

## Article

# Enhancing Early Academic Outcome Prediction in Small Educational Datasets Through Data Augmentation Techniques

Said El Kafhali \*  and Zakaria Soufiane Hafdi

Computer, Networks, Modeling, and Mobility Laboratory (IR2M), Faculty of Sciences and Techniques, Hassan First, University of Settat, Settat 26000, Morocco; z.hafdi@uhp.ac.ma

\* Correspondence: said.elkafhali@uhp.ac.ma

## Abstract

Early prediction of academic outcomes is vital to enabling timely intervention, supporting at-risk students, and improving educational planning and institutional performance. However, this task becomes particularly challenging when data availability is limited, such as in small or graduate-level programs. This study explores the potential of data augmentation techniques, specifically the Synthetic Minority Oversampling Technique, to enhance the performance of machine learning models applied to such constrained educational datasets. We conduct a comparative analysis using four datasets derived from prior research, each representing a distinct educational use case: one focused on predicting academic success in graduate programs, another on student dropout in virtual learning environments, a third on dissertation performance prediction, and a fourth addressing multi-class performance prediction in undergraduate coding courses. By applying consistent machine learning methods in the original and augmented datasets, we systematically evaluate the impact of data augmentation on classification performance using accuracy, precision, recall, and the F1 score. The results demonstrate marked improvements, with accuracy increases up to 21% and precision gains exceeding 25% in some models, notably with KNN and MLP. While not all algorithms benefit equally, our findings highlight data augmentation as a practical and impactful strategy for improving early prediction capabilities in Educational Data Mining (EDM). By leveraging multiple datasets and diverse educational contexts, this contribution provides robust evidence supporting the broader goal of enhancing decision-making and personalized support in digital learning environments.

**Keywords:** Educational Data Mining; E-learning; Learning Management System; machine learning; data augmentation



Academic Editor: Antonio Sarasa Cabezuelo

Received: 3 June 2026

Revised: 23 June 2026

Accepted: 24 June 2026

Published: 1 July 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

## 1. Introduction

At the university level, institutions are constantly striving to accurately predict student performance in a way that ensures they maximize learning efficacy and project potential academic challenges beforehand. Predictive capabilities are vital to ensuring proper allocation of resources and optimizing overall scholastic achievement rates. After the emergence of education data mining techniques, it is now possible to extract meaningful insights from large datasets in education for facilitating personalized intervention and improving education support systems.

Educational Data Mining (EDM) refers to the application of data mining techniques in education. EDM provides institutions with the ability to examine vast amounts of

educational data and recognize patterns that are not readily obvious, predict student performance, and support decision-making [1]. EDM techniques have become more important in the age of large-scale deployment of Learning Management Systems (LMSs), systems designed to manage and perform various forms of educational activities like content supply, assignment uploading, grading, and student–teacher communication [2].

The rapid integration of digital platforms such as Learning Management Systems (LMS), Virtual Learning Environments (VLEs), and Massive Open Online Courses (MOOCs) has revolutionized the delivery and management of education. These technologies, ranging from globally recognized platforms like Moodle to institution-specific VLEs, facilitate content distribution, online assessments, grading, communication, and student engagement tracking. E-learning, enabled through these systems, has become a cornerstone of modern education by promoting accessibility, flexibility, and personalized learning pathways. The widespread adoption of MOOCs has further expanded access to educational resources, allowing institutions to reach large, diverse learner populations. Collectively, these platforms generate rich data streams that offer immense potential for understanding student behaviors and predicting academic outcomes [3]. However, despite this abundance, many educational settings (particularly specialized graduate programs and smaller institutions) still face the challenge of limited data availability. Such small datasets hinder the effectiveness of traditional predictive analytics, which typically rely on high-volume data to ensure model stability and generalizability [4].

In response to the challenges posed by small educational datasets, particularly in graduate-level or specialized programs, data augmentation has emerged as a promising solution within the machine learning landscape. Data augmentation involves the generation of synthetic data points based on the distribution and structure of existing data, thereby expanding the dataset without the need for additional real-world collection [5]. This technique is especially valuable in Educational Data Mining (EDM), where limited records can restrict model performance and generalizability. By enriching sparse datasets, data augmentation helps mitigate overfitting, and improves the robustness of predictive models. When applied effectively, it enables institutions to leverage machine learning even in data-constrained environments, enhancing their ability to forecast student performance, detect dropout risks, and support personalized interventions. In this context, augmentation becomes a key enabler for institutions aiming to derive actionable insights from limited educational data [6].

To address the growing need for effective analysis and prediction in education, machine learning (ML) has become a central tool within the field of EDM. ML algorithms are capable of uncovering complex, non-linear patterns in educational data that may not be apparent through traditional analytical methods [7]. Their adaptability allows them to make predictions, classify student outcomes, and provide real-time insights based on a wide range of features from academic performance to engagement behaviors. Even in cases where data is limited or partially incomplete, ML models can still deliver valuable results through techniques that enhance data representation and model robustness. As a result, ML plays a transformative role in modern education by enabling data-driven decisions, personalized interventions, and early identification of at-risk students [8].

In this paper, we utilize existing datasets from previous studies and compare their original predictive metrics with those obtained after applying data augmentation techniques. This comparative analysis employs identical methods and metrics, enabling a precise and credible evaluation of the impact and effectiveness of data augmentation in terms of improving prediction accuracy.

The key contributions of this research are as follows:

- Evaluation of the effectiveness of data augmentation on small and different educational datasets.
- Comparative analysis of performance metrics before and after data augmentation.
- Comparative analysis of data augmentation strategies.
- Assessment of various ML algorithms to determine the most beneficial augmentation approaches.

The organization of this paper is as follows: Section 2 presents a literature review of recent papers, Section 3 details the methodology and materials used in the study, including descriptions of datasets, implementation of ML models, and performance metrics. Section 4 presents and discusses the experimental results, highlighting the performance gains achieved through data augmentation. Finally, the Section 5 concludes the paper, summarizing key findings and providing directions for future research.

## 2. Related Work

The prediction of students academic outcomes has become a critical area within Educational Data Mining (EDM), attracting considerable interest from researchers and educational institutions. Accurate predictions help educators implement early interventions, improve academic outcomes, and enhance institutional effectiveness. A significant amount of research has focused on various aspects of student performance prediction, ranging from dropout prediction to academic success in diverse educational contexts.

Various studies have tried using ML algorithms with small dataset sizes, which is a common issue for learning in most institutions. Zohair et al. [2] addressed the modeling issue arising from small dataset sizes in disciplines demanding expertise at graduate levels. Their study underscored the possibilities of clustering algorithms in searching for significant predictive factors, which they later demonstrated with Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA) algorithms with significant accuracy even in the presence of limitations on dataset size.

Similarly, Kabathova et al. [4] also addressed the prediction of dropouts from less educational data received from e-learning courses. They emphasized the significance of an appropriate understanding of data and feature selection. Though conducted with less data, their research validated the use of dropout rate prediction with accuracies of 77% to 93% through different ML classifiers.

Cordovilla et al. [9] subsequently extended the predictive modeling paradigm through the merging of external data and LMS data to predict educational attainment for online coding classes. The authors' findings were that simpler models, particularly Logistic Regression, performed extremely effectively and were understandable, indicating the promise of viable implementation in learning systems.

Realinho et al. [10] also explored predicting student dropouts and academic performance by clustering demographic, socioeconomic, and educational data. Their comprehensive approach provided valuable insights to inform educational intervention, indicating the broad applicability and relevance of predictive analytics in higher education.

In addition, Riestra-Gonzalez et al. [11] analyzed massive LMS log data to predict student performance in the initial phases of academic courses. Their study validated the utility of predictive models for course-independent contexts and emphasized the role of early intervention in academic achievement.

Kaensar et al. [12] also utilized vast quantities of Moodle log data during the COVID-19 period to demonstrate that ML models, i.e., Decision Trees and Support Vector Machine, were sufficiently capable of predicting student performance and enabling teachers to better manage online learning environments.

Hung et al. [13] investigated patterns of learning by Educational Data Mining (EDM) in a mixed-mode learning setup, illustrating that clustering and prediction models effectively diagnosed high-risk learners. Their work justified the implementation of predictive analytics in teaching due to timely and focused interventions.

Ciolacu et al. [14] advanced the concept of Education 4.0 through AI-assisted higher education, leveraging smart sensors, wearable devices, and adaptive learning environments. Their innovative approach demonstrated the effectiveness of early recognition systems based on machine learning, highlighting the potential for personalized education experiences.

Furthermore, Villegas-Ch et al. [15] proposed integrating AI and data analysis techniques within LMS environments to improve learning outcomes. Their findings supported the argument for adopting robust, technology-driven educational models, particularly relevant in post-pandemic educational scenarios.

Azimi et al. [16] specifically targeted the challenge posed by small-scale online courses, demonstrating the feasibility and efficacy of ML-based intervention strategies for enhancing student performance. Their results highlighted the viability of applying ML methods even when limited data are available, contributing to early and effective interventions.

Naicker et al. [17] compared Linear Support Vector Machine against classical ML algorithms, confirming Linear Support Vector Machine's suitability in predicting academic performance, while Mueen et al. [18] highlighted Naïve Bayes' robustness in academic performance predictions, outperforming other ML models in undergraduate courses.

Hussain et al. [19] further emphasized ML's capabilities in predicting secondary and intermediate-level academic performances, demonstrating clear potential for ML to significantly influence educational policy and intervention strategies at various educational levels.

Lastly, Bertolini et al. [20] proposed integrating feature selection with cross-validation to enhance predictive models' stability and interpretability, providing actionable insights for educators and stakeholders. Their research underscored the importance of systematically identifying influential predictive features.

Table 1 summarizes representative studies and outlines the reference, core focus, methodology, key findings, and performance metrics for each analyzed study.

**Table 1.** Summary of related works.

| Reference                    | Focus                                           | Methodology                         | Key Findings                                             | Performance Metrics             |
|------------------------------|-------------------------------------------------|-------------------------------------|----------------------------------------------------------|---------------------------------|
| Zohair et al. [2]            | Modeling small datasets for student performance | ML classifiers comparison           | SVM and LDA performed best                               | Accuracy                        |
| Kabathova et al. [4]         | Dropout prediction in e-learning                | ML classifiers comparison           | Achieved 77–93% accuracy despite limited features        | Accuracy, Precision, Recall, F1 |
| Cordovilla et al. [9]        | Success prediction using LMS and external data  | Logistic Regression, SVM, RF, MLP   | Logistic Regression achieved best AUC-ROC (0.9354)       | AUC-ROC                         |
| Riestra-González et al. [11] | Early performance prediction from LMS logs      | Decision Trees, SVM, MLP            | MLP showed best accuracy growing with course progression | Accuracy                        |
| Kaensar et al. [12]          | Prediction using Moodle logs during pandemic    | Decision Trees, SVM, LR             | SVM highest early-stage prediction, DT better overall    | Accuracy                        |
| Hung et al. [13]             | Learning behavior patterns in blended learning  | Random Forest, Clustering, Heatmaps | Early identification of high-risk learners               | F1 Score                        |

Table 1. *Cont.*

| Reference               | Focus                                                  | Methodology                                    | Key Findings                                                   | Performance Metrics    |
|-------------------------|--------------------------------------------------------|------------------------------------------------|----------------------------------------------------------------|------------------------|
| Ciolacu et al. [14]     | Early recognition with AI-assisted learning systems    | Smart sensors, ML algorithms                   | Early prediction models to support adaptive education          | Early Recognition Rate |
| Villegas-Ch et al. [15] | LMS improvement with ML and data analysis              | Machine Learning models                        | Robust LMS development to enhance learning                     | Accuracy               |
| Azimi et al. [16]       | Early intervention in small-scale online courses       | ML-based predictive modeling                   | Effective early intervention strategies recommended            | Accuracy               |
| Naicker et al. [17]     | Performance prediction in school-based education       | Linear SVM, Classical ML algorithms            | Linear SVM outperformed traditional classifiers                | Accuracy               |
| Mueen et al. [18]       | Academic performance prediction (undergraduate)        | Naïve Bayes, Neural Networks, Decision Trees   | Naïve Bayes achieved highest prediction accuracy (86%)         | Accuracy               |
| Hussain et al. [19]     | Academic performance at secondary/intermediate level   | Regression, Decision Tree classifiers          | Efficient prediction of marks and grades using ML              | Accuracy               |
| Bertolini et al. [20]   | Improving feature stability for student success models | Feature Selection, Logistic Regression, GLMNET | Identified stable features for enhanced model interpretability | AUC-ROC                |

In Table 1, while these diverse studies demonstrate various ML methods' effectiveness in educational performance prediction, our research contributes uniquely by directly comparing results from existing datasets with outcomes obtained through data augmentation. Our work explicitly focuses on assessing data augmentation's effectiveness in addressing the limitations posed by small educational datasets, thereby providing a clear evaluation of its utility in educational predictive analytics.

