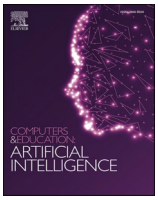
Computers and Education: Artificial Intelligence 6 (2024) 100222

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence



# Enhancing algorithmic assessment in education: Equi-fused-data-based SMOTE for balanced learning

Yasmine Chachoui [a,*], Nabiha Azizi [b], Richard Hotte [c], Tahar Bensebaa [a]

[a] Computer Science Department, LRI Laboratory, Badji Mokhtar Annaba University, Annaba, Algeria
[b] Computer Science Department, Labged Laboratory, Badji Mokhtar Annaba University, Annaba, Algeria
[c] I2A Applied Artificial Intelligence Institute, TELUQ University, Montreal, QC, Canada

ARTICLE INFO

*Keywords:*
Artificial intelligence
Machine learning
Education datasets
Oversampling techniques
SMOTE

ABSTRACT

Recently, there has been a growing interest among researchers in enhancing the efficacy of learning through the utilization of diverse machine learning models within the field of artificial intelligence. However, imbalanced data distributions in educational datasets present a significant challenge to machine learning algorithms. This imbalance can result in biased models, untrustworthy outcomes, and poor performance. Data was gathered from a sample of 2176 first-year novice programming students in this study. Due to an alarming 76% failure rate, the imbalanced dataset was preprocessed before being oversampled with techniques such as SMOTE, SMOTE Borderline, SMOTE-ENN, and ADASYN. The proposed non-redundant synthetic data cooperation approach, named Equi-Fused-Data-based SMOTE, seeks to capitalize on the diversity of the obtained data by combining oversampled datasets. The balanced bagging model was then applied to the combined dataset to demonstrate the robustness of this approach. The promising results demonstrate the effectiveness of the Equi-Fused-Data-based SMOTE model, which achieved a higher Accuracy of 93.85%, a Precision, Recall and F1-score of 92,86%, and an AUC of 98.08%.

## 1. Introduction

In the 21st century, programming has emerged as a critical skill, often referred to as the new literacy. However, acquiring coding skills remains difficult for novice learners, with global failure rates ranging from 25% to 80% (Abdessemed et al., 2018; Gross & Powers, 2005; Lahtinen et al., 2005; Pillay, 2003; Pillay & Vikash, 2005; Price & Barnes, 2015). Despite efforts from researchers in interdisciplinary fields such as environmental development, smart tutors, and serious games, understanding computer programming remains difficult, particularly at the introductory level (Gross & Powers, 2005). This alarming rate of academic failure in introductory programming highlights the critical need to identify and address the root causes. Several studies have looked at individual factors that influence programming performance, such as problem-solving abilities (Sim & Lau, 2018), learner intellectual capacity, mathematical skills, motivation (Yilmaz & Karaoglan Yilmaz, 2023), and the ability to apply effective learning strategies. However, research indicates that learning styles have a smaller impact (Kirschner, 2017; Lu et al., 2003; Nancekivell et al., 2020; Wilkinson et al., 2014). Specific areas of difficulty have also been identified, including tables, structured data types, recursion, pointers, references, and memory manipulation. These difficulties can range from syntax errors to semantic or pragmatic difficulties, which are difficulties in applying programming knowledge to a specific case (Lahtinen et al., 2005; McCall & Kölling, 2019).

Given the complexities of these challenges, as well as the inherent variability in students' backgrounds and prerequisites, assessing programming aptitude solely through traditional methods raises serious questions about validity and fairness. Furthermore, the imbalanced nature of educational datasets, in which the minority class (e.g., students who pass exams) far outnumbers the majority class (e.g., students who fail), complicates accurately predicting student outcomes and developing effective assessment tools (Radwan & Cataltepe, 2017). Furthermore, this imbalance can result in biased models, unreliable results, and poor performance because the overall Accuracy metric does not accurately reflect the model's effectiveness (Fernandez et al., 2018; Wang et al., 2021). Given the need for effective solutions, various sampling techniques have been proposed to address class imbalance, including

* Corresponding author. Université Badji Mokhtar Annaba, Sidi Amar, Annaba, Algeria.
*E-mail addresses:* yasmine.chachoui@univ-annaba.org (Y. Chachoui), azizi@labged.net (N. Azizi), richard.hotte@teluq.ca (R. Hotte), t.g.bensebaa@gmail.com (T. Bensebaa).

https://doi.org/10.1016/j.caeai.2024.100222
Received 1 October 2023; Received in revised form 9 April 2024; Accepted 11 April 2024
Available online 16 April 2024

oversampling and under-sampling (Wang et al., 2021). Under-sampling reduces the number of instances in the majority class, resulting in a more balanced distribution. Thus, it helps to reduce bias that can occur when a single class dominates the training data. There are several under-sampling techniques, including random under-sampling, cluster-based under-sampling, and cost-sensitive approaches (Tarekegn et al., 2021). The specific problem and dataset in question determines the technique used. Oversampling techniques, such as SMOTE (Synthetic Minority Over-sampling Technique) (Chawla et al., 2002), SMOTE Borderline, and ADASYN (Adaptive Synthetic Sampling), have been developed to create synthetic samples or modify existing ones to rebalance the class distribution.

Despite the promising results of these techniques, it is critical to recognize their limitations, such as the risk of overfitting (Fernandez et al., 2018), while under-sampling methods can result in the loss of valuable information (Wongvorachan et al., 2023). Furthermore, previous research has identified the distinct strengths and weaknesses of each oversampling technique, emphasizing the difficulty of selecting a single method to effectively address class imbalance (Tariq et al., 2023; Wongvorachan et al., 2023). However, to our knowledge, the specific efficacy of these techniques in the context of programming aptitude assessment has not been thoroughly investigated. Recognizing the limitations of existing oversampling techniques, the purpose of this study is to fill a gap in the literature by investigating the impact of different sampling techniques on algorithmic performance when dealing with class imbalance in a dataset. In this study, different sampling techniques were assessed, such as SMOTE, SMOTE Borderline, SMOTE-ENN (SMOTE Edited Nearest Neighbors), and ADASYN, affect classification algorithms' performance in an algorithmic learning context. The proposed Equi-Fused-Data approach aims to address the shortcomings of individual techniques by leveraging the strengths of various oversampling methods and improving classification Accuracy, Precision, Recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Equi-Fused-Data uses a non-redundant data cooperation approach to improve classification performance. Furthermore, the study aims to provide insights into the broader impact of class imbalance on various classification algorithms, as well as to assess the efficacy of the proposed Equi-Fused-Data-based SMOTE model.

This study, which is part of a college program aimed at improving algorithmic learning, examines data collected over two semesters from 2176 first-year students enrolled in science and technology from 2020 to 2022. The dataset contains both formative and summative assessments. Formative assessments include grades for behavioral characteristics such as class participation, attendance, and project work. Grades were assigned on a scale of 0–20, with 10 or higher indicating passing. Out of 2176 students, only 518 (23.8%) passed the algorithmic module, while 1658 (76.2%) failed. Given the dataset's imbalance, testing different sampling methods is critical. We focused specifically on SMOTE, a popular technique for creating synthetic samples that is both simple and effective. SMOTE Borderline was also considered, which concentrates on samples near the decision boundary to improve class separation. We included SMOTE-ENN, which uses nearest neighbors to generate synthetic samples while preserving the local data distribution. Finally, the performance of the various techniques was compared to ADASYN, which generates more samples for difficult minority cases.

The main research questions of this study are.

1. Which sampling technique (SMOTE, SMOTE Borderline, SMOTE-ENN, or ADASYN) is most effective for this specific dataset? By comparing the performance of the various techniques, we can determine which one is best suited to address the class imbalance in this educational dataset while ensuring the most accurate and reliable assessment results.
2. How does class imbalance impact the performance of various classification algorithms? Different classification algorithms may react differently to imbalanced datasets. By comparing the performance of various techniques on different algorithms, we can gain valuable insights into the specific challenges that class imbalance presents to each algorithm, allowing us to choose the best algorithm for a given evaluation task.
3. Can the Equi-Fused-Data-based SMOTE model, a non-redundant data cooperation approach, outperform traditional oversampling techniques in terms of classification model Accuracy, F1-score, and AUC when applied to an imbalanced educational dataset from an introductory programming class?

Considering these questions, the main contribution of this paper is to develop machine learning models to predict student performance on summative and formative assessments. Furthermore, to address the issue of imbalanced data, this study investigates several balancing techniques in the literature and proposes a novel approach called Equi-Fused-Data. This approach takes advantage of a new scheme for coordinating the main performing balancing techniques. Experimental comparative studies were also carried out to determine the effectiveness of the proposed model. In addition, to better illustrate the importance of addressing class imbalances in education, Section 2 provides an overview of the research conducted in this field, as well as how various sampling techniques and strategies have been used to reduce class imbalances. Section 3 introduces the proposed method, and Sections 4 and 5 describe and discuss the obtained results. Section 6 concludes the paper and presents future perspectives.

## 2. Related work

### 2.1. Educational assessment and the challenge of class imbalance

Pedagogical assessment tools are essential in modern learning environments because they allow teachers to track students' progress, identify areas for improvement, and guide effective learning strategies. However, class imbalance in educational datasets can have a significant impact on the Accuracy and fairness of these tools. Class imbalance occurs when one class in a dataset is significantly overrepresented in comparison to the others. In education, this could manifest as a dataset with significantly more data points for high-achieving students than low-achieving students. Such imbalances have the potential to mislead traditional classification algorithms, resulting in biased models (Wang et al., 2021). These models benefit the majority class while misclassifying students who require additional assistance. Hence, it undermines effective learning interventions. For example, an imbalanced dataset may result in a model that consistently misclassifies borderline-performing students as high performers. This bias may prevent educators from identifying and supporting at-risk students.

### 2.2. Oversampling techniques: mitigating imbalance for fairer assessment

Researchers have investigated various oversampling techniques to address the problem of class imbalance while also ensuring the fairness and effectiveness of assessment tools in education. These techniques, such as SMOTE (Chawla et al., 2002), attempt to balance the class distribution within a dataset by generating synthetic data points for the minority class. Oversampling techniques aid classification algorithms in learning a more accurate representation of the data by artificially increasing the number of data points for the underrepresented class (for example, students with difficulties). This can result in more equitable and reliable assessment tools that can effectively identify and support students across the performance spectrum.

