

Article

The Effectiveness of Generative Adversarial Network-Based Oversampling Methods for Imbalanced Multi-Class Credit Score Classification

I Nyoman Mahayasa Adiputra ¹, Pei-Chun Lin ^{2,*}  and Paweena Wanchai ¹

¹ College of Computing, Khon Kaen University, Khon Kaen 40002, Thailand; mahayasa.a@kkumail.com (I.N.M.A.); wpaweena@kku.ac.th (P.W.)

² Department of Information Engineering and Computer Science, Feng Chia University, Taichung 40724, Taiwan

* Correspondence: peiclin@fcu.edu.tw

Abstract: Credit score models are essential tools for evaluating creditworthiness and mitigating financial risks. However, the imbalanced nature of multi-class credit score datasets poses significant challenges for traditional classification algorithms, leading to poor performance in minority classes. This study explores the effectiveness of Generative Adversarial Network (GAN)-based oversampling methods, including CTGAN, CopulaGAN, WGAN-GP, and DraGAN, in addressing this issue. By synthesizing realistic data for minority classes and integrating it with majority class data, the study benchmarks these GAN-based methods across classical (KNN, Decision Tree, Logistic Regression) and ensemble machine learning models (XGBoost, Random Forest, LightGBM). Evaluation metrics such as accuracy and F1-score reveal that WGAN-GP consistently achieves superior performance, especially when combined with Random Forest, outperforming other methods in balancing dataset representation and enhancing classification accuracy. The results showed that WGAN-GP + RF achieved 0.873 in accuracy, 0.936 F1-score in the “good” class, 0.806 F1-score in the “poor” class, and 0.816 F1-score in the “standard” class. The findings underscore the potential of GAN-based oversampling in improving multi-class credit score classification and highlight future directions, including hybrid sampling and cost-sensitive learning, to address remaining challenges.

Keywords: credit score classification; multi-class; GAN-based oversampling; imbalanced data; machine learning



Academic Editor: George A. Tsihrintzis

Received: 12 January 2025

Revised: 6 February 2025

Accepted: 7 February 2025

Published: 11 February 2025

Citation: Adiputra, I.N.M.; Lin, P.-C.; Wanchai, P. The Effectiveness of Generative Adversarial

Network-Based Oversampling Methods for Imbalanced Multi-Class Credit Score Classification. *Electronics* **2025**, *14*, 697. <https://doi.org/10.3390/electronics14040697>

Copyright: © 2025 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) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Credit score classification is a method used by financial institutions to evaluate the creditworthiness of individuals or businesses applying for loans or credit [1]. The primary goal of credit score is to determine whether a credit applicant belongs to the creditworthy or non-creditworthy category. This process is not accomplished in a single step; instead, financial institutions often conduct it through multiple stages, including application scoring, behavioral scoring, and collection scoring, among others [2]. Credit scores are evaluated based on credit expert analysis from financial institutions [3]. Machine learning methods might help the expert to make a better decision by classifying a credit score based on data. Identifying a credit score is an important task to make a credit approval process more efficient. One of the problems of credit score classification is imbalanced data on the credit scoring dataset; the imbalanced data caused the classifier to fail to learn the minority

data [4]. To overcome this issue, oversampling methods have been conducted in several studies; techniques such as random oversampling, ADAYSN, and SMOTE are commonly used [5,6]. However, the common oversampling methods had an issue of data distribution that was produced after sampling because the real data are not based on sampling the synthetic data. To establish data distribution, generative adversarial networks (GANs) [7] were proposed and used to solve imbalanced data on multi-class [8].

Despite the growing application of GAN-based oversampling techniques in handling class imbalance, most existing studies have primarily focused on binary classification tasks. The effectiveness of these methods in multi-class credit score classification remains largely underexplored.

This paper addresses the gap by systematically benchmarking multiple GAN architectures to evaluate their impact on imbalanced multi-class data. By comparing different GAN-based oversampling methods, we provide a comprehensive analysis of their strengths and limitations in enhancing classification performance. Our findings offer valuable insights into the suitability of various GAN models for multi-class credit scores, contributing to the development of more robust and generalizable oversampling strategies in financial risk assessment.