### 3. Methodology and Materials

This research focuses on the influence of data augmentation on forecasting performance using ML models that are specific to small datasets. As important as predicting students' performance is to making early intervention and improvement of learning outcomes possible, it is a difficult and challenging endeavor. One of the challenges is that the quantity of data is inadequate, particularly in small or newly discovered colleges where fewer students are enrolled. Additionally, student data privacy and confidentiality issues impose additional constraints on the sharing and availability of data between institutions. Another problem is the inability to develop generalized prediction models owing to heterogeneity between datasets caused by student demographics, education programs, LMS, and country-level education policies. This heterogeneity complicates the development of universal performance measures or deployment of models that are sufficiently efficient in diverse environments. These challenges underscore the necessity of researching countermeasures like data augmentation to mitigate the problem of data sparsity and build robust and precise predictive models in the educational environment.

Additionally, it should be mentioned that datasets used in research on EDM differ greatly based on their origin and purpose. For example, datasets originating from systems such as Moodle, LMS, Virtual Learning Environments (VLEs), or general e-learning systems capture distinct features of student interaction and activity. This diversity impacts the quality and type of features available, e.g., login frequency, learning material watching

duration, forum participation, assignment submission, or marks. In addition, the uses of these datasets are varied; some are used to predict students' performance, while others are used to predict dropout risk. Therefore, the heterogeneity of data sources and the heterogeneity of study goals also further complicate generalization and standardization of prediction models across educational settings.

### 3.1. Overview and Description of the Selected Studies and Datasets

In our study, to verify and validate our results, we utilized four small datasets from three distinct educational institutions, each of which has been previously used and published in research articles. The first dataset focuses on forecasting academic performance among master's degree students (Dataset 1) and the dissertation grade (Dataset 2) [2], the third dataset centers on predicting student dropout likelihood in a virtual learning environment (VLE) (Dataset 3) [4], and the last dataset involves forecasting student performance in introductory coding courses at a Moroccan university (Dataset 4) [6]. By selecting datasets of different educational emphases and applying consistent methodologies, we systematically investigate the effectiveness of data augmentation techniques in enhancing predictive modeling under scenarios of limited data availability. The four selected datasets are described below:

- Dataset 1 and Dataset 2 [2]: These datasets were constructed to investigate the possibility of accurately predicting the performance of graduate students using small datasets. Dataset 1 consists of the academic performance records of 50 master's students, while Dataset 2 focuses on dissertation grades within the same cohort. Both datasets include features covering demographics, academic history, and course-level scores. Several ML algorithms were compared, with Support Vector Machine and Linear Discriminant Analysis identified as particularly effective for predicting student success, though the limited data volume raised challenges related to generalizability.
- Dataset 3 [4]: This dataset aimed to predict student dropout using the limited educational data of 321 records collected from university lectures. The predictive model developed sought to identify at-risk students, even within a small dataset context. The dataset contains four years of schooling data, including student activity features, grades, and activity logs from a VLE. Multiple machine learning classifiers were applied, achieving prediction accuracies ranging from 77% to 93% on unseen cohorts from subsequent academic years.
- Dataset 4 [6]: This recent dataset focuses on predicting student performance in coding courses at Hassan First University of Settat, Morocco. It includes 200 records from second-year students enrolled in foundational coding modules. What distinguishes this dataset is its hybrid approach, combining academic history (first-year university grades) with behavioral data collected by instructors during practical coding sessions. The predictive tasks target three modules: Algorithm and Programming 1, Algorithm and Programming 2, and Data Structures. Performance is assessed using multi-class classification, with key metrics including accuracy, precision, recall, and F1 score. Notably, the study also explores advanced data augmentation techniques (SMOTE) to address class imbalance and data scarcity, providing new insights for the application of machine learning in early coding education.

These datasets with their records being few in number are a common and hard issue in EDM. Their size limitations tend to constrain the capacity of machine learning models to generalize, and hence the stability of the models, as well as introducing bias and reliability problems. Yet the same restraint presents a practical advantage; it presents a perfectly suited environment in which to assess the efficiency of data augmentation schemes. Through the application of augmentation techniques to these small datasets and rigorously comparing

the outcomes with baseline performance measures, one can seriously test if augmentation really produces significantly improved prediction outcomes. One thereby converts the problem of small datasets into a required experimental basis for testing the value of data augmentation techniques in education predictive analytics.

In this study, the term small educational dataset does not refer only to the absolute number of records, but also to the relationship between the number of samples, the number of features, the number of target classes, and the degree of class imbalance. Thus, even datasets with 200 or 321 records may be considered small in EDM when they are divided into training and testing subsets and when some classes contain only a limited number of representative samples [6]. This limitation can reduce model stability, increase sensitivity to data splitting, and affect the generalization capacity of machine learning models [4].

### 3.2. Data Augmentation

Data augmentation is a set of techniques that is used to artificially enlarge the size of a dataset by generating new synthetic instances from existing data. Rather than gathering new real-world data, which may be costly, time-consuming, or even impossible in some cases, augmentation generates additional instances that maintain the underlying pattern and structure of the original data. This is carried out through such methods as random sampling, perturbations, transformations, the Synthetic Minority Oversampling Technique (SMOTE), or even other algorithmic methods. Data augmentation is particularly crucial when working on small datasets since the small number of instances present can hugely limit the ability of a machine learning model to generalize to new instances. Issues such as overfitting arise in small datasets; the model learns noise and not underlying trends, leading to poor predictive capacity. Augmentation increases the diversity and representativeness of samples [6]. At the same time, this technique introduces variability that can possibly enhance the model's robustness, accuracy, and stability concerning variations in the space of training data. In addition, augmentation helps in simulating various situations that make the model more robust even in real scenarios where small perturbations or outliers may prevail. Therefore, in applications such as EDM, where the datasets that can be made available are small in size due to privacy issues, demographic differences, or institutional limitations, data augmentation offers a viable option to increase model performance and deliver more accurate and reliable analysis outcomes [21].

Important aspects of data augmentation are as follows:

- **Synthetic Generation:** Creating new data points using techniques such as oversampling, SMOTE, noise addition, and transformation.
- **Introduction of Variability:** Increasing diversity in data so that machine learning models can generalize better and avoid overfitting.
- **Bias Reduction:** Avoiding biases and keeping models away from over-reliance on a limited quantity of data.
- **Addresses Class Imbalance:** Techniques like SMOTE allow for the creation of additional samples for underrepresented classes, leading to more balanced training.
- **Cost-Effective:** Compared to acquiring new labeled data, augmentation is computationally inexpensive and faster.
- **Facilitates Better Early Predictions:** Especially in education, where early intervention is critical, augmentation can simulate enough data to train models even in early course stages.

There are also some disadvantages, such as the following:

- **Potential Overhead:** Although generally low-cost, large-scale augmentation may increase training time, requiring careful management of computational resources.

- **Complexity in Method Selection:** Choosing appropriate augmentation techniques depends heavily on the data type (numerical, categorical, sequential) and task, requiring expertise.

In this study, we work with two data augmentation techniques: the Synthetic Minority Oversampling Technique (SMOTE) and Generative Adversarial Networks (GANs). Both methods have demonstrated potential for improving model performance in small and imbalanced datasets. Our comparative framework applies both approaches to EDM, allowing us to systematically evaluate and select the most effective augmentation technique for each scenario.

SMOTE offers a significant advantage by generating new synthetic examples through interpolation between existing minority class instances, rather than simply duplicating existing data. This approach enhances the diversity of the dataset while maintaining the underlying structure and distribution of the original data, thus mitigating the risk of overfitting. Furthermore, SMOTE directly addresses class imbalance issues, a common characteristic of educational datasets where successful and at-risk students are often unevenly represented. By enriching the minority class without introducing random noise, SMOTE improves model generalization and robustness, responding effectively to the challenge of limited real-world data availability. Additionally, SMOTE minimizes the potential disadvantage of creating unrealistic data points, as synthetic instances are generated along the feature space between close neighbors, preserving realistic patterns. Despite the general concern that augmentation could introduce bias, the controlled synthetic generation of SMOTE ensures that variability is introduced systematically and meaningfully. Consequently, SMOTE aligns well with the goals of this research, offering a practical, interpretable, and effective augmentation strategy that reinforces the credibility and validity of the comparative analysis between the original and augmented datasets [22].

In parallel, Generative Adversarial Networks (GANs) are leveraged for their capacity to model complex, high-dimensional data distributions and generate highly realistic synthetic samples. Unlike SMOTE, which relies on interpolation, GANs use a game-theoretic approach between a generator and discriminator neural network, learning to create data points that are indistinguishable from real data. This makes GANs particularly attractive for generating synthetic educational records that mimic intricate relationships in real student data. In our comparative experiments, we apply SMOTE and GAN-based augmentation, compare their impact on model performance, and discuss their suitability and limitations for educational predictive modeling [23].

For GAN-based augmentation, a simple fully connected vanilla GAN architecture was implemented for tabular educational data. Before training, the features of the training set were normalized to the range  $[-1, 1]$  using MinMax scaling, while the target label was concatenated with the scaled features to form the input data distribution. The generator receives a random noise vector of dimension 10 and maps it through one hidden fully connected layer with 64 neurons and ReLU activation, followed by a linear output layer whose dimension corresponds to the full data vector, including both features and the target label. The discriminator receives a real or generated data vector as input and passes it through one hidden fully connected layer with 64 neurons and ReLU activation, followed by a single sigmoid output neuron that predicts whether the input sample is real or synthetic. Both networks were trained adversarially using binary cross-entropy loss and the Adam optimizer with a learning rate of 0.0002. The model was trained for 100 epochs using a batch size of 64. After training, 250 synthetic samples were generated. The generated feature values were transformed back to the original scale using the inverse MinMax transformation, while the generated target labels were converted into binary class labels using a thresholding rule.

To prevent data leakage and ensure a fair evaluation of the proposed augmentation strategy, the augmentation process was applied only after the original dataset had been divided into training and testing subsets. For each dataset, the train/test split was performed following the same strategy reported in the original study from which the dataset was derived, in order to preserve comparable experimental conditions. In other words, the original non-augmented dataset was first split into training and testing sets according to the protocol of the corresponding baseline study. Then, SMOTE and GAN-based augmentation were applied exclusively to the training set. The test set was kept unchanged and was not used during synthetic data generation, model fitting, or hyperparameter adjustment. This design ensures that the models were evaluated on real, unseen, and non-augmented data, avoiding any artificial improvement caused by increasing the test set size or allowing information from test samples to influence the augmented training data. Consequently, the reported performance gains reflect the effect of enriching the training data while maintaining an independent evaluation protocol and preserving the experimental conditions of the original studies.

### 3.3. Machine Learning Methods

ML algorithms are increasingly significant applications in EDM for modeling intricate and non-linear patterns common in student learning behavior and academic achievement [1]. Classical statistical methods are certainly helpful for generally locating trends in data but only make quite strong assumptions about normality, linearity of relationships, and independence between the variables concerned that are seldom seen in actual educational environments. Therefore, conventional methods may fail to capture the subtle interactions and diverse learning trajectories among students, especially when dealing with heterogeneous datasets that are riddled with missing values, class imbalance, and demographic diversity. In contrast, more recent ML algorithms such as Support Vector Machines (SVMs), Linear Discriminant Analysis (LDA), Artificial Neural Networks (ANNs), and Long-Short-Term-Memory (LSTM) are more adaptive and flexible. They can automatically detect complex feature interactions, process high-dimensional data, and make more accurate and robust predictions even under challenging circumstances. In applications like student performance prediction or dropout detection, where early and accurate interventions are crucial, the higher predictive power of ML models makes them an indispensable tool. By utilizing their strengths, schools can move beyond surface-level analysis, develop personalized support systems, and ultimately promote better academic success for groups of students that are diverse [24].

The studies referenced below employed the following machine learning algorithms to address the classification and prediction issues posed by small datasets. The key algorithms include:

- K-Nearest Neighbors (KNN): A non-parametric classifier that makes a label prediction for a point based on the most frequent class among its neighbors. It is easy to use and can be used when datasets are small, but might be noisy and k-dependent.
- Logistic Regression (LR): A widely used linear classification model, which is especially suitable for binary outcomes. It estimates the probability of a class label through a logistic function and is valued for its interpretability and efficiency in linearly separable problems.
- Support Vector Machines (SVMs): A robust classifier with very good reputation for good performance in high-dimensional space, whose application with small datasets is possible due to the fact that the algorithm can maximize the margin between data instances.