### 2.3. Sampling techniques and their applications in education

Sampling techniques are essential in addressing class imbalances in educational datasets, which occur when one class (e.g., students who pass an exam) outnumbers the other. This imbalance presents challenges

for traditional classification algorithms, potentially leading to biased models and unreliable results (Wang et al., 2021). To address this issue, various oversampling techniques have proven to be promising solutions for balancing class distributions and improving classification performance. SMOTE is a popular oversampling technique in educational settings (Chawla et al., 2002). SMOTE creates synthetic observations for the minority class in unbalanced data. For each minority class observation, synthetic observations are generated at random between the observation and its K-nearest minority class neighbors. This method is computationally efficient and thus appropriate for large datasets (Fernandez et al., 2018). When comparing the efficacy of oversampling techniques, SMOTE has demonstrated its effectiveness in addressing underachievement rates and behavioral issues in educational settings (Rachburee & Punlumjeak, 2021; Wongvorachan et al., 2023). For example, (Khalaf Hamoud et al., 2022) discovered that the SMOTE improved the Accuracy of both supervised and unsupervised machine learning algorithms for predicting student performance. The Random Forest performed consistently well before and after the SMOTE application in terms of Precision, Recall, and F1-score (83%). SMOTE improves the overall performance of all algorithms, particularly Precision and Recall.

There are several other SMOTE variants. In the study conducted by (Wongvorachan et al., 2023), various resampling techniques were evaluated. The results show that Random Forest performed best when combined with the hybrid approach SMOTE Nominal Continuous and Random Under-Sampling (SMOTE-NC + RUS), both on moderately and extremely imbalanced datasets. For moderately imbalanced datasets, the hybrid approach yielded the highest Accuracy (77%), Precision (79%), Recall (74%), AUC-ROC (86%), and F1-score (77%). While the Random Over-Sampling (ROS) technique achieved (87%) Accuracy despite potential overfitting, the RUS technique performed the worst across all metrics (Accuracy: 70%, Precision 72%, Recall 66%, AUC-ROC: 76%, F1-score: 69%). Even with extremely imbalanced datasets, the hybrid approach performed well, with 90% Accuracy and a AUC-ROC of 96.7%, demonstrating its suitability for educational datasets. (Tariq et al., 2023) emphasized the importance of selecting the right combination of data balancing techniques and classifiers for optimal performance on imbalanced datasets. The study's findings show that combining SMOTETomek oversampling with the K-Nearest Neighbors (kNN) classifier produces the highest Accuracy (83.72%) on a multi-class educational dataset. SMOTETomek reduces noise and generates synthetic data that is similar to real data points, which improves the distance-based approach of kNN. However, SMOTE's noise sensitivity may cause it to underperform. The technique can amplify noise in the data, reducing the model's generalizability, or it can generate redundant data with limited diversity, impairing the model's ability to learn.

SMOTE was originally designed for binary classification problems, but researchers have extended it to handle imbalanced datasets with multiple classes by repeatedly applying it to balance each minority class (Tarekegn et al., 2021). In multi-class classification, the SMOTE technique can be applied by comparing the minority class to the remaining classes using a one-versus-all comparison (Fernandez et al., 2018). However, SMOTE may not be suitable for all classifiers due to noise or overlap between classes, which can lead to misclassification. This is primarily because it ignores the underlying distribution of the data, which can result in unrealistic samples. It is also susceptible to the selection of nearest neighbors and the oversampling ratio (Feng et al., 2021). To address these limitations, researchers created SMOTE variants such as SMOTE Borderline and SMOTE-ENN. SMOTE Borderline improves on the original SMOTE by concentrating on samples that are on the border between classes. This improves its ability to handle imbalanced datasets with overlapping classes (Han et al., 2005). It can generate synthetic samples in feature space regions that are difficult for classifiers to learn, potentially improving the performance of classifiers like kNN, Support Vector Machine (SVM), and Naive Bayes (Intayoad et al., 2019).

According to (Intayoad et al., 2019), the use of SMOTE and its variants (Borderline-SMOTE1, Borderline-SMOTE2, SVM-SMOTE) can, in some cases, improve classification performance for the minority class (failing students). This improvement is especially noticeable for kNN and Naive Bayes classifiers. For example, for SMOTE and Bordelrine-SMOTE1, kNN received F1-scores of (92%) and (93%), respectively. Nonetheless, the SMOTE Borderline can produce noisy samples if the decision boundary is not well defined (Nabus et al., 2022). It can also be computationally costly when the dataset contains a large number of borderline samples. Its effectiveness is highly dependent on the quality of the algorithm used to identify the boundary region. SMOTE-ENN is another variant of SMOTE that uses the Edited Nearest Neighbors rule to combine the strengths of under and oversampling techniques. It removes noisy and misclassified SMOTE samples, thereby improving overall performance (Chawla et al., 2002; Fernández et al., 2018). SMOTE-ENN subsamples the majority class and then uses SMOTE to generate synthetic data for the minority class. This method addresses the issue of overfitting, which frequently occurs when SMOTE is applied to large datasets. However, if informative minority samples are near the majority class, SMOTE-ENN may discard them, reducing the dataset's size. As a result, there is a risk of losing valuable information (Krawczyk, 2016).

SMOTE-ENN proved to be an effective method for increasing the overall Accuracy of the model. Thus, in their study (Nabil et al., 2021), various methods were used to resample the dataset, including SMOTE, ADASYN, ROS, and SMOTE-ENN. The experimental results showed that the deep neural network (DNN) model achieved an Accuracy of (89%), demonstrating the effectiveness of deep learning on a balanced dataset for predicting students' academic performance in the field of educational data mining. Another promising oversampling technique is ADASYN, which dynamically adjusts the ratio of synthetic samples to be generated to the density of the various regions, resulting in a more refined approach to oversampling. In other words, ADASYN is an adaptation of SMOTE that produces more synthetic data points for difficult-to-learn classes by effectively addressing datasets with severe class imbalances (He et al., 2008; Rozi et al., 2023) used the ADASYN technique to assess students' achievement. The findings show that using ADASYN with a stacking approach resulted in an F1-score of (97%). The results indicate that resampling techniques improve classification performance. However, the neighborhood size and dataset distribution can have an impact on ADASYN's performance. It can generate noisy samples if the density ratio between classes is too high. It is also computationally intensive for large datasets.

However, when developing models for educational datasets, it is important to consider the techniques' strengths and limitations. Indeed, despite advances in oversampling techniques, their effectiveness in various educational contexts, such as algorithmic learning, has not been the subject of in-depth study. While previous research has primarily focused on individual techniques, this study aims to close this gap by comparing the effectiveness of various oversampling techniques (SMOTE, SMOTE Borderline, SMOTE-ENN, and ADASYN), as well as further exploring the relationships and determining which combinations of oversampling techniques and algorithms perform optimally in the context of programming skills assessment. In addition, this study introduces a new Equi-Fused-data-based SMOTE model that acknowledges the limitations of existing oversampling techniques. The Equi-Fused-Data approach aims to address the shortcomings of individual techniques by leveraging the strengths of various oversampling methods, improving classification Accuracy, Precision, Recall, F1-score, and AUC.

## 3. Methodology

Fig. 1 depicts the proposed approach, which includes data collection, preprocessing, data partitioning, SMOTE variants and Equi-Fused-Data-based SMOTE to manage the imbalanced dataset, and supervised models

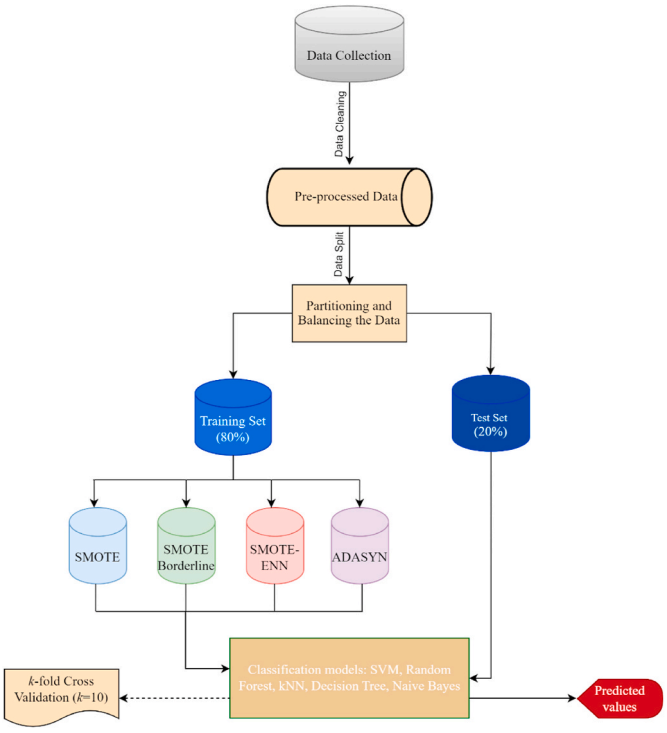



**Fig. 1.** Methodology steps.

to classify the data.

### 3.1. Data collection

The dataset includes assessments (formative and summative) of science and technology students enrolled between 2020 and 2022. The Head of the Department of Science and Technology helped create this dataset by ensuring that the Annaba University's student data guidelines were followed. Students were randomly assigned to groups for evaluation. Each semester consisted of two formative and one summative assessments. Each assessment was carefully designed to address the specific learning objectives and prerequisites. Instructors from Computer Science 1 (CS1) and (CS2) evaluated student performance anonymously. To ensure confidentiality and student privacy, all assessments were anonymized before grading. This means that all identifying information, such as names and student IDs, was removed from the students' assignments. Instead, each student received a unique code that was used to identify their work when it was graded. Following anonymization, the students' papers were randomly assigned to different professors for grading. This ensures that every student's work is evaluated fairly and objectively.

Furthermore, the dataset contains data from four primary features collected during various assessments in (CS1) and (CS2) courses. This study used a total of 2176 instances (rows), with the attributes/variables listed in Table 1. Since the aim is to predict student performance, we faced a multi-class classification problem. Instead of a restrictive binary classification (pass/fail), we chose a three-category system. This option recognizes the nuanced range of student performance (failing, moderate, excelling) found in the dataset. With this finer granularity, we hope to more accurately represent individual performance and provide a more insightful basis for analysis. 36.12% of instances (786) are labeled as "Low" (score 0–6) and represent students who are struggling academically. 63.10% of instances (1373) are labeled "Moderate" (score 7–13), indicating that students are meeting grade level expectations, while the remaining (0.78%) of instances (17) are labeled "High" (score 14–20), indicating that students are excelling in the computer science program. Thus, this dataset is unbalanced.

**Table 1**
Dataset description.

| Attribute | Description | Values | Min | Max | Mean | STD |
|---|---|---|---|---|---|---|
| Grade1 | Mean score for formative assignments (CS1) | Numeric: from 0 to 20 | 0 | 19.50 | 10.72 | 5.39 |
| Grade2 | Mean score for formative assignments (CS2) | Numeric: from 0 to 20 | 0 | 20 | 9.88 | 5.82 |
| Exam1 | Score for summative assessment (CS1) | Numeric: from 0 to 20 | 0 | 18 | 4.70 | 3.46 |
| Exam2 | Score for summative assessment (CS2) | Numeric: from 0 to 20 | 0 | 18.50 | 2.57 | 2.66 |
| Performance | Academic performance | Ordinal: low, moderate, high | – | – | – | – |

Grade1 and Grade2 represent the mean (range 0–20) of various course elements, such as formative assignments and projects in (CS1) and (CS2), respectively. Furthermore, behavioral characteristics related to student engagement were assessed using a combination of class participation, which refers to the frequency and quality of participation in activities and group projects. Attendance records serve as a measure of engagement. Moreover, project work refers to the quality and completion of individual and group projects. These characteristics were also evaluated on a scale of 0–20 and factored into their formative assignment score. Exam 1 and Exam 2 are the scores (range 0–20) from two summative assessments that cover topics such as binary coding and programming basics for Exam 1 and advanced concepts such as strings, loops, and arrays for Exam 2.

