	10.1
	MLP	71.4	87.0	78.4	84.4	71.4	2.7	76.2	87.7	81.5	86.8	76.2	2.7
	LR	84.5	82.6	83.5	90.0	84.5	4.5	89.3	85.2	87.2	92.7	89.3	3.9
	SMO	57.1	87.3	69.1	77.5	57.1	2.1	63.1	88.3	73.6	80.5	63.1	2.1
	ADB	100	92.3	96.0	99.0	100	2.1	100	94.4	97.1	99.3	100	1.5

precision of minority class of I_SMOTE and I_ADASYN are significantly better than the original balancing algorithms.

In detail, for the India dataset, J48 and NNge exhibit perfect recall, precision, and F1-scores of 100% with I_SMOTE and nearly perfect with I_ADASYN. LR shows limited improvement with these methods, evident from its low F1-scores of 5.7% with SMOTE and 11.3% with I_ADASYN. However, the proposed methods improve the performance of LR, SMO, and ADB remarkably.

The results on Uwezo dataset are outstanding, with the best F1-score of 96.9% using I_SMOTE with J48 and 92.2% using I_ADASYN with J48. These results are nearly double the performance of the original SMOTE and ADASYN. The similar improvements are seen in the Vietnam dataset; J48 shows an improvement with an F1-score of 96.0% using I_SMOTE and 97.1% using I_ADASYN. The recall indexes are strongly improved in most cases, and the precisions are the same.

Oman's most notable result is ADB's performance, achieving an F1-score of 63.6% with I_ADASYN. However, the result of I_SMOTE is not good in this dataset.

These achievements highlight the effectiveness of class balancing algorithms in enhancing the predictive performance of various models.

The data balancing method showed that J48 and NNge are the top algorithms in accurately classifying the highest number of dropouts followed by ADB, MLP, and logistic. It is worth noting that both the original SMOTE and ADASYN algorithms, along with the proposed improvements, have varying levels of success in classifying minority classes. The proposed algorithms I_SMOTE and I_ADASYN had improved results in most of the datasets and classifiers; however, there is small number of cases these algorithms could not improve the results. Therefore, additional research is needed to optimize the selection process using roulette wheels to cater to the specific requirements of each problem. Moreover, the algorithm incorporated the calculation of element suitability based on the roulette wheel model in the selection process. As a result, the calculation time has increased compared to the original algorithm.

The average computing times of rebalancing algorithms are depicted in Table 6. The average time is calculated from 50 running times.

It is arbitrary that the two proposed methods have higher running time than the original rebalancing algorithms. New algorithms improve the quality by adding a stochastic selection method that requires more computing time. However, the data rebalancing process is conducted only once during the model development, making the computational time cost acceptable.

Among the two proposed methods, the I_SMOTE is much faster than I_ADASYN. The performance of these algorithms on four datasets 50 times demonstrates the difference of I_SMOTE and I_ADASYN on consuming time.

4.3. Limitation and Discussion. The experimental results show that the proposed algorithms I_SMOTE and I_ADASYN have achieved much better results compared to the original algorithms in minority classification. The improvement can be seen across all metrics. However, there are still certain limitations in the approaches of extended SMOTE and ADASYN. First, the algorithms have heterogeneous impacts on different datasets and classification tasks. This necessitates thorough testing with various algorithms in practical applications. Second, compared to the original algorithms, the improved algorithms require longer computing times. In practical applications, parallel computing methods can be adopted to enhance algorithm performance.

Table 6: Computing time (seconds).

Dataset	SMOTE	ADASYN	I_SMOTE	I_ADASYN
India	0.0261	0.0410	0.807	**1.1533**
Uwezo	0.0792	0.1445	2.0720	**3.1370**
Oman	0.0258	0.0220	0.0391	**0.0866**
Vietnam	0.0292	0.0349	0.0649	**0.0717**

Note: The bold values present the algorithm has longest computing time.

Although there are certain limitations, the proposed algorithms have great potential for real-world applications. Theoretically, the probabilistic selection mechanism and the fitness measure also offer possibilities for further improvement and development in comparison with random selection mechanism.