- Linear Discriminant Analysis (LDA): Most widely used in classification and dimensionality reduction; it tries to maximize differences between different classes and reduce variance within a class.
- Decision Trees (DTs): A non-parametric supervised learning method for classification, providing easily interpretable models.
- Random Forest (RF): An ensemble learning method that merges a sequence of DTs, enhancing predictive precision through majority voting mechanisms.
- Naïve Bayes (NB): A probabilistic classifier based on Bayes' theorem with high suitability to datasets with small samples due to its simplicity and computational efficiency.
- Adaptive Boosting (AB): An ensemble technique that combines multiple weak learners, typically DTs, by iteratively focusing on incorrectly classified instances. It improves accuracy by weighting models and examples adaptively during training.
- Gradient Boosting (GB): An ensemble learning algorithm that builds models sequentially, where each new weak learner, usually a decision tree, is trained to correct the errors made by the previous models. It combines these learners to improve predictive accuracy and reduce overall prediction error.
- Artificial Neural Networks (ANNs): With the capability to model intricate non-linear relationships, this method is sensitive to the quantity and diversity of training data, and data augmentation is highly beneficial.
- Long Short-Term Memory (LSTM): A specialized type of recurrent neural network designed to capture long-term dependencies and temporal patterns in sequential data. LSTM is especially effective for modeling educational sequences and learning trajectories over time.

The field of EDM has witnessed a substantial growth in research activities, with numerous studies applying ML algorithms to analyze and predict various educational outcomes. However, a notable challenge within EDM is the heterogeneity of approaches; each study often works with a different dataset, specific to a particular institution, program, or learning platform, and targets diverse objectives such as predicting academic success, detecting at-risk students, or understanding learning behaviors [25]. As a result, the features extracted, the preprocessing techniques applied, and the evaluation metrics vary widely across studies. Moreover, due to differences in data characteristics and study goals, different machine learning algorithms are found to perform better in different contexts, with no universal consensus on a single best method. Some studies find SVM to be the most effective, while others highlight the strength of DTs, RF, or ANNs, depending on the nature of the data. This variability presents a significant challenge for developing generalized, transferable predictive models within EDM [26]. In our study, we aim to address this complexity by systematically investigating the impact of data augmentation on predictive performance using diverse small educational datasets from different universities, domains, and purposes of use. By applying consistent methodologies and critically comparing the results before and after augmentation, we contribute to a better understanding of how to enhance predictive modeling reliability despite the inherent variability across educational data sources.

To ensure a fair and reliable comparison with the baseline results reported in the original studies, the experimental conditions of the original works were reproduced as closely as possible for each dataset. In particular, the same preprocessing procedures, train/test splitting strategy, feature selection process, and hyperparameter settings reported in the original studies were adopted when re-implementing the corresponding machine learning models. The results obtained after applying data augmentation were then generated using the same models and under the same experimental conditions, so that the main difference between the baseline and augmented experiments was the use of synthetic training

samples. In addition, to reinforce the comparative analysis and provide a more consistent evaluation across all datasets, we applied the same set of machine learning algorithms to each dataset. Consequently, some classifiers included in our experiments were not present in the corresponding original study. These additional models were introduced to strengthen the evaluation framework and to better assess whether the effectiveness of data augmentation depends on the classifier type. For these added models, the hyperparameters and implementation settings are explicitly described to ensure reproducibility and transparency.

### 3.4. Performance Metrics

In this research, several performance metrics are employed to evaluate the effectiveness of the machine learning models.

**Accuracy:** Accuracy represents the ratio of correctly predicted instances to the total number of instances evaluated.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

**Precision:** Precision measures the proportion of true-positive predictions among all instances classified as positive.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

**Recall:** Recall, also known as sensitivity, indicates the proportion of actual positives correctly identified by the model.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

**F1 Score:** The F1 score provides the harmonic mean of precision and recall, offering a balanced assessment of model performance.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

where:

- TP (True Positive): Correctly predicted positive instances.
- TN (True Negative): Correctly predicted negative instances.
- FP (False Positive): Incorrectly predicted positive instances.
- FN (False Negative): Incorrectly predicted negative instances.

Performance metrics in the EDM scenario, such as accuracy, precision, recall, and F1 score, are crucial to assess the efficiency of predictive models and interpret the implications of their outputs. All of these metrics inform us about the potential performance of a model, particularly in learning settings where predicting students at risk early is highly critical [6]. Accuracy gives a general idea of model accuracy overall, but is deceptive in imbalanced data that is common in education. Precision, by focusing on the proportion of true-positive predictions to all predicted positives, helps minimize false alarms and ensures interventions are directed towards students actually at risk. Recall, defined as the ability to classify all actual positive instances accurately, is particularly precious to ensuring that no underperforming student will be left behind. F1 score, being a harmonic mean of precision and recall, is an equilibrated measurement when trade-offs among false positives and false negatives have to be managed well. In educational settings, where punctual and accurate interventions can become the determining element in student performance, relying solely

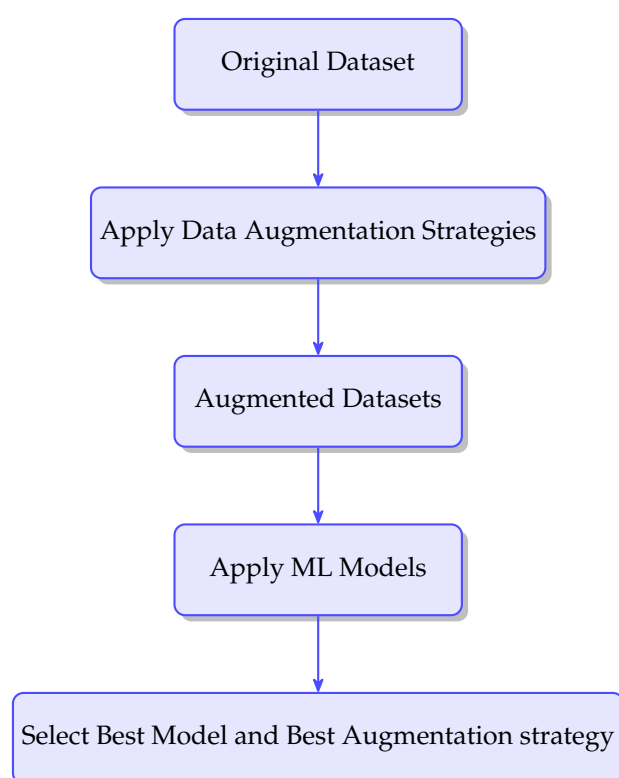
on one measure is insufficient. Combined comprehension through multiple measures of performance ensures that intervention strategies are effective as well as fair, eventually contributing to enhanced learning results and institutional goals [9].

### 3.5. Research Framework and Workflow

In this section, the full discussion of experimental results attained using ML models on the original and augmented datasets created using SMOTE and GANs is provided. It intends to compare how data augmentation—i.e., by means of ascertaining the best between SMOTE and GANs—enhances model performance while working with small educational datasets. To achieve a valid and fair comparison, we use the same performance metrics as were employed in the original works, i.e., precision, recall, accuracy, and F1 score. By comparing pre- and post-augmentation performance, we demonstrate how the data augmentation strategies help to improve predictive performance and allow for early educational intervention.

1. Original datasets obtained from Kabathova et al. [4], Zohair et al. [2], and Hafdi et al. [6].
2. Application of data augmentation methods to generate enhanced datasets.
3. Implementation of the machine learning models as used in the original studies on augmented datasets.
4. Comparative analysis of model performance before and after data augmentation.
5. Selection of the best-performing model based on relevant performance metrics (accuracy, precision, recall, and F1 score), and selection of the best augmentation method.

Figure 1 illustrates the complete process.



**Figure 1.** Study Framework and Workflow.

This structured approach ensures a rigorous comparative analysis to conclusively determine the efficacy of data augmentation techniques in enhancing predictive performance for small dataset.

## 4. Results and Discussion

This section presents a holistic discussion of the experimental results that have been attained through applying ML models on the original and on the augmented datasets. The purpose is to compare the performance of data augmentation strategies, specifically SMOTE and GANs, in enhancing model performance when dealing with small educational datasets. For a fair and equal comparison, we use the same performance measures as were utilized in the original works, i.e., accuracy, precision, recall, and F1 score. By comparing the results before and after augmentation, we aim to demonstrate how synthetic data generation contributes to improving predictive performance and facilitating early educational interventions.

### 4.1. Prediction of Students' Academic Results

In the first referenced study by Zohair et al. [2], the authors investigated student performance prediction using small datasets and evaluated several ML algorithms, including KNN, SVM, NBs, LDA, and ANN. Their analysis was conducted on two different datasets. The first dataset, which we refer to in our study as Dataset 1, focused on predicting final grades for various course modules. The second, referred to in our work as Dataset 2, was used to predict the dissertation grade in a master's program. In terms of performance, the authors reported that SVM achieved the highest accuracy scores, 69.7% and 76.3%, respectively. These results serve as the benchmark for evaluating the impact of data augmentation in our comparative analysis. To strengthen the generality of the comparative analysis, we also included three additional machine learning methods that were not evaluated in the original study, namely DT, LR, and AD. The experimental conditions of the original study were preserved as much as possible, and these additional models were applied under the same preprocessing and train/test split protocol. For the added classifiers, DT was configured with `max_depth = 2` for both datasets, the LR model was used in its default configuration, and AD was trained with five estimators for Dataset 1 and 11 estimators for Dataset 2. All three added models were implemented using a fixed `random_state = 101`. It should also be noted that the original study used accuracy as the main evaluation metric for selecting the best-performing model.

Figure 2 and Tables 2 and 3 illustrate a comparative analysis of model performance across different data augmentation strategies namely SMOTE and GANs alongside the baseline performance without augmentation, for Dataset 1 and Dataset 2. This comprehensive evaluation sheds light on how data generation influences classification accuracy in EDM contexts, particularly when dealing with imbalanced or limited datasets.

**Table 2.** Accuracy comparison across different augmentation techniques and ML algorithms (Dataset 1).

| Augmentation Type  | KNN       | SVM       | NB        | LDA       | ANN       | DT    | LR    | AB  |
|--------------------|-----------|-----------|-----------|-----------|-----------|-------|-------|-----|
| SMOTE Augmentation | 79.2%     | 63.9%     | 47.2%     | 42.4%     | 77%       | 66.7% | 61.6% | 62% |
| GAN Augmentation   | 61.8%     | 57.6%     | 34.7%     | 57%       | 56%       | 66.6% | 53.5% | 64% |
| No Augmentation    | 56.4% [2] | 69.7% [2] | 58.1% [2] | 63.2% [2] | 58.1% [2] | 65.5% | 52%   | 52% |

For Dataset 1, the original (non-augmented) results reported by Zohair et al. [2] indicate that the SVM model achieved the highest accuracy at 69.7%, followed by LDA and NB. However, once SMOTE augmentation was applied, KNN and ANN significantly improved to 79.2% and 77%, respectively. DT, LR, and AD, which were not originally tested in the reference study, also performed well with accuracies above 60%. Interestingly, the GAN augmentation yielded mixed results; while some methods like DT and AD maintained competitive accuracies (66.6% and 64%), others such as NB and ANN experienced perfor-

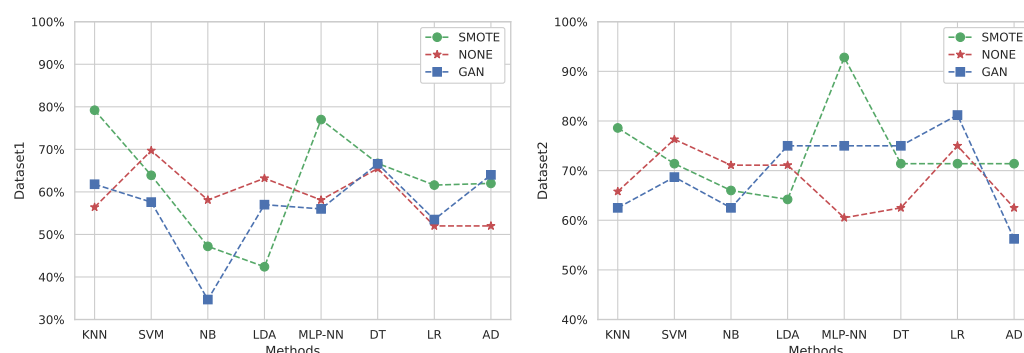
mance drops compared to SMOTE and the baseline, indicating the potential challenges of using GANs for certain classifiers under this dataset's characteristics. Nonetheless, GANs still provided improvement over no augmentation in models like LR (from 52% to 53.5%) and AD (from 52% to 64%).