### 3.2. Data preprocessing and feature selection

To ensure the integrity of the proposed analysis, a comprehensive data cleaning and feature selection were performed. This process included the following steps:

**Data formatting.** The target variable was converted from its original format, which contained categorical values, to a numerical representation using label encoding.

**Identification and removal of duplicates.** Duplicate entries were removed. This process resulted in the removal of 286 instances, reducing the dataset to 1890 instances.

**Correction of typos and inaccuracies.** Any typos or inaccuracies, such as misspellings in the dataset that could potentially affect the results were corrected to maintain data integrity.

**Treatment of outliers.** The interquartile range (IQR) method was used to identify potential outliers. Q1 and Q3 were calculated for each column, and data points outside the range of Q1 -1.5 * IQR and Q3 + 1.5 * IQR were labeled as potential outliers. However, no outliers were identified in this analysis.

**Treatment of missing values.** There were no missing values in the dataset.

**Randomization of data.** To reduce sample bias, the dataset was randomly generated. This step was designed to ensure that the subsequent analysis was conducted on a representative sample free of systemic bias.

**Feature selection.** Two different feature selection techniques were used to identify the most relevant features for predicting student performance: Information Gain and Subset Evaluator Classifier.

The Information Gain algorithm assesses the value of a feature by calculating its information gain in relation to the class (Amrieh et al., 2016). The ranking method allows us to rank the attributes according to their individual scores. The significant attributes identified and their corresponding score values are: Grade 1 ranked first (0.57), Exam 1

ranked second (0.50), Grade 2 ranked third (0.49), and Exam 2 came in last (0.35). The Subset Evaluator Classifier uses the bagging algorithm and the Best-First technique to evaluate subsets based on training data. A classifier is used to estimate the "merit" of a set of attributes. The Best-First algorithm investigates attribute subsets using greedy hill climbing and backtracking (Sharma et al., 2022). This technique selected the same Grade1, Exam1, Grade2, and Exam2 features, each with a merit of 0.94. In both methods, Grade 1, Exam 1, Grade 2, and Exam 2 were identified as key features that correspond to our understanding of student performance. This comprehensive approach captures various aspects of learning and provides an overall picture of progress. Grades 1 and 2 reflect consistent comprehension and application of concepts, whereas Exams 1 and 2 assess comprehensive application of knowledge under time constraints.

### 3.3. Partitioning and balancing the data

This section examines the methodology employed to divide and balance the data for the experiments.

#### 3.3.1. Train/test split

To avoid bias and noise from overlapping instances, the dataset was resampled into two sets: training (80%, 1512 instances) and testing (20%, 378 instances). This common 80-20 split allowed us to collect enough data for training while leaving a sizable portion for an unbiased evaluation of model performance. The testing set remained consistent throughout the experiment trials.

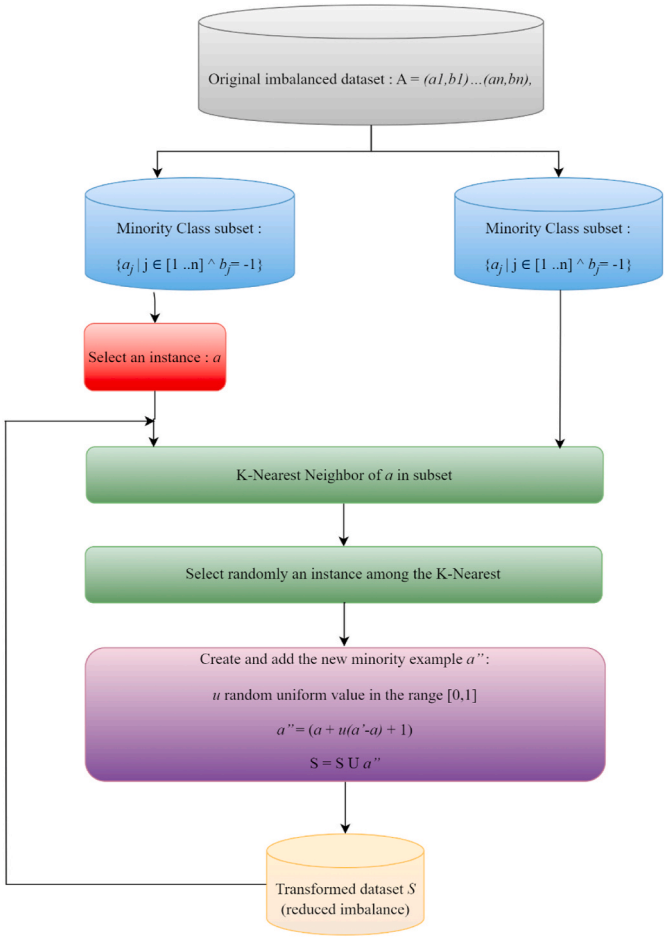



**Fig. 2.** SMOTE algorithm.

#### 3.3.2. Resampling

Variations of the SMOTE technique (Chawla et al., 2002) were used to balance the data. Fig. 2 illustrates the various steps of the SMOTE algorithm. First, we set the total amount of oversampling *N* to achieve an approximate 1:1 class distribution. The second step was an iterative process that began with randomly selecting a positive class from the training set. Next, its kNN (5 b y default) was computed. Then, at random, we selected *N* instances from the *K* to generate new instances via interpolation. To do this, we computed the difference between the feature vectors. The result is the selection of a random point along the "line segment" between the features.

To address the dataset's class imbalance, we applied a variation of the SMOTE technique. While SMOTE effectively oversamples the minority class, it can generate noise. Thus, we investigated SMOTE Borderline, SMOTE-ENN, and ADASYN. First, the SMOTE Borderline prioritizes oversampling near the decision boundary, which can enhance classification performance. Second, SMOTE-ENN uses oversampling and edited nearest neighbor cleanup to remove noisy majority class instances near the minority class. Finally, ADASYN performs adaptive oversampling of minority instances based on their learning difficulty, potentially improving oversampling effectiveness. The following results were obtained.

- SMOTE achieved a nearly balanced distribution (3273 instances, 1091 per class).
- SMOTE Borderline technique yielded results that were identical to those of SMOTE.
- SMOTE-ENN technique resulted in 3061 instances (1091 "high", 997 "low", 973 "moderate").
- ADASYN yielded 3258 instances (1091 each for the "moderate" and "high" classes, 1076 for "low").

### 3.4. Equi-fused-data-based SMOTE model: synthetic data cooperation for improving multi-class learning

The Equi-Fused-Data-based SMOTE model addresses the problem of imbalanced datasets in multi-class learning by combining the benefits of oversampling and ensemble learning. Imbalanced datasets present significant challenges to traditional classification algorithms. Metrics like Accuracy, which are frequently optimized by these algorithms, become misleading in imbalanced scenarios. Instead, AUC and F1-score are important metrics that consider both Precision and Recall.

#### 3.4.1. Model's components

The Equi-Fused-Data-based SMOTE model consists of two components (Fig. 3). First, it employs four well-known oversampling methods: SMOTE, SMOTE Borderline, SMOTE-ENN, and ADASYN. Every technique has its advantages and disadvantages. SMOTE excels at dealing with general imbalance, whereas SMOTE Borderline focuses on samples near the class boundary. SMOTE-ENN uses nearest neighbors to generate data, whereas ADASYN focuses on learning from difficult minority samples. Using these different techniques without duplicates allows us to capture a broader range of potential minority class instances, enriching the training data. The next step is to use balanced bagging, an ensemble technique designed specifically for imbalanced data. Balanced bagging addresses class imbalance within the ensemble by generating multiple sub-models from balanced subsets of oversampled data (Barros et al., 2019). In fact, ensemble learning, which combines the predictions of multiple models, is frequently superior to single models. This approach encourages diversity within the ensemble while reducing the impact of the majority class. Fig. 3 illustrates how the proposed system works by showing the sequence of the various phases.

#### 3.4.2. Theoretical foundation

The proposed Equi-Fused-Data-based SMOTE model capitalizes on the synergy between oversampling and ensemble learning. By

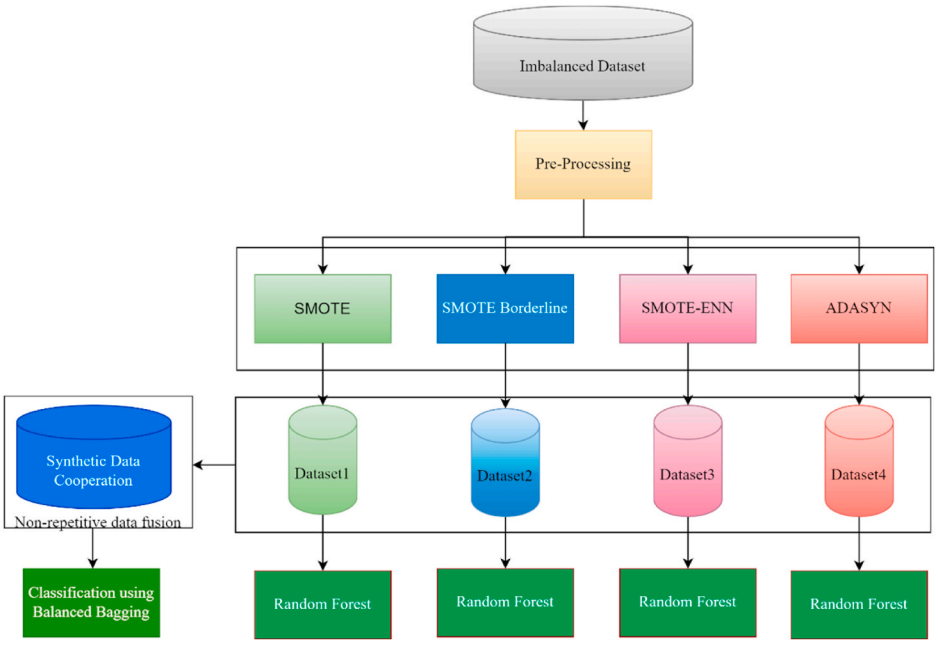



**Fig. 3.** General diagram of the *Equi-Fused-Data-based SMOTE* for the classification of imbalanced datasets.

incorporating different oversampling datasets into the balanced bagging ensemble there are several advantages.

- Improved representation: oversampling improves the representation of minority classes, allowing the ensemble to learn their features more effectively.
- Reduced bias: using multiple oversampling datasets reduces the risk of bias from individual techniques.
- Better generalizability: diversity in the ensemble leads to more robust and generalizable models, reducing the risk of overfitting to a single oversampled dataset.