In this study, we provided a comprehensive benchmark on GAN-based oversampling methods to solve imbalanced multi-class using credit score data. The experiment was conducted on several GAN-based methods: WGAN-GP, CTGAN, DraGAN, and CopulaGAN, each synthetic data produced by GAN applied to classical machine learning: K-Nearest Neighbor (KNN), Decision Tree (DT), Logistic Regression (LR), and ensemble machine learning: XGBoost, Random Forest (RF), Light Gradient-Boosting (LGB). To evaluate the performance of GAN-based oversampling, this study presents overall accuracy, and to get more insight into imbalanced data, this study presents the F1-score metric for each sampling and algorithm combination. Specifically, we seek to address the following research questions pertaining to the credit score challenge:

- RQ1: Are GAN-based oversampling methods improving the performance of multi-class credit score classification?
- RQ2: How effective are GAN-based oversampling methods on credit score classification performance?
- RQ3: What is the most effective combination of GAN-based oversampling methods and machine learning algorithms for credit score classification?

To answer the questions above, we provide and discuss the contributions as follows:

- A benchmark comparison of several GAN-based oversampling method performances.
- An evaluation of classical and ensemble machine learning algorithms before and after GAN-based oversampling data.
- Comparison of real data and synthetic data distribution visualization on GAN-based has the best performance.

The rest of this paper is structured as follows. Section 2 provides a systematic review of related work on credit score classification. In Section 3, the baseline methods of the framework are described in detail. Section 4 presents the experimental results along with relevant discussions. Finally, Section 5 summarizes the contributions of this study.

2. Related Works

Credit score classification is a crucial aspect of financial decision-making processes in various organizations, particularly in banking sectors. Different studies have explored the use of various machine learning techniques to develop effective credit score models. Huang et al. (2007) [9] utilized Support Vector Machines (SVMs) in credit scores, highlight-

ing the importance of data mining approaches. Mohammadi et al. (2016) [10] focused on Artificial Neural Networks, specifically Multilayer Perceptron Neural Network models trained using Back-Propagation algorithms for customer credit risk assessment. Imtiaz et al. (2017) [11] emphasized the significance of parameter tuning and explored credit score classification models with and without imputation techniques. Tripathi et al. (2020) [12] proposed a hybrid credit score model combining feature selection with Binary BAT optimization technique and Radial Basis Function Neural Network (RBFN). Kuppili et al. (2020) [13] introduced a novel spike-generating function in the Leaky Nonlinear Integrate and Fire Model for credit score classification. Parvin et al. (2020) [14] conducted a comparative analysis of base and ensemble classification algorithms for credit scores, highlighting the importance of statistical data analysis in financial institutions. Veeramanikandan et al. (2020) [15] presented a parameter-tuned deep learning model for credit risk assessment and scoring applications, focusing on predicting credit scores for loan applicants. Ahmed et al. (2021) [16] conducted a comparative study of machine learning techniques for credit card fraud detection, emphasizing the importance of selecting the best classification technique. Maurya et al. (2023) [17] proposed a Decision Tree Classifier-based ensemble approach for credit score classification, achieving high accuracy rates and addressing the dynamic nature of credit scores in the financial sector. Overall, these studies demonstrate the diverse approaches and techniques used in credit score classification to enhance financial decision-making processes.

Ziemba et al. (2021) [18] performed a credit risk assessment classification. From their result, a combination of a correlation-based feature selection method and a random forest classifier demonstrates superior performance in addressing the given problem. Sotiropoulos et al. (2024) [19] improved the model's ability to credit score classification by utilizing ensemble methods or more sophisticated feature selection approaches. Hayashi (2022) [20] highlighted the latest methods that have demonstrated higher accuracy compared to ensemble classifiers, their hybrids, rule extraction techniques, and rule-based classifiers.