5. Conclusion

In this study, it aims to use data balancing algorithms to enhance the classification of minority classes in educational data. Two new algorithms were proposed that improve the SMOTE and ADASYN algorithms by introducing new attribute weighting and applying roulette selection to generate new samples of minority class. The effectiveness of the proposed algorithms is evaluated by estimating the performance of popular classification algorithms on original datasets and their rebalancing versions. The results indicate that the proposed I_SMOTE and I_ADASYN algorithms have notably improved the classification of minority classes when using J48, NNge, MLP, LR, SMO, and ADB classification algorithms. It is easy to see that the recall, precision, and F1-score of minority class of I_SMOTE and I_ADASYN are significantly better than the original balancing algorithms. Besides, the I_SMOTE and I_ADASYN improved relatively by 6.6% and 8.0% of ROC area compared to the original SMOTE and ADASYN.

Among these, J48, NNge, and MLP algorithms show the highest accuracy on test datasets. Therefore, our algorithms improve predicting minority classes in student data, especially for those who discontinue education. However, further enhancements to these algorithms are necessary to achieve even better outcomes. In the future, the roulette wheel mechanism will be refined for better performance.

Data Availability Statement

The three public datasets from India, Uwezo, Oman that support the findings of this study are available in Sheik

Mohamed Imran at https://www.kaggle.com/datasets/imrandude/studentdropindia2016, Twaweza East Africa at https://data.humdata.org/dataset/a6fb7ed1-5614-4aa2-b391-3fc2a4a94217, and Raza Hasan at https://zenodo.org/records/5591907. And the Vietnamese data that support the findings of this study is available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest.

Funding

This research was funded by Thuongmai University, Hanoi, Vietnam.

Acknowledgments

This research was funded by Thuongmai University, Hanoi, Vietnam.

References

[1] T. M. Mahmoud, T. Abd-El-Hafeez, and A. Badawy, "A Framework for an E-Learning System Based on Semantic Web" (2013).

[2] Y. Zhang, Y. Yun, R. An, J. Cui, H. Dai, and X. Shang, "Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis," *Frontiers in Psychology* 12 (Dec 2021): 698490, https://doi.org/10.3389/fpsyg.2021.698490.

[3] W. Xiao, P. Ji, and J. Hu, "A Survey on Educational Data Mining Methods Used for Predicting Students' Performance," *Engineering Reports* 4, no. 5 (Dec 2021): https://doi.org/10.1002/eng2.12482.

[4] H. Hassan, N. B. Ahmad, and S. Anuar, "Improved Students' Performance Prediction for Multi-Class Imbalanced Problems Using Hybrid and Ensemble Approach in Educational Data Mining," *Journal of Physics* 1529, no. 5 (May 2020): 052041, https://doi.org/10.1088/1742-6596/1529/5/052041.

[5] T. M. Barros, P. A. SouzaNeto, I. Silva, and L. A. Guedes, "Predictive Models for Imbalanced Data: A School Dropout Perspective," *Education Sciences* 9, no. 4 (Nov 2019): 275, https://doi.org/10.3390/educsci9040275.

[6] L. C. Yu, C. Lee, H Pan, et al., "Improving Early Prediction of Academic Failure Using Sentiment Analysis on Self-Evaluated Comments," *Journal of Computer Assisted Learning* 34, no. 4 (Aug. 2018): 358–365, https://doi.org/10.1111/jcal.12247.

[7] A. Siddique, A. Jan, F. Majeed, A. I. Qahmash, N. N. Quadri, and M. O. A. Wahab, "Predicting Academic Performance Using an Efficient Model Based on Fusion of Classifiers," *Applied Sciences* 11, no. 24 (Dec 2021): 11845, https://doi.org/10.3390/app112411845.

[8] M. Yağcı, "Educational Data Mining: Prediction of Students' Academic Performance Using Machine Learning Algorithms," *Smart Learning Environments* 9, no. 1 (Dec 2022): 11, https://doi.org/10.1186/s40561-022-00192-z.

[9] S. D. A. Bujang, A. Selamat, R Ibrahim, et al., "Multiclass Prediction Model for Student Grade Prediction Using Machine Learning," *IEEE Access* 9 (2021): 95608–95621, https://doi.org/10.1109/ACCESS.2021.3093563.