This approach has its theoretical foundation in studies that show the efficacy of oversampling and ensemble learning in imbalanced learning. The works of (Chawla et al., 2002; Fernandez et al., 2018) demonstrate the benefits of SMOTE for imbalanced datasets, while (Barros et al., 2019) emphasized the benefits of ensemble methods such as balanced bagging, where the method outperformed multi-layer perceptron (MLP) and Decision Tree in performance on G-mean and Unbalanced Accuracy Ratio (UAR) metrics, avoiding the Accuracy paradox in educational dataset context. Furthermore, (Pristyanto et al., 2021) showed that combining oversampling techniques such as ADASYN with ensemble learning can effectively solve imbalanced problems. The Equi-Fused-Data-based SMOTE model was empirically evaluated on imbalanced multi-class datasets by comparing its performance to individual oversampling techniques and other ensemble methods. The model's effectiveness was evaluated using metrics such as AUC-ROC, F1-score, and overall Accuracy. By integrating these theoretical foundations, Equi-Fused-Data demonstrates that it is a theoretically sound and promising approach to overcoming the challenges of imbalanced multi-class learning.

### 3.4.3. Problem formulation

In the context of imbalanced multi-class learning, the Equi-Fused-Data-based SMOTE model addresses class imbalance by combining the benefits of oversampling and ensemble learning. To clarify the theoretical foundation of this approach, the problem must be mathematically defined, considering dataset combination and duplicate exclusion. Let *Di* (for $i = 1, 2, \ldots, n$) denote the dataset obtained from the *i-th* oversampling technique (e.g., SMOTE, SMOTE Borderline, SMOTE-ENN, and ADASYN). Each $D_i$ contains instances that belong to multiple classes ($c$ classes) with an imbalanced distribution. Here, $c$ denotes the number of classes, and $n$ is the total number of oversampling techniques used.

Each instance $x_j^i$ in $D_i$ is a $d$-dimensional feature vector associated with a class label $y_j^i \in \{1, 2, \ldots, c\}$.

Each oversampling technique generates a dataset $D_i$. Let *fi (x, y)* denote the oversampling function used by the $i$-th technique, where $x$ is a data point and $y$ is its class label.

$$D_i = f_i(x_j, y_j) \,|\, (x_j, y_j) \in D_{original}, j = 1, 2, \ldots, |D_{original}|$$

Where $D_{original}$ is the original imbalanced dataset.

The combination of these datasets involves merging them into a unified dataset, denoted as $D_{combined}$. This fusion aims to combine instances from all classes to ensure a more balanced representation of the entire spectrum of classes.

$$D_{combined} = \bigcup_{i=1}^{m} D_i$$

However, as there may be overlaps between the datasets, the resulting combined dataset may contain duplicate instances. To eliminate redundancies and maintain data integrity, redundant instances (duplicates) present in $D_{combined}$ are removed. The resulting unique dataset is denoted as $D_{unique}$. This process ensures that each instance in the dataset is unique and retains the diversity necessary for effective learning.

$$D_{\text{unique}} = D_{\text{combined}} \setminus \text{duplicates}$$

Furthermore, to leverage ensemble learning, balanced bagging is also used with a base classifier such as Random Forest. In this ensemble technique, multiple sub-models are created, each trained on a balanced subset of the data obtained by oversampling. Let ($B$=*{B1, B2, …,Bm}*) represent the ensemble of $m$ base classifiers, where each $B_i$ is trained on a balanced subset of $D_{unique}$. This balancing process ensures a fair representation of all classes during training.

$$B_i = TrainClassifier(D_{subset}, parameters)$$

where $D_{subset}$ is a balanced subset of $D_{unique}$ and *parameters* denote the hyperparameters of the Random Forest classifier.

#### 3.4.4. Implementation details

For this dataset, four balanced datasets were merged, yielding 8155 instances. To remove duplicates, the combined dataset was further processed, yielding a final size of 8136 instances. Then, each of these datasets was classified using Random Forest. Balanced bagging (Barros et al., 2019) was then implemented with a Random Forest classifier. The ensemble included 20 base estimators (trees). This decision was made to achieve a balance between model complexity and computational efficiency. The sampling strategy was set to "not majority', which means that the class distribution was balanced with the sampling technique to ensure equality during training. In addition, the sampling was done without replacement to ensure the dataset's integrity. The random seed for reproducibility was set to 42.

### 3.5. Training

#### 3.5.1. Hyperparameter tuning

In this study, data was classified using a variety of supervised machine learning algorithms, including SVM, Random Forest, Decision Tree, kNN, and Naive Bayes. Table 2 provides information on the algorithms and parameter tuning.

- SVM is a supervised machine learning algorithm that separates data points into different classes by finding an optimal hyperplane in a high-dimensional feature space (Cortes & Vapnik, 1995).
- Random Forest is an ensemble learning method that combines multiple Decision Trees to make predictions by averaging the outputs of each tree, providing robustness and reducing overfitting (Breiman, 2001).
- Decision Tree is a Tree-like model that uses a sequence of binary decisions based on input features to classify or predict outcomes by following a series of if-else conditions (Kumar et al., 2022).
- kNN is an instance-based learning algorithm that classifies new instances based on the majority vote of its k nearest neighbors in the training set (Quan, 2020).
- Naive Bayes is a probabilistic classifier that applies Bayes' theorem with the assumption of independence between features to predict the class probabilities of a given data sample (Saritas & Yasar, 2019).

#### 3.5.2. Cross-validation

To ensure consistent results and reduce overfitting, all algorithms underwent 10-fold cross-validation. The dataset was divided into ten subsets, nine for training and one for testing in each iteration. Similarly, the Equi-Fused-Data-based SMOTE model was 10-fold cross-validated on the combined synthetic dataset, with the same test set used for evaluation.

#### 3.5.3. Model's evaluation

The performance of each algorithm was evaluated using a range of evaluation metrics commonly employed in classification tasks: false positive rate (FPR), Accuracy, Precision, Recall, F1-score, and AUC-ROC.

- The False Positive rate is the proportion of incorrectly predicted positive instances out of all negative instances in all classes:

$$FPR = \frac{FP}{(FP + TN)}$$

where, *FP* is the number of false positives (incorrectly predicted positive instances) and *TN* are true negatives (correctly predicted negative instances).

- Accuracy is the overall correctness of the model's predictions:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

where, *TP* are true positives (correctly predicted positive instances) and *FN* are false negatives.

(incorrectly predicted negative instances).

- Precision represents the proportion of predicted positives that are truly positive:

$$Precision = \frac{TP}{(TP + FP)}$$

- Recall indicates the proportion of actual positive instances correctly identified by the model:

$$Recall = \frac{TP}{(TP + FN)}$$

- F1-score combines Precision and Recall into a single metric, providing a balanced view of the model's performance:

$$F1\ Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

- The AUC-ROC curve evaluates the model's ability to distinguish between different classes. We employed a combination of micro-AUC and weighted averaging for a multi-class classification problem.

The micro-AUC approach treats all classes equally, regardless of their distribution in the data. It considers all true positives and negatives from all classes to generate a single ROC curve and its corresponding AUC. The weighted AUC method first computes individual AUC values for each class using their ROC curves. These individual AUCs are then combined into a single weighted average, with weights assigned based on the relative frequency of each class in the training data. The result considers the prevalence of each class in the data to determine its significance. We combined the model's performance across all classes, taking into account both their individual discriminatory power (via micro-averaging) and their relative importance in the data (via weighted averaging), to gain a more nuanced understanding of the model's performance in a multi-class setting.

**Table 2**
Learning models' configuration.

| Model | Parameters |
|---|---|
| SVM | Kernel = RBF (Radial Basis Function), Probability = True |
| Random Forest | Number of estimators: 100, Criterion: Gini, Max Depth: None, Minimum sample per leaf: 1, Bootstrap samples: True |
| Decision Tree | Criterion: Gini, Splitter: best, Max Depth: None, Minimum samples per Split: 2, Minimum samples per leaf: 1 |
| kNN | K = 5, Distance metric: Euclidean distance |
| Naive Bayes | Gaussian Naive Bayes |

## 4. Results

### 4.1. Imbalanced dataset

In the first experiment, the original dataset for several algorithms yielded reasonable results, ranging from 89.15% to 92.85%. However, it had a classification problem, with the "High" and "Moderate" classes consistently misclassified. Table 3 displays the results, while Fig. 4 depicts a comparative analysis of various models applied to an imbalanced dataset. The SVM model achieved an Accuracy, Precision, Recall and F1-score of 91%, an FPR of 6.32% and an excellent AUC of 98.19%. In contrast, the Random Forest, Naive Bayes, and kNN models achieved accuracies ranging from 89.15% to 92.85%, with Precision, Recall, and F1-score values ranging from 89.12% to 92.94%. Finally, the Decision

**Table 3**
Performance comparison of machine learning models on the imbalanced dataset.

| Model | Accuracy | FPR | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|---|
| SVM | 91.79% | 6.32% | 91.85% | 91.80% | 91.55% | 98.19% |
| Random Forest | 91.70% | 6.91% | 91.75% | 91.79% | 91.74% | 97.24% |
| Decision Tree | 89.15% | 8.62% | 89.12% | 89.15% | 89.13% | 87.06% |
| kNN | 92.85% | 5.56% | 92.94% | 92.85% | 92.84% | 96.38% |
| Naive Bayes | 92.06% | 6.55% | 92.08% | 92.05% | 92.02% | 90.17% |

Tree had the lowest values for Accuracy, Precision, and Recall, with an F1-score of around 89% and an AUC of 87.06%. The findings show that while the SVM had the highest AUC 98.19% and overall performance, it is important to consider the potential biases caused by the imbalanced dataset. In addition, all classifiers except the Decision Tree had relatively high AUC scores, ranging from 90.17% to 98.19%.

### 4.2. SMOTE dataset

Our second experiment aimed to investigate the SMOTE technique's performance for various classifiers. The results detailed in Table 4 lead to the following conclusions. First, using SMOTE to balance the dataset improves Accuracy, reduces false positives by 4.07%–6.06% for most models, with the exception of the Decision Tree (7.92%), and maintains AUC scores for the majority of models, as shown in Fig. 5. For example, when applied to the imbalanced dataset, the SVM model achieved 94% Accuracy, Recall, and F1-score, as opposed to 91%. However, SMOTE introduces randomness into the generation of synthetic data points, which can result in slightly different results each time it is used. Second, slightly better results were observed for the Decision Tree model, with Accuracy, Precision, Recall, and F1-score of around 90%. Overall, balancing data with SMOTE improved model performance in the majority of cases.