The common problem of credit score classification is a multi-class imbalanced problem. Studies of this problem have been conducted for several years, and the most common solution was ensemble learning [21–24]. Ensemble learning is a machine learning technique that combines multiple models to reinforce the performance of a single model [25]. Another approach was a feature selection technique called Bolasso (Bootstrap-Lasso) [26], which works by selecting consistent and relevant features, shortlisting features, and then applied to machine learning algorithms; Bolasso with Random Forest algorithm achieved the highest AUC compared to the other feature selection techniques. In recent years, the use of Generative Adversarial Networks (GANs) in oversampling methods for imbalanced datasets has gained significant attention in the field of machine learning. Nekooimehr et al. (2016) [27] introduced the Adaptive Semi-supervised Weighted Oversampling (A-SUWO) method to address imbalanced datasets. Douzas et al. (2018) [28] proposed a heuristic oversampling method based on K-means and SMOTE to improve imbalanced learning. Gangwar et al. (2019) [29] presented a novel GAN-based oversampling method for data in credit card fraud detection, which outperformed traditional methods such as SMOTE and ADASYN in terms of precision, F1-score, and reduction in false positives. Salazar et al. (2020) [30] explored new applications of GAN-based oversampling methods and compared the results with SMOTE. Engelmann et al. (2020) [31] focused on conditional Wasserstein GAN-based oversampling for tabular data, emphasizing the importance of downstream classification tasks. Li et al. (2021) [32] combined K-means clustering with GAN for fault diagnosis, showing improved diagnostic accuracy for minority-class samples. Dablain et al. (2021) [33] introduced DeepSMOTE, an oversampling algorithm for deep learning models that does not require a discriminator. Wang et al. (2022) [34] proposed

an intelligent identification method for the distribution of network line–transformer relationships using GAN processing of unbalanced data. García-Vicente et al. (2023) [35] evaluated synthetic categorical data generation techniques for predicting cardiovascular diseases, including CTGAN. Liu et al. (2024) [36] addressed the joint issue of imbalance and concept drift in anomaly detection using an ensemble learning method with GAN-based sampling and consistency check. Overall, GAN-based oversampling methods have shown promise in addressing imbalanced datasets across various domains, offering improvements in predictive performance and diagnostic accuracy.

Machine learning methods, particularly those utilizing Generative Adversarial Networks (GANs), have been increasingly employed in various domains for tasks such as image-level domain transfer, attack detection, landslide data balancing, and medical data generation. Cai et al. (2019) [37] introduced a Supervised Class Distribution Learning framework for GAN-based imbalanced classification to address biased and inaccurate classifiers. Li et al. (2020) [38] evaluated GAN-based image-level transfer methods on convolutional neural network models, highlighting the challenges in generalization when models are trained on one dataset and tested on another. Additionally, Li et al. (2020) [39] proposed a hybrid oversampling model using GAN to enhance attack detection performance in anomaly-based NIDS. Al-Najjar et al. (2021) [40] presented an integrated approach for landslide data balancing and spatial prediction using GAN, demonstrating its effectiveness in comparison to standard methods like SMOTE. Abedi et al. (2022) [41] investigated the use of GAN models to generate synthetic data to improve classification accuracy, particularly in scenarios with limited data availability. Furthermore, Duan et al. (2022) [42] explored privacy-preserving data synthesis for tabular data using a federated generative model, emphasizing the challenges posed by multimodal distributions and imbalanced attributes in decentralized environments. In the medical domain, Serte et al. (2022) [43] proposed a data-efficient deep network for COVID-19 detection on CT images, showcasing the potential of artificial intelligence techniques in medical image analysis. Rafiei et al. (2023) [44] conducted a study on integrating synthetic data into electronic medical records to enhance machine learning predictions, specifically focusing on fluid overload prediction in ICU patients. Moreover, Kaabachi et al. (2024) [45] introduced a method for evaluating privacy risks in GAN by leveraging the discriminator outputs of the standard GAN architecture. Overall, the literature demonstrates the diverse applications of machine learning methods with GAN datasets across various domains, highlighting their potential for improving classification accuracy, data synthesis, privacy preservation, and predictive modeling in challenging scenarios. While many studies on multi-class credit score classification were conducted by ensemble learning, the effectiveness of the GAN method on multi-class credit score classification cases remains underexplored. Based on the effectiveness of the GAN method in multi-class classification tasks, this study presents a framework and benchmarking analysis of several GAN methods on imbalanced multi-class credit score problems.

3. Methods

3.1. Research Framework

The GAN-based oversampling method in this study is illustrated in Figure 1. The objective of the framework is to improve machine learning performance on imbalanced datasets through oversampling; balancing the data and evaluating across multiple classifiers provide a robust approach for accurate and fair credit scores.