[10] P. Kumar, R. Bhatnagar, K. Gaur, and A. Bhatnagar, "Classification of Imbalanced Data: Review of Methods and Applications," *IOP Conference Series: Materials Science and Engineering* 1099, no. 1 (Mar 2021): 012077, https://doi.org/10.1088/1757-899X/1099/1/012077.

[11] A. M. E. Koshiry, T. Abd El-Hafeez, A. Omar, and E. H. I. Eliwa, "A Prediction System Using AI Techniques to Predict Students' Learning Difficulties Using LMS for Sustainable Development at KFU," *Data Science and Algorithms in Systems* 597 (2023): 22–36, https://doi.org/10.1007/978-3-031-21438-7_2.

[12] M. Khairy, T. M. Mahmoud, and T. Abd-El-Hafeez, "The Effect of Rebalancing Techniques on the Classification Performance in Cyberbullying Datasets," *Neural Computing and Applications* 36, no. 3 (Jan 2024): 1049–1065, https://doi.org/10.1007/s00521-023-09084-w.

[13] M. H. A. Hamid, M. Yusoff, and A. Mohamed, "Survey on Highly Imbalanced Multi-Class Data," *International Journal of Advanced Computer Science and Applications* 13, no. 6 (2022): https://doi.org/10.14569/IJACSA.2022.0130627.

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique," *Jair* 16 (Jun 2002): 321–357, https://doi.org/10.1613/jair.953.

[15] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," in *2008 IEEE International Joint Conference on Neural Networks, Hong Kong, China* (Piscataway: IEEE, Jun 2008), 1322–1328, https://doi.org/10.1109/IJCNN.2008.4633969.

[16] V. S. Spelmen and R. Porkodi, "A Review on Handling Imbalanced Data," in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), Coimbatore* (Piscataway: IEEE, Mar 2018), 1–11, https://doi.org/10.1109/ICCTCT.2018.8551020.

[17] H. He and E. A. Garcia, "Learning From Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering* 21, no. 9 (Sep 2009): 1263–1284, https://doi.org/10.1109/TKDE.2008.239.

[18] N. O. A. Abdouli, Z. Aung, W. L. Woon, and D. Svetinovic, "Tackling Class Imbalance Problem in Binary Classification Using Augmented Neighborhood Cleaning Algorithm," *Lecture Notes in Electrical Engineering* 339 (2015): 827–834, https://doi.org/10.1007/978-3-662-46578-3_98.

[19] N. M. Aslam, I. U. Khan, L. H. Alamri, and R. S. Almuslim, "An Improved Early Student's Academic Performance Prediction Using Deep Learning," *International Journal of Emerging Technologies in Learning (iJET)* 16, no. 12 (Jun 2021): 108, https://doi.org/10.3991/ijet.v16i12.20699.

[20] A. Yaqin, M. Rahardi, and F. F. Abdulloh, "Accuracy Enhancement of Prediction Method Using SMOTE for Early Prediction Student's Graduation in XYZ University," *International Journal of Advanced Computer Science and Applications* 13, no. 6 (2022): https://doi.org/10.14569/IJACSA.2022.0130652.

[21] V. Flores, S. Heras, and V. Julian, "Comparison of Predictive Models With Balanced Classes Using the SMOTE Method for the Forecast of Student Dropout in Higher Education," *Electronics* 11, no. 3 (Feb 2022): 457, https://doi.org/10.3390/electronics11030457.

[22] D. Thammasiri, D. Delen, P. Meesad, and N. Kasap, "A Critical Assessment of Imbalanced Class Distribution Problem: The Case of Predicting Freshmen Student Attrition," *Expert Systems with Applications* 41, no. 2 (Feb 2014): 321–330, https://doi.org/10.1016/j.eswa.2013.07.046.

[23] J. Dhar and A. K. Jodder, "An Effective Recommendation System to Forecast the Best Educational Program Using Machine Learning Classification Algorithms," *ISI* 25, no. 5 (Nov 2020): 559–568, https://doi.org/10.18280/isi.250502.

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

[25] P. Vaishnavi, Ch. Prathima, V. Rakesh, P. Sujala, P. A. Nitin, and D. R. K. Yadav, "Employing the SMOTE Technique, a Machine Learning Model for Predicting Student Grades," in *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)* (Piscataway: IEEE, 2023), 349–354.