### 4.3. SMOTE borderline dataset

Our third experiment investigated the effectiveness of SMOTE Borderline on various models. Table 5 summarizes the results. It can be observed that while both SMOTE and SMOTE Borderline generally outperformed the imbalanced data, some other interesting findings emerged. First, the SVM model improved the most with SMOTE Borderline, achieving higher Accuracy (94.17%), lower FPR (3.31%) and maintaining AUC (97.98%) compared to the imbalanced (91%, 6.32%, and 0.98%). This suggests that SMOTE Borderline may be especially useful for models that are sensitive to classifying data points near class boundaries, such as SVM. Second, other models produce mixed results. While some models, such as Random Forest and kNN, maintain similar performance with Accuracy, Precision, Recall, and an F1-score of 91%, a positive rate of 5% with both SMOTE techniques, others, such as Naive Bayes and Decision Trees, show slightly lower Accuracy with SMOTE Borderline than with SMOTE. These findings show that the effectiveness of SMOTE Borderline varies depending on the algorithm used and its sensitivity to the underlying imbalance, even after using balancing techniques. Fig. 6 depicts the SMOTE Borderline performance.

### 4.4. SMOTE-ENN dataset

The results in Table 6 demonstrate that SMOTE-ENN outperforms SMOTE and SMOTE Borderline. In all models, SMOTE-ENN outperforms the other two techniques in terms of Accuracy, Precision, Recall, F1-score, and AUC. This indicates that SMOTE-ENN enhances overall classification performance. First, SVM, like SMOTE Borderline, improves the most with SMOTE-ENN, reaching an Accuracy of 93.65% and an AUC of 98.2%. This finding lends credence to the idea that techniques focusing on borderline samples could be particularly useful for SVM. Second, other models show consistent improvements. Although the degree of improvement varies, all models exhibit higher performance metrics with SMOTE-ENN than with SMOTE and SMOTE Borderline, as

**Table 4**
Performance comparison of machine learning models on the SMOTE balanced dataset.

| Model | Accuracy | FPR | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|---|
| SVM | 94.17% | 4.07% | 94.86% | 94.17% | 94.29% | 98.21% |
| Random Forest | 92.06% | 6.06% | 92.09% | 92.06% | 92.06% | 97.06% |
| Decision Tree | 90.74% | 7.92% | 90.66% | 90.74% | 90.65% | 88.12% |
| kNN | 92.32% | 4.70% | 92.92% | 92.30% | 92.45% | 96.59% |
| Naive Bayes | 90.47% | 4.99% | 92.55% | 90.40% | 91% | 92.51% |

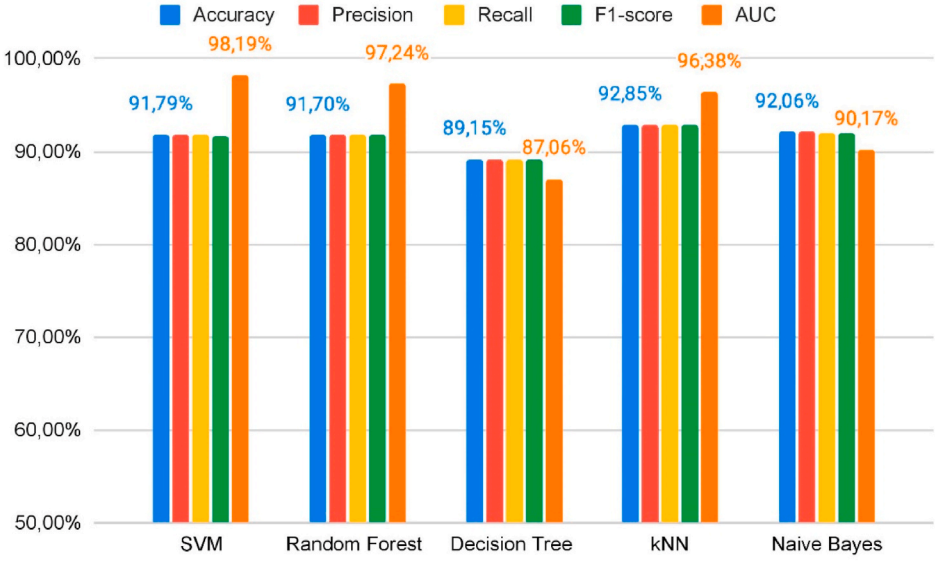



**Fig. 4.** Classification models' performance on the imbalanced dataset.

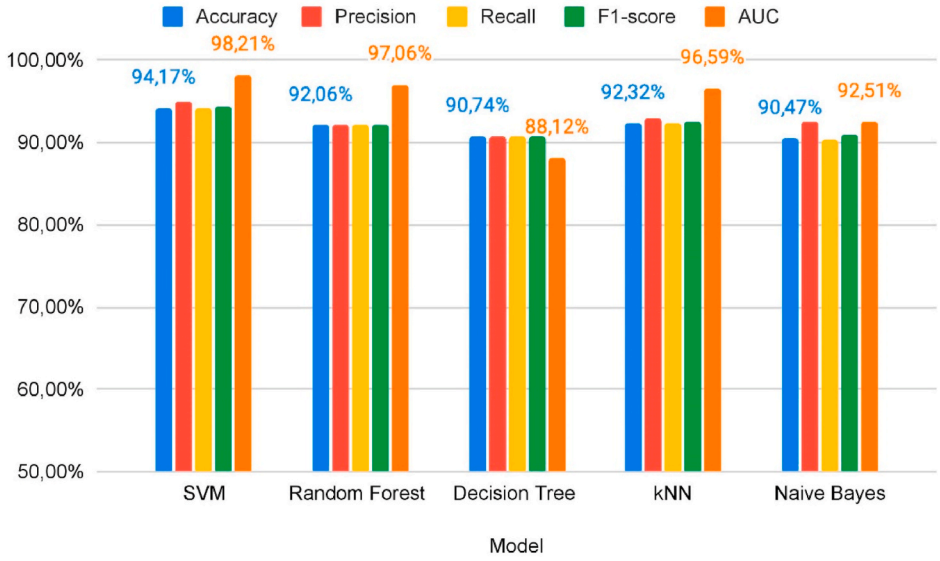



**Fig. 5.** Classification models' performance on the SMOTE balanced dataset.

**Table 5**
Performance comparison of machine learning models on the SMOTE Borderline balanced dataset.

| Model | Accuracy | FPR | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|---|
| SVM | 94.17% | 3.31% | 95.12% | 94.18% | 94.31% | 97.98% |
| Random Forest | 91.79% | 5.89% | 91.91% | 91.70% | 91.82% | 97.24% |
| Decision Tree | 89.41% | 8.46% | 89.40% | 89.41% | 89.39% | 87.30% |
| kNN | 91.53% | 5.07% | 92.32% | 91.50% | 91.69% | 96.24% |
| Naive Bayes | 87.03% | 6.07% | 90.88% | 87% | 87.70% | 90.89% |

shown in Fig. 7. SMOTE-ENN addresses potential issues related to SMOTE and SMOTE Borderline. While SMOTE has the risk of over-sampling irrelevant data, and SMOTE Borderline may not be effective in all borderline cases, SMOTE-ENN combines oversampling and nearest neighbor selection. This selection process introduces more relevant and representative data points than random oversampling in SMOTE, potentially leading to improved model learning and performance.

### 4.5. ADASYN dataset

In the fifth experiment, the performance of the ADASYN technique was assessed using various classifiers. Based on the results in Table 7, it

**Table 6**
Performance comparison of machine learning models on the SMOTE-ENN balanced dataset.

| Model | Accuracy | FPR | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|---|
| SVM | 93.65% | 3.11% | 94.7% | 93.65% | 93.80% | 98.20% |
| Random Forest | 90.74% | 5.92% | 91.24% | 90.70% | 90.84% | 97.19% |
| Decision Tree | 90.74% | 6.13% | 91.25% | 90.70% | 90.87% | 90.80% |
| kNN | 92.06% | 4.66% | 92.83% | 92% | 92.20% | 95.43% |
| Naive Bayes | 90.21% | 4.42% | 92.99% | 92.20% | 90.77% | 93.36% |

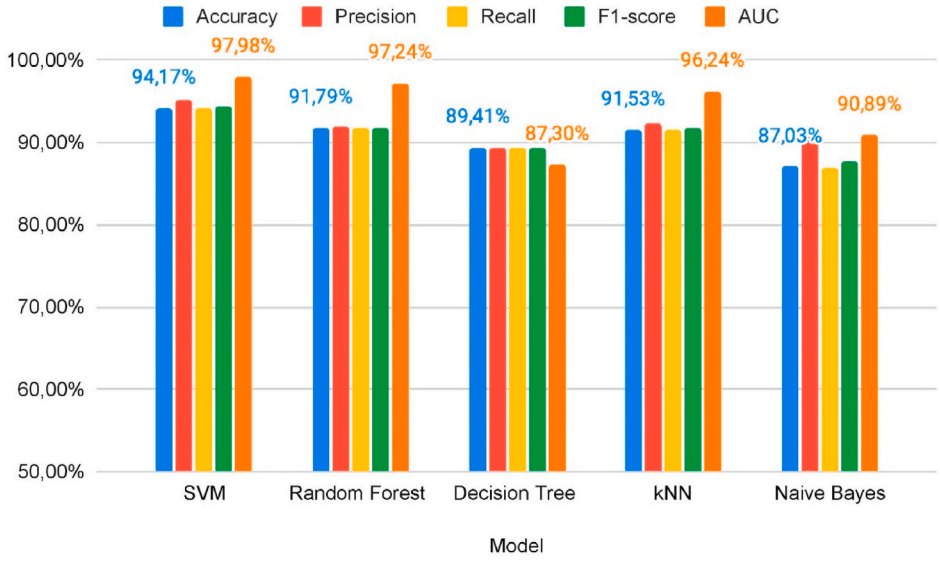



**Fig. 6.** Classification models' performance on the SMOTE Borderline balanced dataset.

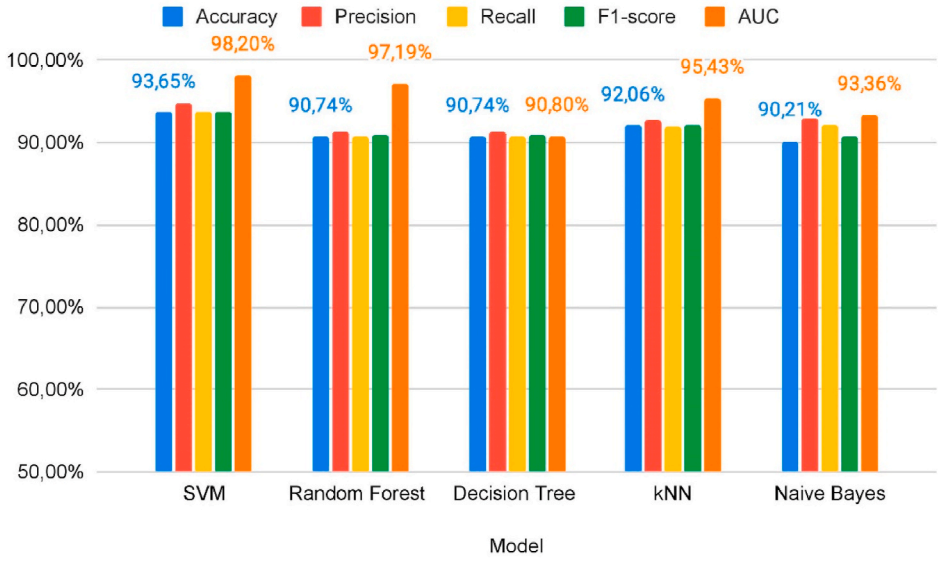



**Fig. 7.** Classification models' performance on the SMOTE-ENN balanced dataset.

can be noted that ADASYN performs similarly with SVM in terms of Accuracy, Precision, Recall, and F1-score (93.91%). While some models, such as Random Forest and kNN, perform similarly across all techniques, others, such as Naive Bayes and Decision Tree, have slightly lower Accuracy of 89.41% for Naive Bayes and higher FPR of 4.93%–7.07% with ADASYN when compared to SMOTE-ENN (90%, 4.42%–6.13%). As previously stated, SMOTE-ENN's selection of informative neighbors during oversampling may result in more relevant data points, leading to improved performance in most cases. On the other hand, ADASYN focuses on oversampling instances of the minority class based on their difficulty level, which may benefit models such as SVM that struggle with imbalanced data, as illustrated in Fig. 8. However, it may not work as well for all models or data characteristics. Indeed, different models may have varying sensitivities to the specific data distributions generated by each balancing technique, which explains why some models produce mixed results with ADASYN.