Turning to Dataset 2, the non-augmented results once again highlight SVM as the leading method with 76.3%, followed closely by LR and LDA at 75% and 71.1%. After applying SMOTE, substantial performance gains were observed, especially for ANN (rising from 60.5% to 92.8%) and KNN (from 65.8% to 78.6%). The DT and AB classifiers also benefited from SMOTE, achieving 71.4% and 71.3%, respectively. In contrast, NB performance decreased significantly with SMOTE. The GAN augmentation approach, on the other hand, provided a more balanced outcome. Although it did not outperform SMOTE in most models, it yielded high accuracies for LR (81.2%), LDA (75%), and DT/ANN (75% each). Notably, NB accuracy rose to 62.5% with the GAN technique performing better than SMOTE (35.7%), demonstrating that GANs may better preserve class structure for certain classifiers.

**Table 3.** Accuracy comparison across different augmentation techniques and ML algorithms (Dataset 2).

| Augmentation Type  | KNN       | SVM       | NB        | LDA       | ANN       | DT    | LR    | AB     |
|--------------------|-----------|-----------|-----------|-----------|-----------|-------|-------|--------|
| SMOTE Augmentation | 78.6%     | 71.4%     | 35.7%     | 64.2%     | 92.8%     | 71.4% | 71.4% | 71.3%  |
| GAN Augmentation   | 62.5%     | 68.7%     | 62.5%     | 75%       | 75%       | 75%   | 81.2% | 56.25% |
| No Augmentation    | 65.8% [2] | 76.3% [2] | 71.1% [2] | 71.1% [2] | 60.5% [2] | 65.5% | 75%   | 62.5%  |

Overall, the visualized comparison in Figure 2 confirms that data augmentation plays a pivotal role in enhancing model performance when dealing with small or imbalanced educational datasets. SMOTE generally produced more consistent and higher accuracy boosts across most classifiers, especially in Dataset 2. However, GAN augmentation proved to be a promising alternative, particularly for models like LR and LDA, and may be better suited to complex feature–target relationships. These findings underline the importance of selecting the appropriate augmentation technique based on the dataset and the classification algorithm used.



**Figure 2.** Accuracy comparison of data augmentation using the first and second datasets.

The experimental results derived from datasets reinforce the value and complexity of data augmentation in EDM. For Dataset 1, SMOTE-based augmentation significantly improved the performance of several models, particularly KNN and ANN, which rose to 79.2% and 77%, respectively, surpassing the original top-performing SVM (69.7%). The newly added models DT, LR, and AB also demonstrated competitive results under SMOTE, confirming the method's efficacy in enhancing learning with imbalanced or limited samples. However, the introduction of GAN-based augmentation led to more varied outcomes; while

DT and AB maintained stable performance (66.6% and 64%), models such as SVM, NB, and LDA suffered declines, suggesting that GAN-generated data may introduce feature distributions not easily captured by all classifiers. These results support previous findings by Zohair et al. [2], who emphasized the role of data quality and volume in determining the reliability of predictions using models like SVM and LDA.

Dataset 2 further illustrated the contrast between augmentation strategies. SMOTE led to the highest accuracy for ANN (85.7%) and KNN (78.6%), strongly outperforming their baseline scores of 60.5% and 65.8%. GAN augmentation, in turn, showed promising gains for linear models like LR and LDA achieving 81.2% and 75% respectively, while causing accuracy drops in AB and SVM. This outcome aligns with prior studies such as Riestra-González et al. [11], which highlighted the strength of neural networks when exposed to rich and balanced data, and findings from Bertolini et al. [20], who stressed that algorithmic performance depends heavily on dataset characteristics and preprocessing steps. The diverging results between SMOTE and GANs underline the importance of choosing augmentation methods tailored to the model type and data context. Overall, while SMOTE delivered more stable improvements across models, GANs offered unique advantages for specific classifiers, further reinforcing the idea that augmentation must be strategically aligned with the target model and task.

#### 4.2. Prediction of Student Dropout Using the Third Dataset

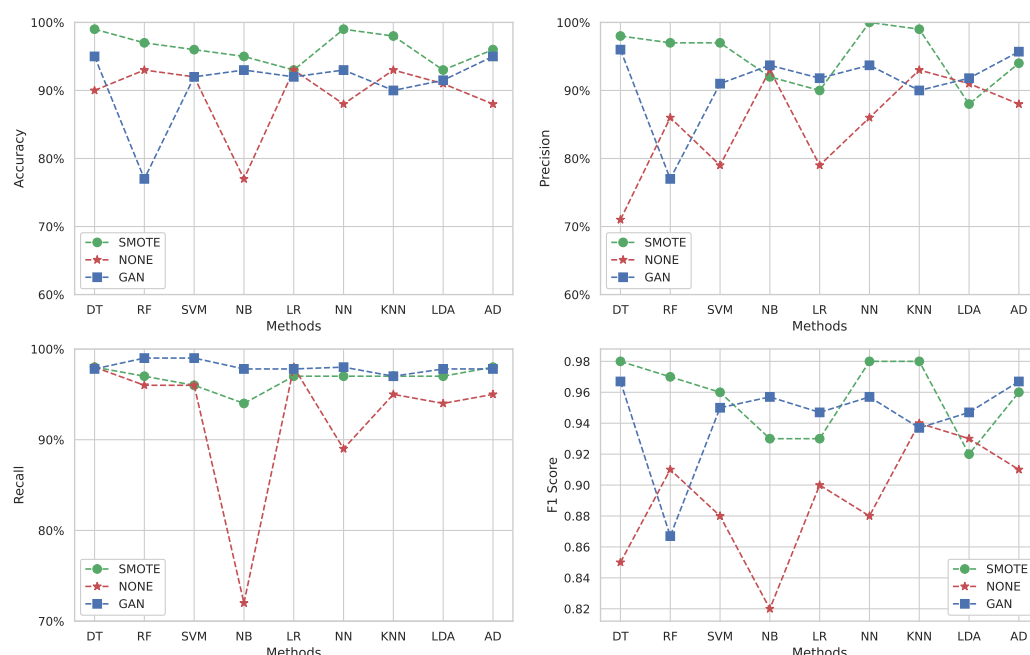
The third dataset, derived from the work of Kabathova et al. [4], centers on the early prediction of student dropout within university level e-learning environments. In their study, Kabathova et al. evaluated a range of machine learning classifiers using a compact set of features readily available to course instructors, achieving promising accuracy rates between 77% and 93% across different models. Their approach placed particular emphasis on core performance metrics such as accuracy, precision, recall, and F1 score measures that are especially pertinent in the context of small and imbalanced educational datasets. These metrics established a robust benchmark for classifier reliability and model consistency, and in this work, we adopt their results as a baseline to assess the impact and benefits of data augmentation methods within our comparative framework. For this dataset, the ML models were implemented under the same experimental conditions as those reported in the original study, particularly regarding preprocessing, train/test splitting, and evaluation metrics. To reinforce the comparative analysis, additional classifiers were also included using fixed and transparent configurations. Specifically, KNN was implemented with  $n\_neighbors = 3$ , LDA was kept with its default configuration, and AB was trained with  $n\_estimators = 1$ . All remaining parameters were kept at their default values to ensure a simple and reproducible comparison between the original and augmented datasets.

The first subfigure in Figure 3 and Table 4 present the accuracy comparison across multiple machine learning models under the same three augmentation conditions: No augmentation, SMOTE, and GAN-based augmentation. In the original study by Kabathova et al. [4], LR and KNN achieved the best performance with 93% accuracy, while DT and ANN reached 90% and 88%, respectively. With the introduction of SMOTE in our pipeline, notable improvements were observed across nearly all models, particularly DT and ANN, both reaching 99%, followed by RF and SVM at 97% and 96%, respectively. NB also improved dramatically from 77% to 95%. Additionally, models not originally evaluated in the study, such as AB, LDA, and KNN, were included to generalize the evaluation. These also showed enhanced accuracy using SMOTE, demonstrating the technique's effectiveness. On the other hand, the use of GAN-based data augmentation revealed different trends. Although GANs also improved most model performances compared to the baseline, the gains were generally less pronounced than with SMOTE. For instance, GANs led to strong

accuracies for DT (95%), NB (93%), and AB (95%), but RF dropped significantly to 77%. Nonetheless, GAN augmentation still proved beneficial in several cases, such as for LDA and LR, both maintaining accuracies above 91%. These results demonstrate that while GANs are a promising alternative, their performance varies more depending on the model and may require finer tuning to match the consistency of SMOTE in accuracy enhancement.

**Table 4.** Accuracy comparison across ML algorithms and augmentation strategies for the third dataset.

| Augmentation Type  | DT     | RF        | SVM     | NB      | LR      | ANN     | KNN | LDA   | AB  |
|--------------------|--------|-----------|---------|---------|---------|---------|-----|-------|-----|
| SMOTE Augmentation | 99%    | 97%       | 96%     | 95%     | 93%     | 99%     | 98% | 93%   | 96% |
| GAN Augmentation   | 95%    | 77%       | 92%     | 93%     | 92%     | 93%     | 90% | 91.5% | 95% |
| No Augmentation    | 90%[4] | 92.5% [4] | 92% [4] | 77% [4] | 93% [4] | 88% [4] | 93% | 91%   | 88% |



**Figure 3.** Comparison for dropout prediction metrics (Dataset 3).

The precision results in Table 5 and the second subfigure in Figure 3 indicate that SMOTE consistently yields the highest scores, particularly for ANN (99.8%), KNN (99%), and DT (98%). This confirms SMOTE's strength in generating clean, distinguishable synthetic samples for imbalanced classes. GAN augmentation, while less effective than SMOTE in precision, still showed competitive results, especially with NB (93.7%), AB (95.7%), and LDA (91.8%). It outperformed the original non-augmented precision in several cases, such as LR (91.8% vs. 79%) and SVM (91% vs. 79%), showing its potential when combined with appropriate models.

**Table 5.** Precision comparison across ML algorithms and augmentation strategies for the third dataset.

| Augmentation Type  | DT     | RF     | SVM    | NB     | LR     | ANN    | KNN | LDA   | AB    |
|--------------------|--------|--------|--------|--------|--------|--------|-----|-------|-------|
| SMOTE Augmentation | 98%    | 97%    | 97%    | 92%    | 91%    | 99.8%  | 99% | 88%   | 94%   |
| GAN Augmentation   | 96%    | 77%    | 91%    | 93.7%  | 91.8%  | 93.7%  | 90% | 91.8% | 95.7% |
| No Augmentation    | 71%[4] | 86%[4] | 79%[4] | 93%[4] | 79%[4] | 86%[4] | 93% | 91%   | 88%   |

In terms of recall presented in Table 6 and the third subfigure in Figure 3, GANs demonstrated exceptional performance across nearly all models, with scores reaching or

exceeding 97.8% for DT, NB, LDA, and AB. Notably, RF and SVM achieved a perfect recall of 99% under GAN augmentation, surpassing even the SMOTE and baseline results. These findings suggest that GAN-augmented models are particularly effective at identifying positive cases in dropout prediction, enhancing model sensitivity.

**Table 6.** Recall comparison across ML algorithms and augmentation strategies for the third dataset.

| Augmentation Type  | DT      | RF     | SVM    | NB     | LR      | ANN     | KNN | LDA   | AB    |
|--------------------|---------|--------|--------|--------|---------|---------|-----|-------|-------|
| SMOTE Augmentation | 98%     | 97%    | 96%    | 94%    | 97%     | 97%     | 97% | 97%   | 98%   |
| GAN Augmentation   | 97.8%   | 99%    | 99%    | 97.8%  | 97.8%   | 98%     | 97% | 97.8% | 97.8% |
| No Augmentation    | 98% [4] | 96%[4] | 96%[4] | 72%[4] | 98% [4] | 89% [4] | 95% | 94%   | 95%   |

The F1 score evaluated in the last subfigure in Figure 3 and in Table 7 further supports these insights. While SMOTE again achieved the highest values across most models, with ANN, DT, and KNN reaching the best score with 0.98, GAN augmentation also resulted in robust and balanced scores. For example, AB and NB recorded an F1 score of 0.967 and 0.957, respectively, closely aligning with SMOTE-enhanced models. These consistent results across recall and F1 metrics indicate that GANs are especially strong at achieving balanced performance in classification tasks, making them a promising data augmentation strategy alongside SMOTE.