[26] S. Alija, E. Beqiri, A. S. Gaafar, and A. K. Hamoud, "Predicting Students Performance Using Supervised Machine Learning Based on Imbalanced Dataset and Wrapper Feature Selection," *Informatica* 47, no. 1 (Mar. 2023): https://doi.org/10.31449/inf.v47i1.4519.

[27] C. F. Rodríguez-Hernández, M. Musso, E. Kyndt, and E. Cascallar, "Artificial Neural Networks in Academic Performance Prediction: Systematic Implementation and Predictor Evaluation," *Computers and Education: Artificial Intelligence* 2 (2021): 100018, https://doi.org/10.1016/j.caeai.2021.100018.

[28] C. Huang, J. Zhou, J. Chen, J. Yang, K. Clawson, and Y. Peng, "A Feature Weighted Support Vector Machine and Artificial Neural Network Algorithm for Academic Course Performance Prediction," *Neural Computing & Applications* 35, no. 16 (2023): 11517–11529, https://doi.org/10.1007/s00521-021-05962-3.

[29] E. C. Coppo, R. S. Caetano, L. M. De Lima, and R. A. Krohling, "Student Dropout Prediction Using 1D CNN-LSTM with Variational Autoencoder Oversampling," in *2022 IEEE Latin American Conference on Computational Intelligence (LA-CCI)* (Piscataway: IEEE, Nov. 2022), 1–6, https://doi.org/10.1109/LA-CCI54402.2022.9981340.

[30] Y. Pristyanto, I. Pratama, and A. F. Nugraha, "Data Level Approach for Imbalanced Class Handling on Educational Data Mining Multiclass Classification," in *2018 International Conference on Information and Communications Technology* (Piscataway: IEEE, 2018), 310–314.

[31] H. S. Bako, F. U. Ambursa, B. S. Galadanci, and M. Garba, "Predicting Timely Graduation of Postgraduate Students Using Random Forests Ensemble Method," *Fudma Journal of Sciences* 7, no. 3 (2023): 177–185, https://doi.org/10.33003/fjs-2023-0703-1773.

[32] K. Deepika and N. Sathyanarayana, "Hybrid Model for Improving Student Academic Performance," *International Journal of Advanced Research in Engineering and Technology* 11, no. 10 (2020): 768–779, https://doi.org/10.34218/IJARET.11.10.2020.078.

[33] I. D. Shetty, D. Shetty, and S. Roundhal, "Student Performance Prediction," *International Journal of Computer Applications Technology and Research* 8, no. 5 (2019): 157–160, https://doi.org/10.7753/ijcatr0805.1003.

[34] A. Fernandez, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for Learning From Imbalanced Data: Progress and Challenges, Marking the 15-Year Anniversary," *Journal of Artificial Intelligence Research* 61 (Apr. 2018): 863–905, https://doi.org/10.1613/jair.1.11192.

[35] R. Bhargava, M. Mathuria, N. Bhargava, and G. Sharma, "Decision Tree Analysis on J48 Algorithm for Data Mining," *International Journal of Advanced Research in Computer Science and Software Engineering* (Jun. 2013).

[36] S. L. Salzberg, "C4.5: Programs for Machine Learning" (1994).

[37] U. Adhikari, T. H. Morris, and S. Pan, "Applying Non-Nested Generalized Exemplars Classification for Cyber-Power Event and Intrusion Detection," *IEEE Transactions on Smart Grid* 9, no. 5 (2018): 3928–3941, https://doi.org/10.1109/TSG.2016.2642787.

[38] T. G. Nick and K. M. Campbell, "Logistic Regression," *Methods in Molecular Biology* 404 (2007): 273–301, https://doi.org/10.1007/978-1-59745-530-5_14.

[39] J. C. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines" (Jul 1998).

[40] "Student-Drop-India2016," https://www.kaggle.com/datasets/imrandude/studentdropindia2016.

[41] "Uwezo 2017 Dataset," https://data.humdata.org/dataset/a6fb7ed1-5614-4aa2-b391-3fc2a4a94217.

[42] "Oman Dataset," https://zenodo.org/records/5591907.