### 4.6. Equi-fused-data-based SMOTE dataset

The experimental results in Table 8 demonstrate the Balanced Bagging model's promising performance with the Equi-Fused-Data-based SMOTE technique. This approach outperformed several metrics, including Accuracy (93.85%), Precision (92.86%), Recall (92.8%), F1-score (92.86%), and AUC (98.08%), all while maintaining a low FPR (5.35%). These metrics collectively indicate good classification performance. This approach's strength stems from its ability to combine multiple oversampled versions of data, potentially capturing a broader range of patterns and improving class balance. Notably, the performance of balanced bagging with a Random Forest classifier is consistent with that of the SMOTE-ENN technique, which was previously identified as the most effective technique for this particular dataset. These findings indicate that the Equi-Fused-Data-based SMOTE model has significant potential as a viable strategy for addressing multi-class classification problems.

**Table 7**
Performance comparison of machine learning models on the ADASYN balanced dataset.

| Model | Accuracy | FPR | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|---|
| SVM | 93.91% | 2.84% | 94.94% | 93.91% | 94.05% | 98.14% |
| Random Forest | 92.32% | 5.60% | 92.43% | 92.32% | 92.34% | 97.15% |
| Decision Tree | 91% | 7.07% | 91% | 91% | 91% | 89.39% |
| kNN | 92.06% | 4.83% | 92.71% | 92.06% | 92.20% | 96.67% |
| Naive Bayes | 89.41% | 4.93% | 92.41% | 89.42% | 90% | 92.60% |

The results in Table 9 and Fig. 9 compare Random Forest's performance on various datasets: imbalanced, oversampled with SMOTE, SMOTE Borderline, SMOTE-ENN, ADASYN, and the novel Equi-Fused-Data-based SMOTE. Notably, the Equi-Fused-Data-based SMOTE model outperforms the imbalanced and oversampled datasets, with an accuracy of (93.85%) versus (91.70%) and approximately (91%), respectively. The results also show a higher Precision, Recall, and F1-score (92%), the lowest FPR (5.35%), and a higher AUC (98.08%). This model may overcome the limitations of single oversampling techniques by incorporating diverse synthetic samples derived from multiple versions. This comprehensive representation of the data most likely improves model learning. Furthermore, the Equi-Fused-Data strategy is specifically designed for multi-class problems, which may provide advantages over techniques based primarily on binary classification.

## 5. Discussion

### 5.1. Class imbalance in student performance assessment

As shown by the first experiment, class imbalance is a significant challenge in the field of education, where datasets frequently show an uneven distribution of student performance categories. This imbalance biases machine learning models toward the majority class, resulting in misleadingly high overall Accuracy but potentially poor performance in identifying minority-class students. These findings emphasize the critical need to account for class imbalance when using machine learning for educational assessment.

### 5.2. Effectiveness of oversampling techniques

The first objective of this study is to demonstrate the efficacy of various oversampling techniques in addressing class imbalance issues in the educational dataset. The results show that the SMOTE, SMOTE Borderline, SMOTE-ENN, and ADASYN techniques are effective at addressing the class imbalance problem, resulting in consistent performance across training and test sets and preventing overfitting. The proposed analysis demonstrates that SMOTE produces promising results, particularly for Random Forest and SVM models, which have significant

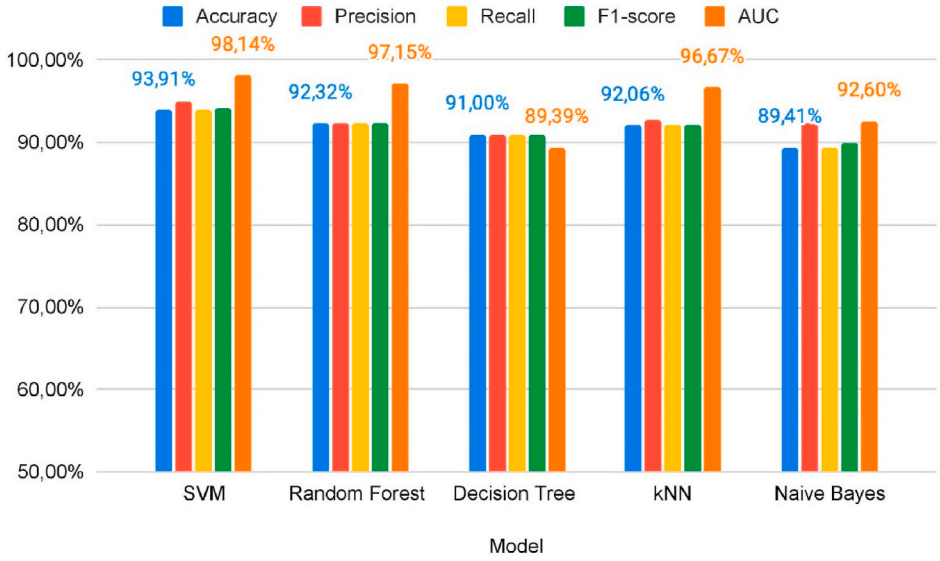



**Fig. 8.** Classification models' performance on the ADASYN balanced dataset.

**Table 8**
Balanced bagging performance on the Equi-Fused-Data-based SMOTE balanced dataset.

| Metric | Value |
|---|---|
| Accuracy | 93.85% |
| FPR | 5.35% |
| Precision | 92.86% |
| Recall | 92.80% |
| F1-score | 92.86% |
| AUC | 98.08% |

**Table 9**
Comparative analysis of classification models with imbalanced, oversampled, and Equi-Fused-Data datasets.

| Dataset | Accuracy | FPR | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|---|
| Imbalanced | 91.70% | 6.91% | 91.75% | 91.79% | 91.74% | 97.24% |
| SMOTE | 92.06% | 6.06% | 92.09% | 92.06% | 92.06% | 97.06% |
| SMOTE Borderline | 91.79% | 5.89% | 91.91% | 91.70% | 91.82% | 97.24% |
| SMOTE-ENN | 90.74% | 5.92% | 91.24% | 90.70% | 90.84% | 97.19% |
| ADASYN | 92.32% | 5.60% | 92.43% | 92.32% | 92.34% | 97.15% |
| Equi-Fused-Data | 93.85% | 5.35% | 92.86% | 92.80% | 92.86% | 98.08% |

AUC values (e.g., 98% for SVM). This finding is consistent with the findings of other studies, such as (Rozi et al., 2023; Tariq et al., 2023; Wongvorachan et al., 2023), which reported reasonable performance of SMOTE while emphasizing that it is dependent on factors such as data distribution, noise level, and the classifier used. For example, in this dataset, the combination of SMOTE and SVM achieved the best performance, whereas, in the study by (Tariq et al., 2023), the best classifier was kNN, implying that different classifiers may respond differently to oversampling. Some algorithms, such as kNN that rely on local distances may benefit more from SMOTE because it provides a clearer decision boundary. Furthermore, in the work of (Khalaf Hamoud et al., 2022), Random Forest outperformed the other algorithms in terms of Precision, Recall, and F1-score (81%). Random Forest produced consistent results (92%), indicating that the use of SMOTE improved the model's performance in many cases, but its effectiveness remains dependent on a variety of factors.

Similarly, the SMOTE Borderline improved the efficiency of SVM models by increasing Accuracy, precision, recall, F1-score, and AUC. However, its effectiveness varies by algorithm, implying that it is not inherently superior to SMOTE. Further research into its application to various educational datasets is needed. SMOTE-ENN outperformed other techniques, achieving impressive AUC values for both SVM and Random Forest models (98.2% and 97.19%, respectively). Furthermore, it produced lower FPR, demonstrating its efficacy in improving classification across multiple models. Finally, ADASYN provided satisfactory performance results. While most models achieved AUC values ranging from 92.6% to 98.14%, the Decision Tree classifier had an even lower AUC value (89.39%). This implies that ADASYN may be less effective for some algorithms than other methods. Similar to the results of (Rozi et al., 2023), ADASYN improved classification performance, but its effectiveness, like SMOTE's, is dependent on the specific data characteristics and models used. For example, in this dataset, ADASYN performed best with SVM, Random Forest, and kNN, whereas in the work of (Rozi et al., 2023), the stacking algorithm performed best overall. Figs. 10 and 11 show a comprehensive comparison of the Accuracy and AUC of the various models. Most models, with the exception of Naive Bayes, have higher overall Accuracy when using oversampling techniques. In terms of AUC, these techniques produce consistent or better results across all models.

### 5.3. *Equi-fused-data-based SMOTE*

The second objective of this study was to investigate the effectiveness of a non-redundant data cooperation model for improving overall performance. The proposed Equi-Fused-Data-based SMOTE model effectively addresses the class imbalance issue, as shown in Table 9, and outperforms other techniques in terms of Accuracy, Precision, Recall, F1-Score over (92%), AUC (98.08%), and FPR (5.35%) (Fig. 12). Thus, these results demonstrate the potential benefits of combining elements from different oversampling techniques to produce a more comprehensive and effective solution for identifying minority classes in educational datasets. Furthermore, the stability of balanced bagging observed in the experiments suggests that ensemble techniques continue to be effective in dealing with class imbalances in multi-class problems by providing robust and stable predictions, as supported by studies such as (Barros et al., 2019; Pristyanto et al., 2021). This finding highlights the potential of this method for educational assessment.