**Table 7.** F1 Score comparison across ML algorithms and augmentation strategies for the third dataset.

| Augmentation Type  | DT       | RF       | SVM      | NB       | LR       | ANN      | KNN   | LDA   | AB    |
|--------------------|----------|----------|----------|----------|----------|----------|-------|-------|-------|
| SMOTE Augmentation | 0.98     | 0.97     | 0.96     | 0.93     | 0.94     | 0.98     | 0.98  | 0.92  | 0.96  |
| GAN Augmentation   | 0.967    | 0.867    | 0.95     | 0.957    | 0.947    | 0.957    | 0.937 | 0.947 | 0.967 |
| No Augmentation    | 0.85 [4] | 0.91 [4] | 0.88 [4] | 0.82 [4] | 0.90 [4] | 0.88 [4] | 0.94  | 0.93  | 0.91  |

The comparison between our proposed approaches and the original work by Kabathova et al. [4] demonstrates clear improvements across all major performance metrics, showcasing the effectiveness of data augmentation strategies, particularly SMOTE and GANs in enhancing dropout prediction using ML models. With SMOTE, models such as DT and ANN reached 99% accuracy, significantly surpassing the original values of 90% and 88%, respectively. Even with GAN augmentation, these models retained high performance levels (DT at 95%, ANN at 93%), validating the robustness of synthetic data in supporting generalization from limited samples. Precision gains were especially notable; DT improved from 71% to 98% with SMOTE, and maintained 96% with GANs. Similarly, SVM and RF saw consistent precision boosts, confirming SMOTE's effectiveness in managing class imbalance. For recall, the GAN strategy stood out with superior results across many models, including RF (99%), SVM (99%), and LDA (97.8%), underscoring GANs' strength in capturing all at-risk cases. These patterns were reinforced in the F1 scores, where models like DT and NB attained 0.98 and 0.957, respectively, suggesting reduced false alarms.

These models demonstrated strong compatibility with augmentation techniques. Under SMOTE, KNN achieved outstanding metrics—98% accuracy, 99% precision, 97% recall, and an F1 score of 0.98—placing it among the top-performing models. AB also achieved balanced performance across all metrics with SMOTE (96%, 94%, 98%, and 0.96, respectively) and even more competitive results under the GAN strategy. LDA, while often considered more limited in complex datasets, consistently delivered high recall (97%) and respectable F1 scores (0.92 with SMOTE, 0.947 with GANs), confirming its usefulness when combined with data enrichment. These additions support the versatility and scalability of SMOTE and GANs as augmentation solutions across a wide range of algorithms.

Our findings align with recent work in EDM. Cordovilla et al. [9] emphasized the importance of preprocessing strategies like SMOTE in improving prediction with interpretable models such as LR, which also showed reliable results in our study. Likewise, the study by González et al. [11] validated the strong performance of models like MLP (noted here as ANN) and KNN when trained on structured and augmented LMS logs, echoing our results. Notably, the inclusion of GANs in this study presents a new perspective, as few works in EDM have integrated generative models into augmentation workflows. Although GANs' impact varied across models, their success in boosting recall and producing balanced F1 scores signals a promising direction for future research. Overall, our work pushes forward the agenda of augmentation-aware modeling pipelines for effective and early intervention in educational systems.

#### 4.3. Prediction of Students' Academic Success in Programming Courses—Dataset 4

In this subsection, we present the comparative results obtained on the fourth dataset, sourced from Hafdi et al. [6], which focuses on predicting student outcomes in foundational coding modules at Hassan First University of Settat, Morocco. The dataset includes 200 records of second-year students and is characterized by a hybrid feature set that merges academic performance from the first year with weekly behavioral observations collected during practical sessions. Unlike previous datasets, this one addresses a multi-class classification challenge, where the targets (Algorithm and Programming 1, Algorithm and Programming 2, and Data Structures, respectively noted Target 1, Target 2, and Target 3) are categorized into three distinct risk or achievement levels according to institutional standards. To robustly evaluate the impact of data augmentation under these challenging conditions of limited and imbalanced data, we systematically compare the performance of ML models trained with two prominent augmentation methods SMOTE and GANs, as well as a baseline scenario without augmentation. The results are reported across the key classification metrics of accuracy, precision, recall, and F1 score for each of the three targets, allowing for a comprehensive assessment of how different augmentation strategies can enhance predictive modeling for multi-class educational outcomes. And as mentioned earlier all the MLs methods run in the same conditions as original study (train/test split, hyperparameters, etc.).

The analysis for Target 1 in the fourth dataset investigates the impact of two data augmentation strategies SMOTE and GANs, as well the no augmentation strategy, across a broad spectrum of machine learning algorithms. The following discussion evaluates performance in terms of accuracy, precision, recall, and F1 score, referencing the detailed results provided in Tables 8–11 and the Figure 4.

**Table 8.** Accuracy comparison for Target 1 across ML algorithms and augmentation strategies.

| Augmentation Type  | KNN        | DT         | RF         | SVM        | LR         | AB         | GB         | ANN        | LSTM       | BN         | LDA        |
|--------------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| SMOTE Augmentation | 73%<br>[6] | 76%<br>[6] | 79%<br>[6] | 79%<br>[6] | 83%<br>[6] | 62%<br>[6] | 82%<br>[6] | 79%<br>[6] | 82%<br>[6] | 78%<br>[6] | 75%<br>[6] |
| GAN Augmentation   | 69%        | 68%        | 68%        | 70%        | 64%        | 56%        | 64%        | 73%        | 75%        | 67%        | 67%        |
| No Augmentation    | 68%<br>[6] | 71%<br>[6] | 73%<br>[6] | 71%<br>[6] | 73%<br>[6] | 66%<br>[6] | 62%<br>[6] | 78%<br>[6] | 76%<br>[6] | 71%<br>[6] | 75%<br>[6] |

The accuracy results in Table 8 clearly demonstrate the advantage of SMOTE augmentation, which consistently enhances classification performance across nearly all models. Notably, LR achieves the highest accuracy (83%), followed closely by GB (82%), LSTM (82%), RF (79%), and SVM (79%). In contrast, models trained on the original data (no augmentation) generally achieve lower accuracy, with values most often in the low 70% range. The GAN augmentation approach, while showing moderate success with some

models (e.g., KNN and GB), underperforms in comparison to SMOTE, and in several cases, delivers lower accuracy than the non-augmented baseline (for example, LR drops to 64%, GB to 64%). These results indicate that, for this target, SMOTE is a more robust strategy for improving model accuracy.

For precision presented in Table 9, SMOTE again delivers the highest scores for most algorithms. LR (84%), GB (82%), and LSTM (83%) achieve the top results, reflecting a strong reliability in positive predictions. Interestingly, some methods such as ANN and LSTM perform similarly or better without augmentation, suggesting these algorithms either adapt well to the original data or are less tolerant to synthetic data. GAN augmentation yields mixed effects: LSTM and KNN achieve respectable precision (75%, 69%), while LR (54%) and LDA (46%) see considerable decreases. Thus, while SMOTE consistently enhances precision, GANs' benefit appears more algorithm-dependent.

**Table 9.** Precision comparison for Target 1 across ML algorithms and augmentation strategies.

| Augmentation Type  | KNN        | DT         | RF         | SVM        | LR         | AB         | GB         | ANN        | LSTM       | BN         | LDA        |
|--------------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| SMOTE Augmentation | 73%<br>[6] | 77%<br>[6] | 82%<br>[6] | 80%<br>[6] | 84%<br>[6] | 65%<br>[6] | 82%<br>[6] | 80%<br>[6] | 83%<br>[6] | 79%<br>[6] | 75%<br>[6] |
| GAN Augmentation   | 79%        | 60%        | 69%        | 79%        | 54%        | 78%        | 75%        | 81%        | 82%        | 78%        | 46%        |
| No Augmentation    | 79%<br>[6] | 68%<br>[6] | 82%<br>[6] | 80%<br>[6] | 81%<br>[6] | 59%<br>[6] | 75%<br>[6] | 85%<br>[6] | 83%<br>[6] | 58%<br>[6] | 68%<br>[6] |

Recall analysis using Table 10 underscores SMOTE's strength in identifying positive cases. With SMOTE, LR (84%), GB (83%), and LSTM (83%) attain the highest recall, confirming their ability to identify a greater share of true positives. GAN augmentation produces competitive recall in DT (78%), RF (77%), and BN (70%), but results are considerably weaker for methods such as AB (42%), GB (43%), and LDA (45%). Without augmentation, recall values remain moderate (peaking at LDA: 63% and LSTM: 62%). These findings affirm that SMOTE reliably improves recall, while GANs' effect varies more significantly across algorithms.

**Table 10.** Recall comparison for Target 1 across ML algorithms and augmentation strategies.

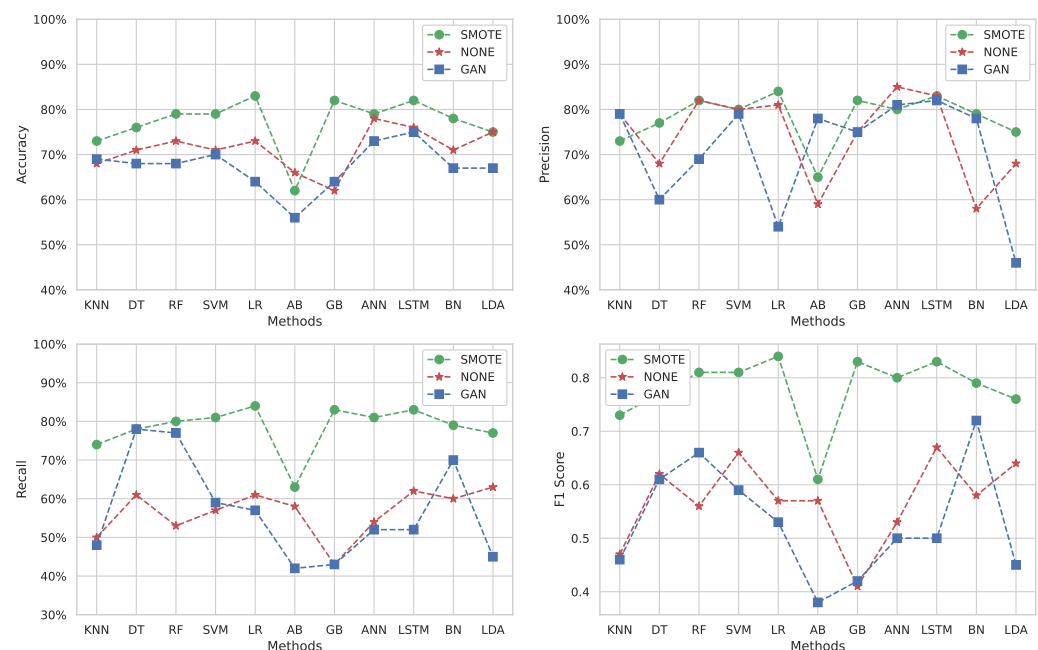
| Augmentation Type  | KNN        | DT         | RF         | SVM        | LR         | AB         | GB         | ANN        | LSTM       | BN         | LDA        |
|--------------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| SMOTE Augmentation | 74%<br>[6] | 78%<br>[6] | 80%<br>[6] | 81%<br>[6] | 84%<br>[6] | 63%<br>[6] | 83%<br>[6] | 81%<br>[6] | 83%<br>[6] | 79%<br>[6] | 77%<br>[6] |
| GAN Augmentation   | 48%        | 78%        | 77%        | 59%        | 57%        | 42%        | 43%        | 52%        | 52%        | 70%        | 45%        |
| No Augmentation    | 50%<br>[6] | 61%<br>[6] | 53%<br>[6] | 57%<br>[6] | 61%<br>[6] | 58%<br>[6] | 43%<br>[6] | 54%<br>[6] | 62%<br>[6] | 60%<br>[6] | 63%<br>[6] |

**Table 11.** F1 Score comparison for Target 1 across ML algorithms and augmentation strategies.

| Augmentation Type  | KNN         | DT          | RF          | SVM         | LR          | AB          | GB          | ANN         | LSTM        | BN          | LDA         |
|--------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| SMOTE Augmentation | 0.73<br>[6] | 0.77<br>[6] | 0.81<br>[6] | 0.81<br>[6] | 0.84<br>[6] | 0.61<br>[6] | 0.83<br>[6] | 0.80<br>[6] | 0.83<br>[6] | 0.79<br>[6] | 0.76<br>[6] |
| GAN Augmentation   | 0.46        | 0.61        | 0.66        | 0.59        | 0.53        | 0.38        | 0.42        | 0.50        | 0.50        | 0.72        | 0.45        |
| No Augmentation    | 0.47<br>[6] | 0.62<br>[6] | 0.56<br>[6] | 0.66<br>[6] | 0.57<br>[6] | 0.57<br>[6] | 0.41<br>[6] | 0.53<br>[6] | 0.67<br>[6] | 0.58<br>[6] | 0.64<br>[6] |

When considering the F1 score detailed in Table 11, the SMOTE technique again yields the most impressive results. The highest F1 scores are seen for LR (0.84), GB (0.83), LSTM (0.83), and ANN (0.80) under SMOTE. GAN augmentation produces moderate F1 scores for BN (0.72) and DT (0.61), but generally lags behind, especially for AB (0.38), GB (0.42), and

LDA (0.45). The non-augmented models rarely surpass an F1 score of 0.67. These patterns support the conclusion that SMOTE consistently provides the greatest improvement in model performance for this educational multiclassification task, while GAN-generated data may require further optimization to reach comparable results.