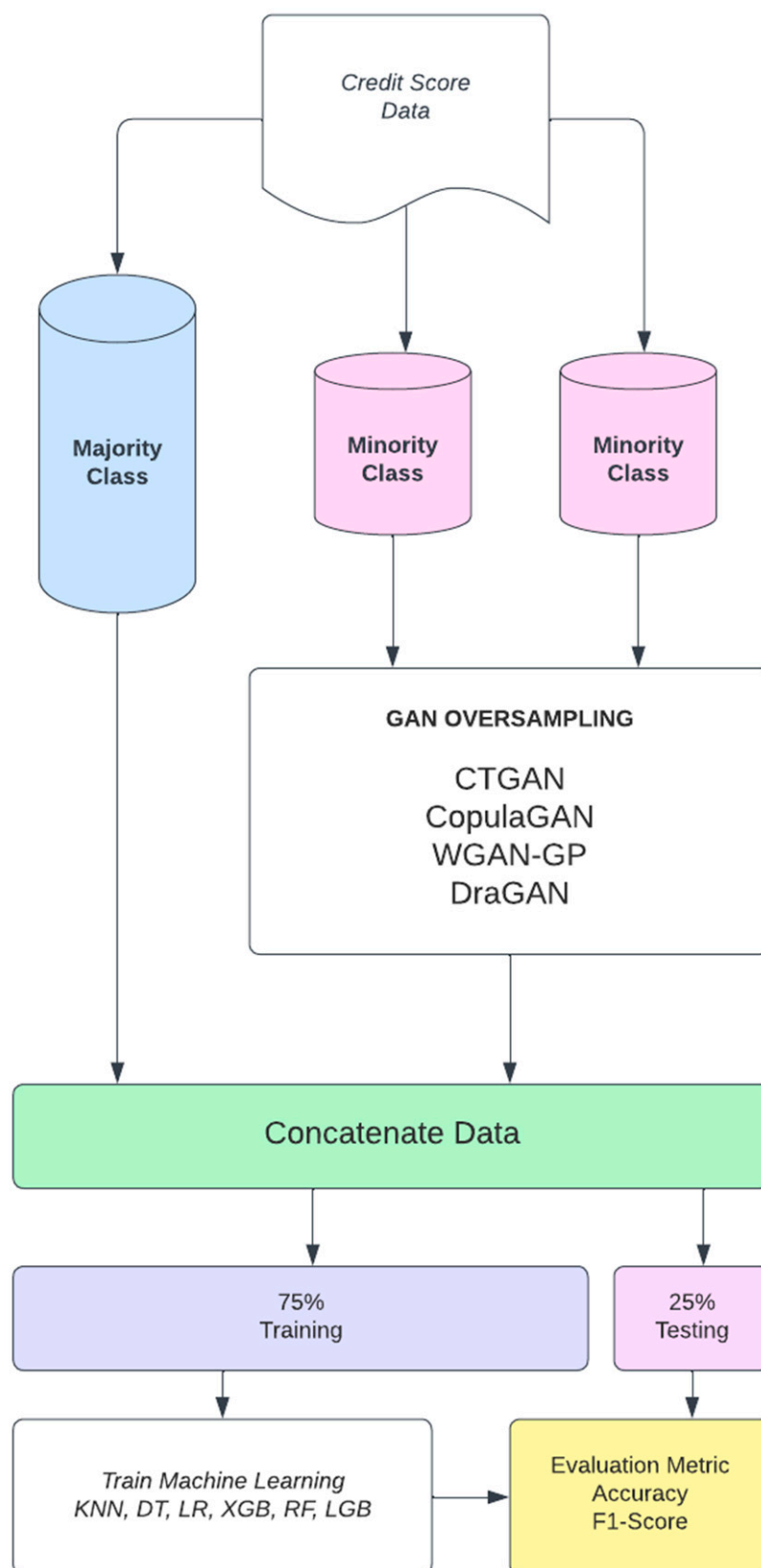


Figure 1. GAN-based oversampling method on the multi-class framework.

The framework begins with distinguishing data into majority and minority classes. As the problem of this study is multi-class, the minority data will be more than one class. The minority classes' data undergo oversampling through GAN. Four GAN-based models are applied: CTGAN, CopulaGAN, WGAN-GP, and DraGAN. These GAN models are

employed to synthesize additional instances of the minority class, enhancing the representation of the minority group in the data. The generated synthetic and real minority class data are combined with the original majority class data, creating a balanced dataset. The last step is evaluation metrics using accuracy and F1-score, which provided a comprehensive insight, especially on imbalanced data.

Different types of GAN and machine learning methods allow for benchmarking the effectiveness of the GAN method on specific types of machine learning and imbalanced multi-class data.