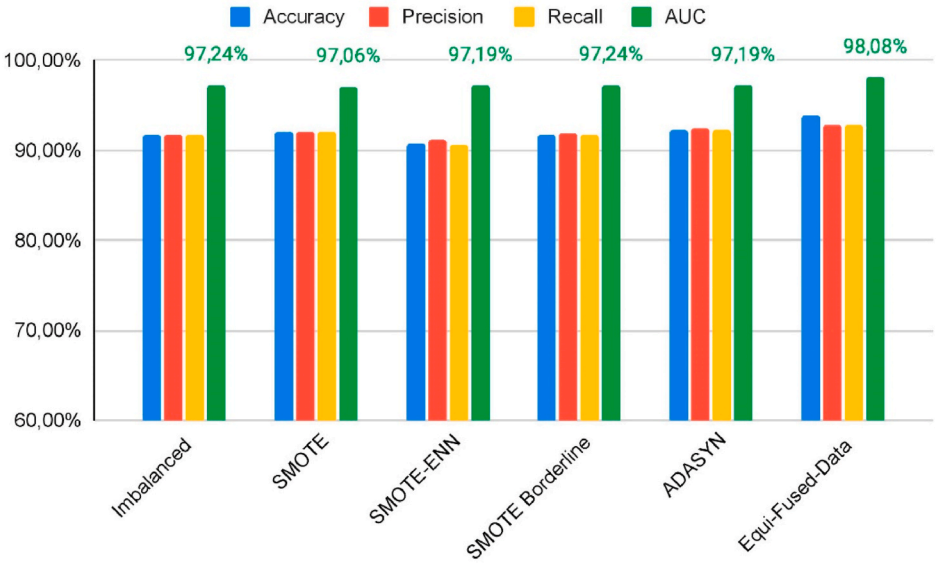



**Fig. 9.** Comparative analysis of balanced bagging vs. Random Forest on different datasets.

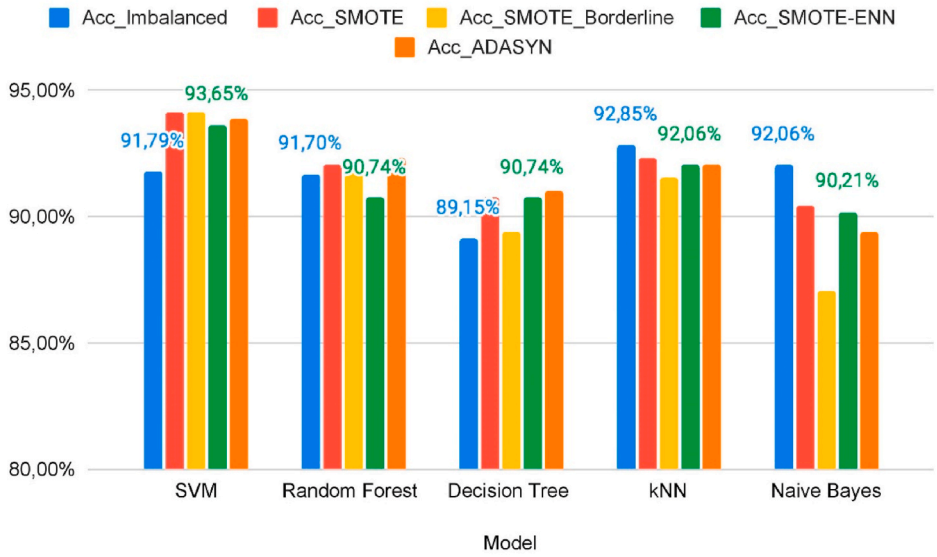



**Fig. 10.** Accuracy comparison: imbalanced dataset vs oversampled balanced datasets.

### 5.4. Educational implication

The findings show that the proposed Equi-Fused-Data-based SMOTE model, as well as oversampling techniques and classification models, are effective in addressing class imbalance in educational assessments. These findings highlight the importance of considering not only Accuracy but also Precision, Recall, FPR, and AUC when assessing model performance. This is especially important in education, where errors in classification can have serious consequences for students. These findings are promising for educators because they can help to develop more accurate tools for assessing student performance and progress. Early identification of students at risk of failure or in need of additional assistance can result in targeted interventions and, ultimately, better educational outcomes.

### 5.5. Limitations

While the proposed Equi-Fused-Data-based SMOTE and balanced bagging method demonstrated promising results in addressing class imbalance for educational assessment, some limitations must be addressed. The current study utilized a specific dataset that concentrated on a specific educational context (algorithmic performance) and student population. The applicability to other educational settings with different student populations, learning objectives, and assessment types requires further investigation. The dataset's size (2176 instances) limits the statistical results' generalizability and necessitates validation using larger and more diverse datasets. While the oversampling techniques used in this study, including SMOTE, SMOTE Borderline, SMOTE-ENN, and ADASYN, were effective, they did have potential limitations, such as introducing their own biases into the data, which could affect the model's generalizability.

Furthermore, as previously stated, their efficacy varies depending on the specific data distribution and characteristics of class imbalance. Investigating alternative approaches, such as cost-sensitive learning, may provide useful insights into their efficacy in the context of educational assessment. While balanced bagging performs well, it is computationally expensive when compared to simpler oversampling techniques, particularly when dealing with large datasets. Investigating

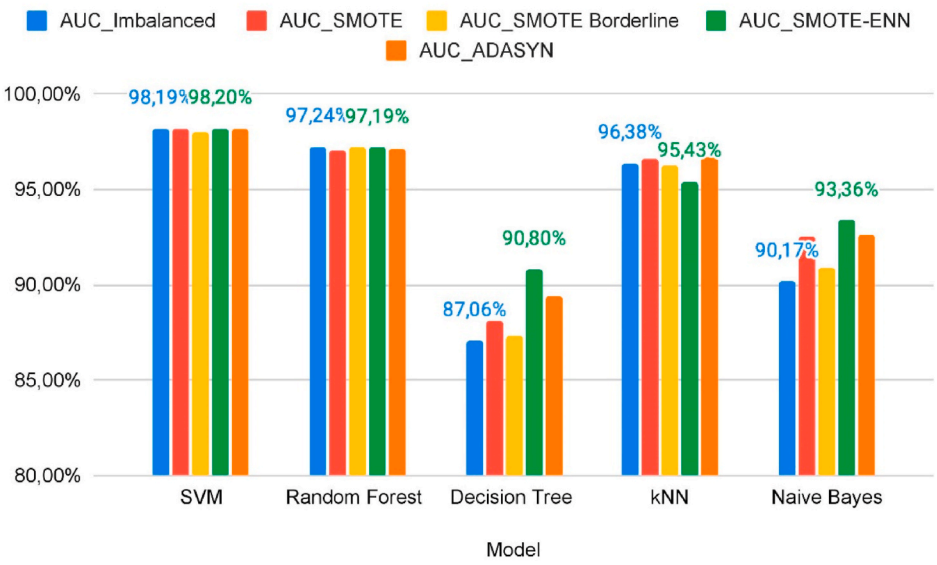



**Fig. 11.** AUC comparison: imbalanced dataset vs oversampled balanced datasets.

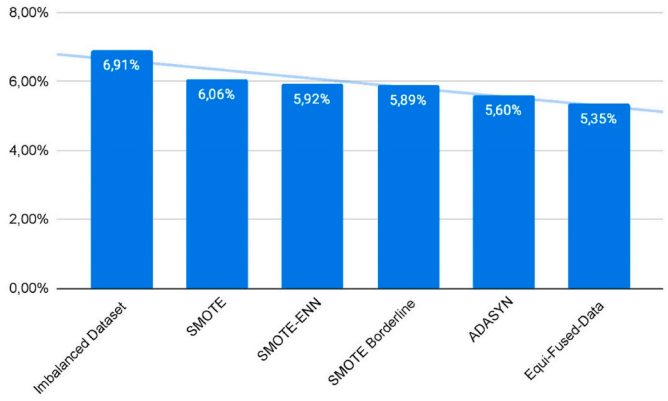



**Fig. 12.** Comparative analysis of false positive rates across different oversampled datasets.

optimization strategies to improve scalability is, therefore, critical. In addition, this study primarily addressed class imbalances using oversampling techniques. However, incorporating domain-specific knowledge into the assessment process, such as student demographics or cognitive profiles, has the potential to improve model accuracy and fairness.

## 6. Conclusion

This study addressed the critical issue of class imbalance in educational datasets, specifically focusing on a dataset of student performance in introductory programming. Unaddressed imbalance can result in biased and unreliable assessment results. We investigated the efficacy of several sampling techniques in mitigating this imbalance in the context of multi-class classification. These techniques included SMOTE, SMOTE Borderline, SMOTE-ENN, ADASYN, and the recently proposed Equi-Fused-Data-based SMOTE model. Three main research questions guided this investigation.

1. Which sampling technique (SMOTE, SMOTE Borderline, SMOTE-ENN, or ADASYN) is most effective for this specific dataset?

Our comparison of different techniques, including SMOTE, SMOTE Borderline, SMOTE-ENN, and ADASYN, revealed that SMOTE, SMOTE Borderline and ADASYN were successful in mitigating class imbalance and improving classification model performance. These approaches enabled the models to learn from a more balanced dataset, resulting in improved generalization and higher performance. However, SMOTE-ENN demonstrated promise by producing encouraging results while preserving the dataset's global information, thereby improving overall performance.

2. How does class imbalance impact the performance of various classification algorithms?

We found that the different classification algorithms respond differently to class imbalance, a result that is consistent with previous studies (Rozi et al., 2023; Tariq et al., 2023). SVM and Random Forest were found to be suitable choices for this particular problem and robust to imbalanced data.

3 Can the Equi-Fused-Data-based SMOTE model, a non-redundant data cooperation approach, outperform traditional oversampling techniques in terms of classification model Accuracy, F1-score, and AUC when applied to an imbalanced educational dataset from an introductory programming class?

The Equi-Fused-Data-based SMOTE model tackles class imbalance in multi-class classification while delivering high performance. It shows significant improvements in Accuracy (93.85%), Precision, Recall, F1-score (92%), and AUC (98.08%). Additionally, it has a lower FPR (5.35%) than individual sampling methods. The low false positive rate reduces the number of misclassified instances in the minority class, resulting in more accurate and equitable classification, which is especially important in education. In contrast to previous studies (Khalaf Hamoud et al., 2022; Wongvorachan et al., 2023) that used single sampling methods, the proposed model employs a data collaboration strategy. This method combines the advantages of several sampling techniques to produce a more robust solution for identifying minority classes. As a result, it effectively addresses the issue of underrepresentation of minority classes while also improving classification performance.

Based on the findings highlighted, this study makes two major contributions. First, the Equi-Fused-Data-based SMOTE model was introduced, which marks a significant step forward in addressing class imbalances in educational datasets. This builds on previous research on

reducing class imbalances and contributes to ongoing efforts to improve learning outcomes in programming education. Beyond Accuracy, the findings highlight the importance of incorporating multiple assessment metrics when dealing with class imbalances. Furthermore, we discuss the potential benefits of combining elements from different over-sampling techniques to develop a more comprehensive and effective solution for identifying minority classes in educational datasets. This is a promising direction for future research, with potential applications in real-world educational assessment scenarios.