**Figure 4.** Comparison for Target 1 across ML algorithms and augmentation strategies.

In summary, the comparative analysis reveals that the combination of the SMOTE augmentation strategy with LR, LSTM, and GB yields the most effective results for predicting Target 1. These models consistently achieve the highest scores across all key performance metrics—accuracy, precision, recall, and F1 score—when trained on data enhanced by SMOTE. This finding underscores the suitability of these algorithms, particularly when paired with robust oversampling, for multi-class classification tasks in EDM scenarios.

The evaluation of Target 2 in the fourth dataset examines the effectiveness of data augmentation strategies applied to the same set of machine learning algorithms as Target 1. The detailed results are presented in Tables 12–15 and visualized in Figure 5. The analysis covers performance across accuracy, precision, recall, and F1 score, providing a clear perspective on the most robust approaches for this multi-class prediction task.

Regarding accuracy in Table 12, SMOTE augmentation consistently boosts classification performance for all models. SVM and LSTM reach the highest accuracy (86%), with ANN and BN close behind at 83%. The non-augmented baseline lags, with most scores in the 60–70% range, peaking at 78% for ANN and LSTM. GAN augmentation underperforms relative to SMOTE and the baseline, with the best accuracy for BN (73%) and LSTM (68%). This reinforces SMOTE's strength in enhancing the predictive accuracy of machine learning models for Target 2.

**Table 12.** Accuracy comparison for Target 2 across ML algorithms and augmentation strategies.

| Augmentation Type  | KNN        | DT         | RF         | SVM        | LR         | AB         | GB         | ANN        | LSTM       | BN         | LDA        |
|--------------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| SMOTE Augmentation | 79%<br>[6] | 77%<br>[6] | 79%<br>[6] | 86%<br>[6] | 71%<br>[6] | 77%<br>[6] | 75%<br>[6] | 83%<br>[6] | 86%<br>[6] | 83%<br>[6] | 76%<br>[6] |
| GAN Augmentation   | 56%        | 52%        | 52%        | 61%        | 56%        | 45%        | 50%        | 62%        | 68%        | 73%        | 53%        |
| No Augmentation    | 63%<br>[6] | 59%<br>[6] | 72%<br>[6] | 68%<br>[6] | 68%<br>[6] | 61%<br>[6] | 68%<br>[6] | 78%<br>[6] | 78%<br>[6] | 66%<br>[6] | 69%<br>[6] |

When examining precision in Table 13, SMOTE once again demonstrates superiority, especially with SVM and LSTM (both 86%), and ANN and BN (83%). In contrast, models without augmentation show more variability, with only a few surpassing 80% (LSTM, 92%; GB, 89%; AB, 75%). GAN-based augmentation provides moderate gains for SVM (75%), but lags for most others; notably, LR (53%) and LDA (49%) see poor precision. These findings suggest that GANs are not the appropriate augmentation technique to enhance precision.

**Table 13.** Precision comparison for Target 2 across ML algorithms and augmentation strategies.

| Augmentation Type  | KNN        | DT         | RF         | SVM        | LR         | AB         | GB         | ANN        | LSTM       | BN         | LDA        |
|--------------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| SMOTE Augmentation | 80%<br>[6] | 79%<br>[6] | 78%<br>[6] | 86%<br>[6] | 72%<br>[6] | 80%<br>[6] | 75%<br>[6] | 83%<br>[6] | 86%<br>[6] | 83%<br>[6] | 76%<br>[6] |
| GAN Augmentation   | 68%        | 66%        | 66%        | 75%        | 53%        | 64%        | 61%        | 63%        | 72%        | 71%        | 49%        |
| No Augmentation    | 75%<br>[6] | 59%<br>[6] | 75%<br>[6] | 74%<br>[6] | 74%<br>[6] | 75%<br>[6] | 89%<br>[6] | 80%<br>[6] | 92%<br>[6] | 77%<br>[6] | 66%<br>[6] |

Recall analysis using Table 14 further emphasizes the positive impact of SMOTE. SVM (87%), LSTM (86%), and GB and BN (83%) all achieve high recall rates, demonstrating excellent sensitivity for correctly identifying true cases. The recall for non-augmented models is generally lower, peaking at LDA (61%) and ANN (60%), and even lower for GAN-augmented models except for BN (77%) and LSTM (63%). These results highlight that, for recall, SMOTE provides a decisive advantage.

**Table 14.** Recall comparison for Target 2 across ML algorithms and augmentation strategies.

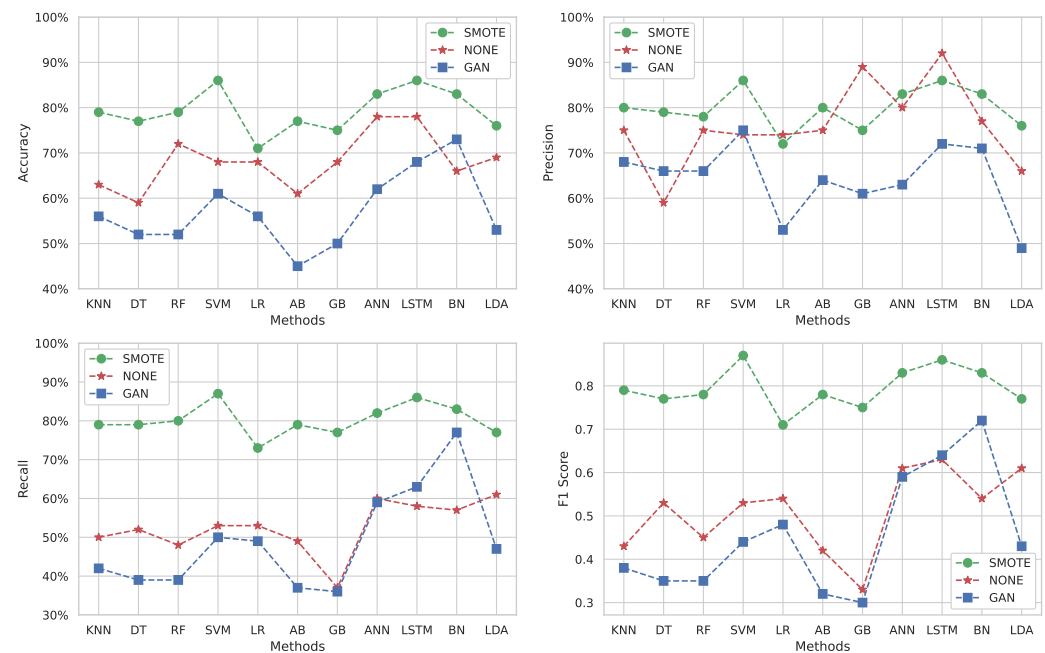
| Augmentation Type  | KNN        | DT         | RF         | SVM        | LR         | AB         | GB         | ANN        | LSTM       | BN         | LDA        |
|--------------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| SMOTE Augmentation | 79%<br>[6] | 79%<br>[6] | 80%<br>[6] | 87%<br>[6] | 73%<br>[6] | 79%<br>[6] | 77%<br>[6] | 82%<br>[6] | 86%<br>[6] | 83%<br>[6] | 77%<br>[6] |
| GAN Augmentation   | 42%        | 39%        | 39%        | 50%        | 49%        | 37%        | 36%        | 59%        | 63%        | 77%        | 47%        |
| No Augmentation    | 50%<br>[6] | 52%<br>[6] | 48%<br>[6] | 53%<br>[6] | 53%<br>[6] | 49%<br>[6] | 37%<br>[6] | 60%<br>[6] | 58%<br>[6] | 57%<br>[6] | 61%<br>[6] |

In terms of the F1 score presented in Table 15, SMOTE leads again, yielding the top results for SVM (0.87), LSTM (0.86), ANN and BN (0.83%), and KNN (0.79). Without augmentation, F1 scores are mostly below 0.63, while GAN augmentation only provides competitive F1 results for BN (0.72) and LSTM (0.64). For most other methods, GANs remain noticeably inferior to the baseline. This consistent trend confirms SMOTE's effectiveness in balancing precision and recall.

**Table 15.** F1 Score comparison for Target 2 across ML algorithms and augmentation strategies.

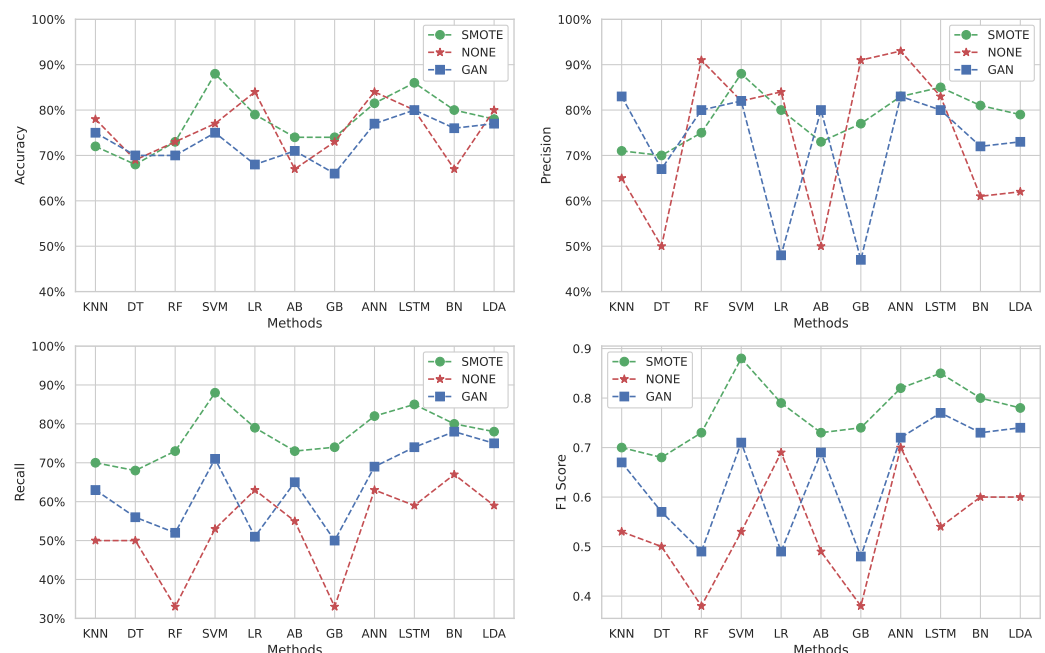
| Augmentation Type  | KNN         | DT          | RF          | SVM         | LR          | AB          | GB          | ANN         | LSTM        | BN          | LDA         |
|--------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| SMOTE Augmentation | 0.79<br>[6] | 0.77<br>[6] | 0.78<br>[6] | 0.87<br>[6] | 0.71<br>[6] | 0.78<br>[6] | 0.75<br>[6] | 0.83<br>[6] | 0.86<br>[6] | 0.83<br>[6] | 0.77<br>[6] |
| GAN Augmentation   | 0.38        | 0.35        | 0.35        | 0.44        | 0.48        | 0.32        | 0.30        | 0.59        | 0.64        | 0.72        | 0.43        |
| No Augmentation    | 0.43<br>[6] | 0.53<br>[6] | 0.45<br>[6] | 0.53<br>[6] | 0.54<br>[6] | 0.42<br>[6] | 0.33<br>[6] | 0.61<br>[6] | 0.63<br>[6] | 0.54<br>[6] | 0.61<br>[6] |

Overall, the results clearly indicate that SMOTE augmentation, when combined with SVM, LSTM, and ANN, offers the best predictive performance for Target 2, reinforcing its value for multi-class academic outcome prediction in EDM.