Building on the Equi-Fused-Data model's findings, a comprehensive evaluation of various sampling techniques yields valuable insights for future research on algorithmic learning assessment methods. This is accomplished by emphasizing the critical importance of taking into account class imbalances when developing more equitable and reliable assessments in programming education. Furthermore, it paves the way for the development of more effective and reliable assessment tools, which will lead to better learning outcomes in programming education. While limitations such as generalizability and scalability warrant further investigation, this study paves the way for future advances in personalized learning approaches. For future work, we propose using the Equi-Fused-Data-based SMOTE to recommend personalized learning resources tailored to individual student needs, which could lead to further improvement in learning outcomes.

## Statement on open data and ethics

The study was approved by an ethical committee with ID: 2024/6,024,753. Informed consent was obtained from all participants, and their privacy rights were strictly observed. The participants were protected by hiding their personal information during the research process. They knew that the participation was voluntary and they could withdraw from the study at any time. There is no potential conflict of interest in this study. The data can be obtained by sending request e-mails to the corresponding author with permission of Badji Mokhtar Annaba University, Annaba, Algeria.

## Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## Author contributions

Richard Hotte: Writing – original draft, Validation, Supervision, Writing – review & editing. Tahar Bensebaa: Validation, Supervision, Conceptualization. Yasmine Chachoui.: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization, Data curation. Nabiha Azizi: Writing – review & editing, Validation, Supervision, Methodology, Formal analysis, Conceptualization

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## List of Acronyms

ADASYN Adaptative Synthetic Sampling
AUC –ROC Area Under the Receiver Operating Characteristic Curve
CS1 Computer Science 1
CS2 Computer Science 2
FPR False Positive Rate
kNN K-Nearest Neighbors
ROS Random Over-Sampling
RUS Random Under-Sampling
SMOTE Synthetic Minority Over-sampling Technique
SMOTE-ENN SMOTE Edited Nearest Neighbors
SVM Support vector Machine

## References

Abdessemed, M., Bensebaa, T., Belhaoues, T. E., & Bey, A. (2018). Automatic exercise sequencing-based algorithmic skills. *International Journal of Innovation and Learning, 23*(1), 104–121. https://doi.org/10.1504/ijil.2018.088788

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining educational data to predict student's academic performance using ensemble methods. *International Journal of Database Theory and Application, 9*(8), 119–136. https://doi.org/10.14257/ijdta.2016.9.8.13

Barros, T. M., SouzaNeto, P. A., Silva, I., & Guedes, L. A. (2019). Predictive models for imbalanced data : A school dropout perspective. *Education Sciences, 9*(4), 275. https://doi.org/10.3390/educsci9040275

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote : Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

Feng, S., Keung, J., Yu, X., Xiao, Y., & Zhang, M. (2021). Investigation on the stability of SMOTE-based oversampling techniques in software defect prediction. *Information and Software Technology, 139*. https://doi.org/10.1016/j.infsof.2021.106662

Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data : Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research, 61*, 863–905. https://doi.org/10.1613/jair.1.11192

Gross, P., & Powers, K. (2005). Evaluating assessments of novice programming environments. In *Proceedings of the 2005 international workshop on computing education research - icer '05* (pp. 99–110). https://doi.org/10.1145/1089786.1089796

Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-SMOTE : A new over-sampling method in imbalanced data sets learning. In D.-S. Huang, X.-P. Zhang, & G.-B. Huang (Eds.), *Advances in intelligent computing* (Vol. 3644, pp. 878–887). Springer Berlin Heidelberg. https://doi.org/10.1007/11538059_91.

He, H., Yang, B., Garcia, E. A., & Shutao, L. (2008). Adasyn : Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE international joint conference on neural networks* (pp. 1322–1328). IEEE World Congress on Computational Intelligence). https://doi.org/10.1109/IJCNN.2008.4633969

Intayoad, W., Kamyod, C., & Temdee, P. (2019). Synthetic minority over-sampling for improving imbalanced data in educational web usage mining. *ECTI Transactions on Computer and Information Technology, 12*(2), 118–129. https://doi.org/10.37936/ecti-cit.2018122.133280

Khalaf Hamoud, A., Baqr Mohammed Kamel, M., Sahl Gaafar, A., Salah Alasady, A., Majeed Humadi, A., Akeel Awadh, W., & Mohammed Dahr, J. (2022). A prediction model based machine learning algorithms with feature selection approaches over imbalanced dataset. *Indonesian Journal of Electrical Engineering and Computer Science, 28*(2), 1105–1116. https://doi.org/10.11591/ijeecs.v28.i2.pp1105-1116

Kirschner, P. A. (2017). Stop propagating the learning styles myth. *Computers & Education, 106*, 166–171. https://doi.org/10.1016/j.compedu.2016.12.006

Krawczyk, B. (2016). Learning from imbalanced data : Open challenges and future directions. *Progress in Artificial Intelligence, 5*(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

Kumar, M., Bajaj, K., Sharma, B., & Narang, S. (2022). A comparative performance assessment of optimized multilevel ensemble learning model with existing classifier models. *Big Data, 10*(5), 371–387. https://doi.org/10.1089/big.2021.0257

Lahtinen, E., Kirsti, A.-M., & Hannu-Matti, J. (2005). A study of the difficulties of novice programmers. In *Proceedings of the 10th annual SIGCSE conference on innovation and technology in computer science education (ITiCSE '05)* (pp. 14–18). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1067445.1067453.

Lu, J., Yu, C.-S., & Liu, C. (2003). Learning style, learning patterns, and learning performance in a WebCT-based MIS course. *Inf Manage, 420*(3), 497–507. https://doi.org/10.1016/S0378-7206(02)00064-2

McCall, D., & Kölling, M. (2019). A new look at novice programmer errors. *ACM Transactions on Computing Education, 19*(4), 1–30. https://doi.org/10.1145/3335814

Nabil, A., Seyam, M., & Abou-Elfetouh, A. (2021). Prediction of students' academic performance based on courses' grades using deep neural networks. *IEEE Access, 9*, 140731–140746. https://doi.org/10.1109/ACCESS.2021.3119596

Nabus, H. S. Y., Ali, A., Hassan, S., Shamsuddin, S. M., Mustapha, I. B., & Saeed, F. (2022). Adaptive generation-based approaches of oversampling using different sets of base and nearest neighbor's instances. *International Journal of Advanced Computer Science and Applications, 13*(4). https://doi.org/10.14569/IJACSA.2022.0130461

Nancekivell, S. E., Shah, P., & Gelman, S. A. (2020). Maybe they're born with it, or maybe it's experience : Toward a deeper understanding of the learning style myth. *Journal of Educational Psychology, 112*(2), 221–235. https://doi.org/10.1037/edu0000366

Pillay, N. (2003). Developing intelligent programming tutors for novice programmers. *ACM SIGCSE Bulletin, 35*(2), 78–82. https://doi.org/10.1145/782941.782986

Pillay, N., & Vikash, R. J. (2005). An investigation into student characteristics affecting novice programming performance. *ACM SIGCSE Bulletin, 37*(4), 107–110. https://doi.org/10.1145/1113847.1113888

Price, T. W., & Barnes, T. (2015). Comparing textual and block interfaces in a novice programming environment. In *Proceedings of the eleventh annual international conference on international computing education research - icer* (pp. 91–99). https://doi.org/10.1145/2787622.2787712, 15.

Pristyanto, Y., Nugraha, A. F., Pratama, I., Dahlan, A., & Wirasakti, L. A. (2021). Dual approach to handling imbalanced class in datasets using oversampling and ensemble learning techniques. In *2021 15th international conference on ubiquitous information management and communication (IMCOM)* (pp. 1–7). https://doi.org/10.1109/IMCOM51814.2021.9377420

Quan, Y. (2020). Development of computer aided classroom teaching system based on machine learning prediction and artificial intelligence KNN algorithm. *Journal of Intelligent and Fuzzy Systems, 39*(2), 1879–1890. https://doi.org/10.3233/JIFS-179959

Rachburee, N., & Punlumjeak, W. (2021). Oversampling technique in student performance classification from engineering course. *International Journal of Electrical and Computer Engineering, 11*(4), 3567–3574. https://doi.org/10.11591/ijece.v11i4.pp3567-3574

Radwan, A. M., & Cataltepe, Z. (2017). Improving performance prediction on education data with noise and class imbalance. *Intelligent Automation & Soft Computing*, 1–8. https://doi.org/10.1080/10798587.2017.1337673

Rozi, A. F., Wibowo, A., & Warsito, B. (2023). Resampling technique for imbalanced class handling on educational dataset. *JUITA: Jurnal Informatika, 11*(1), 77. https://doi.org/10.30595/juita.v11i1.15498

Saritas, M. M., & Yasar, A. (2019). Performance analysis of ANN and naive Bayes classification algorithm for data classification. *International Journal of Intelligent Systems and Applications in Engineering, 7*(2), 88–91. https://doi.org/10.18201/ijisae.2019252786

Sharma, R. M., Agrawal, C. P., Kumar, V., & Mulatu, A. N. (2022). CFSBFDroid : Android malware detection using CFS + best first search-based feature selection. *Mobile Information Systems, 2022*, 1–15. https://doi.org/10.1155/2022/6425583

Sim, T. Y., & Lau, S. L. (2018). Online tools to support novice programming : A systematic review. In *2018 IEEE Conference on E-Learning, e-Management and e-Services (IC3e)* (pp. 91–96). https://doi.org/10.1109/IC3e.2018.8632649

Tarekegn, A. N., Giacobini, M., & Michalak, K. (2021). A review of methods for imbalanced multi-label classification. *Pattern Recognition, 118*. https://doi.org/10.1016/j.patcog.2021.107965

Tariq, M. A., Sargano, A. B., Iftikhar, M. A., & Habib, Z. (2023). Comparing different oversampling methods in predicting multi-class educational datasets using machine learning techniques. *Cybernetics and Information Technologies, 23*(4), 199–212. https://doi.org/10.2478/cait-2023-0044

Wang, S., Dai, Y., Shen, J., & Xuan, J. (2021). Research on expansion and classification of imbalanced data based on SMOTE algorithm. *Scientific Reports, 11*. https://doi.org/10.1038/s41598-021-03430-5

Wilkinson, T., Boohan, M., & Stevenson, M. (2014). Does learning style influence academic performance in different forms of assessment? *Journal of Anatomy, 224*(3), 304–308. https://doi.org/10.1111/joa.12126

Wongvorachan, T., He, S., & Bulut, O. (2023). A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. *Information, 14*(1), 54. https://doi.org/10.3390/info14010054

Yilmaz, R., & Karaoglan Yilmaz, F. G. (2023). The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation. *Computers and Education: Artificial Intelligence, 4*, Article 100147. https://doi.org/10.1016/j.caeai.2023.100147