**Figure 5.** Comparison for Target 2 across ML algorithms and augmentation strategies.

The evaluation of Target 3 explores the comparative effectiveness of augmentation strategies across multiple MLs algorithms. The comprehensive results, presented in Tables 16–19 and Figure 6, shed light on the best approaches for multi-class academic performance prediction in coding courses.



**Figure 6.** Comparison for Target 3 across ML algorithms and augmentation strategies.

For accuracy detailed in Table 16, SVM stands out as the top performer under SMOTE augmentation, achieving 88%, followed closely by LSTM (86%) and ANN (81.5%). The accuracy for most models under SMOTE is comparable or superior to the non-augmented baseline, except for KNN and LR, where non-augmented data achieves higher scores (78% for KNN and 84% for LR). GAN augmentation provides moderate accuracy, with the highest being 80% for LSTM, 77% for ANN and LDA, and 76% for BN, though overall it

does not surpass SMOTE or the baseline in most cases. These trends indicate that SMOTE is the most robust augmentation strategy for boosting accuracy in this multi-class setting.

**Table 16.** Accuracy comparison for Target 3 across ML algorithms and augmentation strategies.

| Augmentation Type  | KNN        | DT         | RF         | SVM        | LR         | AB         | GB         | ANN          | LSTM       | BN         | LDA        |
|--------------------|------------|------------|------------|------------|------------|------------|------------|--------------|------------|------------|------------|
| SMOTE Augmentation | 72%<br>[6] | 68%<br>[6] | 73%<br>[6] | 88%<br>[6] | 79%<br>[6] | 74%<br>[6] | 74%<br>[6] | 81.5%<br>[6] | 86%<br>[6] | 80%<br>[6] | 78%<br>[6] |
| GAN Augmentation   | 75%<br>[6] | 70%<br>[6] | 70%<br>[6] | 75%<br>[6] | 68%<br>[6] | 71%<br>[6] | 66%<br>[6] | 77%<br>[6]   | 80%<br>[6] | 76%<br>[6] | 77%<br>[6] |
| No Augmentation    | 78%<br>[6] | 69%<br>[6] | 73%<br>[6] | 77%<br>[6] | 84%<br>[6] | 67%<br>[6] | 73%<br>[6] | 84%<br>[6]   | 80%<br>[6] | 67%<br>[6] | 80%<br>[6] |

When considering precision in Table 17, SMOTE again delivers the highest values for SVM (88%), LSTM (85%), ANN (83%), and BN (81%). GAN augmentation, however, achieves notable precision for KNN (83%), RF (80%), and LSTM (80%), while significantly underperforming for LR (48%) and GB (47%). The baseline models exhibit very high precision for ANN (93%), GB (91%), ANN (93%), and LSTM (83%). This indicates that while SMOTE generally provides superior precision, models applied to the original dataset can be competitive for certain algorithms but are inconsistent overall.

Recall results presented in Table 18 reinforce SMOTE's effectiveness, with SVM (88%), LSTM (85%), and ANN (82%) again achieving the highest scores. GAN augmentation is strongest for LSTM (74%), ANN (69%), and BN (78%), but overall recall remains lower than SMOTE, except for BN where GANs attain their highest value (78%). The baseline shows substantially lower recall for most models, with only LDA (59%) and LSTM (67%) reaching moderate scores. These findings confirm that SMOTE is most reliable for maximizing the identification of true positives.

**Table 17.** Precision comparison for Target 3 across ML algorithms and augmentation strategies.

| Augmentation Type  | KNN        | DT         | RF         | SVM        | LR         | AB         | GB         | ANN        | LSTM       | BN         | LDA        |
|--------------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| SMOTE Augmentation | 71%<br>[6] | 70%<br>[6] | 75%<br>[6] | 88%<br>[6] | 80%<br>[6] | 73%<br>[6] | 77%<br>[6] | 83%<br>[6] | 85%<br>[6] | 81%<br>[6] | 79%<br>[6] |
| GAN Augmentation   | 83%<br>[6] | 67%<br>[6] | 80%<br>[6] | 82%<br>[6] | 48%<br>[6] | 80%<br>[6] | 47%<br>[6] | 83%<br>[6] | 80%<br>[6] | 72%<br>[6] | 73%<br>[6] |
| No Augmentation    | 65%<br>[6] | 50%<br>[6] | 91%<br>[6] | 82%<br>[6] | 84%<br>[6] | 50%<br>[6] | 91%<br>[6] | 93%<br>[6] | 83%<br>[6] | 61%<br>[6] | 62%<br>[6] |

**Table 18.** Recall comparison for Target 3 across ML algorithms and augmentation strategies.

| Augmentation Type  | KNN        | DT         | RF         | SVM        | LR         | AB         | GB         | ANN        | LSTM       | BN         | LDA        |
|--------------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| SMOTE Augmentation | 70%<br>[6] | 68%<br>[6] | 73%<br>[6] | 88%<br>[6] | 79%<br>[6] | 73%<br>[6] | 74%<br>[6] | 82%<br>[6] | 85%<br>[6] | 80%<br>[6] | 78%<br>[6] |
| GAN Augmentation   | 63%<br>[6] | 56%<br>[6] | 52%<br>[6] | 71%<br>[6] | 51%<br>[6] | 65%<br>[6] | 50%<br>[6] | 69%<br>[6] | 74%<br>[6] | 78%<br>[6] | 75%<br>[6] |
| No Augmentation    | 50%<br>[6] | 50%<br>[6] | 33%<br>[6] | 53%<br>[6] | 63%<br>[6] | 55%<br>[6] | 33%<br>[6] | 63%<br>[6] | 59%<br>[6] | 67%<br>[6] | 59%<br>[6] |

Finally, the F1 score in Table 19 shows that SMOTE continues to provide the leading results, with SVM (0.88), LSTM (0.85), ANN (0.82), and BN (0.80) attaining the highest values. GAN augmentation offers good F1 performance for LSTM (0.77), ANN (0.72), and LDA (0.74), but generally lags behind SMOTE. The non-augmented strategy, while occasionally strong (e.g., ANN, 0.70), does not consistently match the performance achieved through the SMOTE technique. This pattern underscores SMOTE's consistent and substantial advantage in overall predictive performance for Target 3.

**Table 19.** F1 Score comparison for Target 3 across ML algorithms and augmentation strategies.

| Augmentation Type  | KNN         | DT          | RF          | SVM         | LR          | AB          | GB          | ANN         | LSTM        | BN          | LDA         |
|--------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| SMOTE Augmentation | 0.70<br>[6] | 0.68<br>[6] | 0.73<br>[6] | 0.88<br>[6] | 0.79<br>[6] | 0.73<br>[6] | 0.74<br>[6] | 0.82<br>[6] | 0.85<br>[6] | 0.80<br>[6] | 0.78<br>[6] |
| GAN Augmentation   | 0.67        | 0.57        | 0.49        | 0.71        | 0.49        | 0.69        | 0.48        | 0.72        | 0.77        | 0.73        | 0.74        |
| No Augmentation    | 0.53<br>[6] | 0.50<br>[6] | 0.38<br>[6] | 0.53<br>[6] | 0.69<br>[6] | 0.49<br>[6] | 0.38<br>[6] | 0.70<br>[6] | 0.54<br>[6] | 0.60<br>[6] | 0.60<br>[6] |

In summary, across all metrics, the combination of SMOTE augmentation with SVM, LSTM, and ANN models proves most effective for Target 3, reinforcing the superiority of SMOTE for multi-class student performance prediction in EDM.

The results of the comparative analysis across all three targets in the fourth dataset provide important insights into the relative strengths of different machine learning models and data augmentation strategies for multi-class educational prediction tasks. Notably, models such as LSTM, SVM, and ANN consistently rank among the top performers, particularly when combined with SMOTE augmentation. For each target, these models achieve leading scores in accuracy, precision, recall, and F1 score, demonstrating their strong capability to learn complex patterns in educational data, especially when class imbalance and limited data size are addressed through robust augmentation techniques. LSTM's recurrent architecture appears especially well-suited for capturing sequential or behavioral trends, while SVM and ANN offer reliable classification boundaries and flexible feature learning, respectively. The ensemble models, such as GB, also show competitive results for certain targets, but overall, the superiority of LSTM, SVM, and ANN is clear. These findings confirm that, for multi-class prediction in educational settings, leveraging deep learning and advanced augmentation such as SMOTE yields the most reliable and generalizable performance, suggesting a strong direction for future research and practical applications in early student success identification.

These findings are further corroborated by the comprehensive study conducted by Hafdi et al. [6], which rigorously evaluated the effectiveness of SMOTE-based data augmentation in multi-class and binary classification contexts for early prediction of student performance. Their research demonstrated that SMOTE consistently enhances the performance of various machine learning algorithms—mirroring the results observed in our experiments, and confirmed the statistical significance of these improvements using robust statistical validation methods such as *t*-tests. By systematically comparing the augmented and original datasets, Hafdi et al. established that SMOTE enables early detection of at-risk students with higher reliability, thus reinforcing its suitability and efficacy as a data augmentation technique within EDM. The alignment of our results with those of Hafdi et al. [6] provides additional confidence in recommending SMOTE as a preferred augmentation strategy for predictive analytics in the educational domain.

The comparative analysis of our experiments across four diverse educational datasets yields several key insights regarding the efficacy of data augmentation techniques and model selection in EDM tasks.

For Dataset 1, the combination of SMOTE with KNN (79.2%) and ANN (77%) delivered the highest accuracy, notably surpassing the original baseline and the GAN-augmented results, with GANs generally failing to outperform the non-augmented scenario. In Dataset 2, SMOTE again demonstrated clear superiority, driving ANN accuracy to 92.8% an improvement of more than 30% compared to the baseline while KNN also benefited (78.6%). Here, GANs' best outcome (75% with LR) did exceed the original dataset but still lagged behind SMOTE.

Turning to Dataset 3, SMOTE consistently yielded remarkable results, achieving 99% accuracy and an F1 score of 0.98 with ANN, and 98% accuracy with KNN. GAN augmentation also surpassed the baseline with DT and AB (both at 95%), and produced strong precision, recall, and F1 metrics for select models. However, these gains of the GAN strategy make it promising in other scenarios and necessitate further research.

For the fourth dataset involving multi-class prediction in coding courses, the trends held firm. For Target 1, LR and LSTM, when paired with SMOTE, reached the best results, while GAN augmentation generally led to performance declines relative to the baseline, except in a few isolated cases. In Target 2, SVM and LSTM (both 86% accuracy and F1 scores of 0.87 and 0.86, respectively) clearly benefited from SMOTE, while GANs mostly resulted in lower metrics, except for a minor increase with BN. Regarding Target 3, SMOTE once more dominated, with SVM leading, followed by LSTM, and an overall stable performance increase across all methods. GAN augmentation surpassed the baseline in select cases (LSTM, BN, and LDA), but its impact remained inconsistent and model-dependent.

Despite the growing popularity of GANs for synthetic data generation in many domains, our results highlight important limitations when applying GANs to EDM tasks. Firstly, GANs typically require large amounts of data to train effectively, whereas most EDM applications are constrained by small, highly structured datasets. This often leads to unstable training, mode collapse, or the generation of synthetic samples that poorly capture the true data distribution. In our experiments, the GAN-augmented strategy often resulted in decreased or inconsistent performance across most ML models, particularly in precision and recall metrics. Additionally, GAN-generated data can introduce artifacts or unrealistic patterns that do not reflect genuine student behaviors, thereby undermining model interpretability and reliability, both of which are essential in sensitive educational contexts. Consequently, while GANs offer potential for complex and high-dimensional data domains, their application to small-scale, tabular, and highly contextualized EDM datasets remains challenging and unreliable.

In summary, the experimental findings establish SMOTE as the most reliable and effective augmentation strategy for small and imbalanced educational datasets, due to its ability to generate synthetic samples that bridge minority and majority classes while preserving original data structure. GANs, while promising in theory, were generally less effective in the constrained and tabular EDM setting, producing less stable improvements and often underperforming compared to SMOTE. Collectively, this study provides a broad and generalizable overview of data augmentation and model selection for student outcome prediction in EDM, especially for institutions or researchers working with limited or imbalanced datasets and seeking robust, early identification of at-risk students across varying educational environments.

#### *4.4. Statistical Validation Using Paired t-Test*

To further validate the robustness of the reported performance improvements and to ensure that the observed gains were not only due to a favorable random train/test split or random initialization, we added a statistical validation step based on a paired *t*-test. This analysis was conducted on the best-performing model identified for each dataset. For each selected model, the experiment was repeated five times under the same preprocessing, augmentation, training, and evaluation conditions, while changing only the random seed used for the data split and model initialization when applicable. This allowed us to observe the stability of the model performance across different random configurations.

For each dataset, we compared the performance of the best model trained using the augmented training data with the corresponding baseline model. The paired *t*-test was applied because the baseline and augmented results were obtained under matched

experimental conditions using the same sequence of random seeds. Therefore, each baseline result was directly paired with its corresponding augmented result. The purpose of this test was to examine whether the improvement obtained after data augmentation was statistically meaningful rather than being only the result of random variation.

This statistical validation was repeated for the best-performing method of each dataset. The obtained results provide additional evidence about the stability of the selected models and their ability to capture meaningful predictive patterns from the educational features rather than relying on random effects. However, given the small size of the datasets and the limited number of repeated runs, the statistical test is interpreted as a complementary validation step.

The paired *t*-test results further support the stability of the best-performing SMOTE-augmented models across repeated runs. Since accuracy was the common metric reported across the original studies, it was selected for this statistical comparison. For Dataset 1, the KNN model with SMOTE augmentation achieved the best accuracy compared with both the original baseline and the GAN-augmented results. The paired *t*-test yielded a *t*-statistic of  $-24.65$  and a *p*-value of  $1.61 \times 10^{-5}$ , indicating a statistically significant improvement. For Dataset 2, the ANN model with SMOTE augmentation showed the strongest performance, with a *t*-statistic of  $-28.61$  and a *p*-value of  $8.88 \times 10^{-6}$ . Similarly, for Dataset 3, the ANN model with SMOTE augmentation outperformed the corresponding baseline and GAN-based augmentation, with a *t*-statistic of  $-14.52$  and a *p*-value of  $1.31 \times 10^{-4}$ . For Dataset 4, the statistical validation was conducted separately for each prediction target. For Target 1, the LR model with SMOTE augmentation achieved the best performance, with a *t*-statistic of  $-20.12$  and a *p*-value of  $3.60 \times 10^{-5}$ . For Target 2, the SVM model with SMOTE augmentation obtained a *t*-statistic of  $-57.67$  and a *p*-value of  $5.41 \times 10^{-7}$ . For Target 3, the SVM model again achieved the best SMOTE-augmented performance, with a *t*-statistic of  $-32.86$  and a *p*-value of  $5.11 \times 10^{-6}$ . In all cases, the *p*-values are well below the conventional significance threshold of 0.05, suggesting that the observed improvements are unlikely to be due only to random splitting or random initialization. The negative sign of the *t*-statistics reflects the direction of the comparison according to the subtraction order used in the paired test; in practical terms, the results indicate that the SMOTE-augmented models consistently achieved higher accuracy than their corresponding baselines across the repeated runs.

## 5. Conclusions

This research contributes to the field of Educational Data Mining (EDM) by examining the potential of data augmentation techniques to improve early academic outcome prediction in small and imbalanced educational datasets. The findings suggest that data augmentation can improve the predictive performance of machine learning models in several educational contexts, particularly when SMOTE is used. Compared with GAN-based augmentation, SMOTE generally produced more stable and consistent improvements across datasets and classifiers. This can be explained by the fact that SMOTE generates synthetic samples through interpolation between existing minority-class instances, which helps preserve the local structure of small tabular datasets. In contrast, GAN-based augmentation showed more mixed results, probably because GANs usually require larger datasets to learn stable data distributions and may generate less reliable samples when applied to small, structured, and highly contextual educational data.

The results also show that the benefit of data augmentation is not uniform across all models. Some classifiers, such as KNN, ANN, LR, and SVM, benefited more clearly from SMOTE in specific datasets or target variables, while other models showed limited or inconsistent improvements. This indicates that the effectiveness of augmentation depends

on several factors, including dataset size, class distribution, feature characteristics, target variable, classifier type, and evaluation protocol. Therefore, data augmentation should not be considered a universally effective solution, but rather a strategy that must be carefully adapted to the characteristics of each educational prediction task.

To further validate the observed improvements, a paired *t*-test was added using five repeated runs with different random seeds for the best-performing SMOTE-augmented models. Accuracy was selected for this validation because it was the common metric reported across the original studies. The statistical results supported the stability of the best SMOTE-based improvements. These results suggest that the improvements obtained with SMOTE are unlikely to be due only to random splitting or random initialization. Nevertheless, these statistical tests should be interpreted as complementary validation, given the limited size of the datasets and the small number of repeated runs.

Despite these promising findings, several limitations should be acknowledged. First, the study relies on previously published datasets, and although the original experimental conditions were reproduced as closely as possible, differences between datasets in terms of institutional context, student population, features, target variables, and data collection procedures limit the generalizability of the findings. Second, the small size of the datasets remains a central challenge, especially for complex generative methods such as GANs. Third, the evaluation focused mainly on standard classification metrics, while future work could include additional validation strategies, interpretability analysis, and external testing on independent datasets.

Future research could explore more advanced and context-aware augmentation strategies, including ensemble-based augmentation, hybrid SMOTE variants, and improved generative models designed specifically for small tabular educational datasets. In addition, integrating richer behavioral data, such as LMS activity logs, clickstream patterns, submission behavior, and engagement indicators, may provide deeper insights into student learning processes. Finally, validating the proposed approaches across larger multi-institutional datasets would help develop more generalizable and scalable predictive models for early identification of at-risk students in diverse academic environments.

**Author Contributions:** S.E.K.: Conceptualization, Methodology, Supervision, Visualization, Investigation, Project administration, Writing—original draft, Formal analysis, Writing—review and editing. Z.S.H.: Conceptualization, Methodology, Software, Formal analysis, Writing—original draft, Visualization, Investigation, Data Curation, Implementation. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The datasets used in this study were obtained from previously published studies and/or publicly available academic sources. No new primary dataset was collected. The data were used only for secondary analysis and comparative evaluation, and their availability remains subject to the conditions defined by the original authors or data sources.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Hafdi, Z.S.; El Kafhali, S. Student Performance Prediction in Learning Management System Using Small Dataset. In *Proceedings of the International Conference on Artificial Intelligence and Computer Vision*; Springer: Cham, Switzerland, 2023; pp. 197–205.
2. Zohair, A.; Mahmoud, L. Prediction of Student's performance by modelling small dataset size. *Int. J. Educ. Technol. High. Educ.* **2019**, *16*, 27. [[CrossRef](#)]

3. Hafdi, Z.S.; El Kafhali, S. A Hybrid Ensemble Learning Framework for Early Dropout Prediction in Learning Management Systems. In *Proceedings of the Artificial Intelligence and Cognitive Sciences for Emerging Technologies: Proceedings of the International Conference on Artificial Intelligence and Cognitive Science for Emerging Technologies (AICSET 2025)*; Springer Nature: Cham, Switzerland, 2026; Volume 2, pp. 3–14.
4. Kabathova, J.; Drlik, M. Towards predicting student's dropout in university courses using different machine learning techniques. *Appl. Sci.* **2021**, *11*, 3130. [\[CrossRef\]](#)
5. Chowdhury, M.M.; Ayon, R.S.; Hossain, M.S. An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced BRFSS dataset. *Healthc. Anal.* **2024**, *5*, 100297.
6. Hafdi, Z.S.; El Kafhali, S. A Comparative Evaluation of Machine Learning Methods for Predicting Student Outcomes in Coding Courses. *AppliedMath* **2025**, *5*, 75. [\[CrossRef\]](#)
7. Song, Z.; Sung, S.H.; Park, D.M.; Park, B.K. All-year dropout prediction modeling and analysis for university students. *Appl. Sci.* **2023**, *13*, 1143. [\[CrossRef\]](#)
8. Santiago-Muñoz, F.; Larrañaga, M.; Elorriaga, J.A.; Arruarte, A. Identification of Patterns to Prevent the Failure of Engineering Students. *Int. J. Artif. Intell. Educ.* **2025**, *35*, 3356–3405. [\[CrossRef\]](#)
9. Arévalo Cordovilla, F.E.; Peña Carrera, M. Comparative analysis of machine learning models for predicting student success in online programming courses: A study based on LMS data and external factors. *Mathematics* **2024**, *12*, 3272. [\[CrossRef\]](#)
10. Realinho, V.; Machado, J.; Baptista, L.; Martins, M.V. Predicting student dropout and academic success. *Data* **2022**, *7*, 146. [\[CrossRef\]](#)
11. Riestra-González, M.; del Puerto Paule-Ruiz, M.; Ortin, F. Massive LMS log data analysis for the early prediction of course-agnostic student performance. *Comput. Educ.* **2021**, *163*, 104108. [\[CrossRef\]](#)
12. Kaensar, C.; Wongnin, W. Analysis and Prediction of Student Performance Based on Moodle Log Data using Machine Learning Techniques. *Int. J. Emerg. Technol. Learn.* **2023**, *18*, 184–203. [\[CrossRef\]](#)
13. Hung, H.C.; Liu, I.F.; Liang, C.T.; Su, Y.S. Applying educational data mining to explore students' learning patterns in the flipped learning approach for coding education. *Symmetry* **2020**, *12*, 213. [\[CrossRef\]](#)
14. Ciolacu, M.; Tehrani, A.F.; Binder, L.; Svasta, P.M. Education 4.0-Artificial Intelligence assisted higher education: Early recognition system with machine learning to support students' success. In *Proceedings of the 2018 IEEE 24th International Symposium for Design and Technology in Electronic Packaging (SIITME)*; IEEE: New York, NY, USA, 2018; pp. 23–30.
15. Villegas-Ch, W.; Román-Cañizares, M.; Palacios-Pacheco, X. Improvement of an online education model with the integration of machine learning and data analysis in an LMS. *Appl. Sci.* **2020**, *10*, 5371. [\[CrossRef\]](#)
16. Azimi, S.; Popa, C.G.; Cucić, T. Improving Students Performance in Small-Scale Online Courses—A Machine Learning-Based Intervention. *arXiv* **2020**, arXiv:2012.01187.
17. Naicker, N.; Adeliyi, T.; Wing, J. Linear support vector machines for prediction of student performance in school-based education. *Math. Probl. Eng.* **2020**, *2020*, 4761468. [\[CrossRef\]](#)
18. Mueen, A.; Zafar, B.; Manzoor, U. Modeling and predicting students' academic performance using data mining techniques. *Int. J. Mod. Educ. Comput. Sci.* **2016**, *8*, 36–42. [\[CrossRef\]](#)
19. Hussain, S.; Khan, M.Q. Student-performulator: Predicting students' academic performance at secondary and intermediate level using machine learning. *Ann. Data Sci.* **2023**, *10*, 637–655. [\[PubMed\]](#)
20. Bertolini, R.; Finch, S.J.; Nehm, R.H. Enhancing data pipelines for forecasting student performance: Integrating feature selection with cross-validation. *Int. J. Educ. Technol. High. Educ.* **2021**, *18*, 1–23. [\[CrossRef\]](#)
21. Sulak, S.A.; Koklu, N. Predicting Student Dropout Using Machine Learning Algorithms. *Intell. Methods Eng. Sci.* **2024**, *3*, 91–98.
22. Dang, X.T. Improve the Effectiveness of Predicting Student Dropouts Based on Deep Learning and SMOTE Models. In *Machine Learning and Other Soft Computing Techniques: Biomedical and Related Applications*; Springer: Cham, Switzerland, 2024; pp. 169–181.
23. Shang, Y.; Su, X.; Xiao, Z.; Chen, Z. Campus sentiment analysis with gan-based data augmentation. In *Proceedings of the 2021 13th International Conference on Advanced Infocomm Technology (ICAIT)*; IEEE: New York, NY, USA, 2021; pp. 209–214.
24. Yağcı, M. Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learn. Environ.* **2022**, *9*, 11. [\[CrossRef\]](#)
25. Ersozlu, Z.; Taheri, S.; Koch, I. A review of machine learning methods used for educational data. *Educ. Inf. Technol.* **2024**, *29*, 22125–22145. [\[CrossRef\]](#)
26. Roslan, M.B.; Chen, C. Educational data mining for student performance prediction: A systematic literature review (2015–2021). *Int. J. Emerg. Technol. Learn. (iJET)* **2022**, *17*, 147–179.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.