3.2. GAN Method

A generative adversarial network (GAN) is a neural network that generates synthetic data [7]. GAN has two main parts: generator and discriminator. The generator's main objective is to generate a random instance, and then the discriminator has a function to discriminate whether the generated instance is real or fake.

Figure 2 illustrates the GAN architecture, and the discriminator determines the synthetic data by calculating how similar the generated data are. The loss function of GAN is derived from the Binary Cross-Entropy (BCE) Loss. It is defined as seen in Formula (1), which aims to minimize the generator by producing fake data and maximizing the discriminator by determining real data from generated instances. Recently, the GAN method had various types based on specific purposes; this study observed the performance across types of GAN methods on multi-class classification problems. The GAN methods are CTGAN [46], CopulaGAN [47], WGAN-GP [48], and DraGAN [49]. The comparison of the GAN methods in various aspects is described in Table 1.

$$\min_G \max_D V(G, D) \quad (1)$$

where the value function $V(G, D)$ is:

$$V(G, D) = E_{x \sim p_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log(1 - D(G(z)))]$$

where

$x \sim p_{data}(x)$ is a real sample from the true data distribution.

$z \sim p_z(z)$ is a random noise sample from a prior distribution (e.g., Gaussian or uniform).

$G(z)$ is the synthetic data (fake data) generated from a latent noise variable z .

$D(x)$ is the discriminator's probability that x is real data.

$D(G(z))$ is the discriminator's probability that $G(z)$ is real (but it's actually fake).

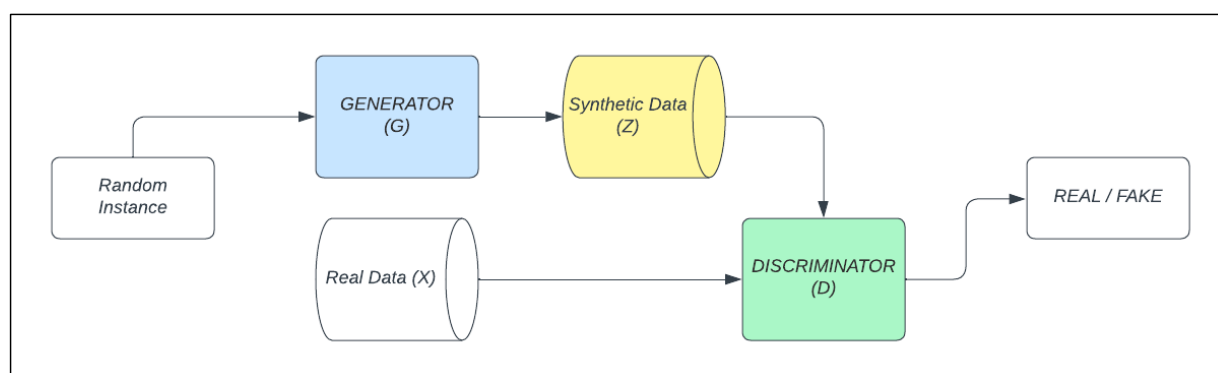


Figure 2. Generative adversarial network architecture.

Table 1. GAN methods aspect comparison.

Aspect	CTGAN	CopulaGAN	WGAN-GP	DraGAN
Data handling	Design for tabular data	Tabular data leveraging copulas to model relationships	Various types of data	Often applied for time-series and image data
Gradient penalty	No gradient penalty	No gradient penalty	Enforces gradient norm of 1 over interpolated points	Enforces gradient regularization near real data points
Regularization focus	Regularizes the generation of categorical data via a mode-specific sampling	Copula-based statistical modeling	Global Lipschitz regularization	Localized regularization around real data samples
Core algorithm	Conditional GAN (cGAN) with a mode-specific normalization	Combines copula statistical modeling with GAN-based	Wasserstein loss with a gradient penalty	Penalizes gradients near noisy real data points
Strengths	Effective for tabular datasets, especially with mixed data types	Explicitly models dependencies using copulas	Globally stabilizes training, widely applicable across domains	Better localized gradient handling
Weaknesses	Requires careful tuning for high-dimensional datasets	Performance depends on the copula choice and is less flexible for high-dimensional	Gradient penalty may fail with poorly generated samples	Requires noise injection and is less effective in global regularization

3.3. Evaluation Metrics

This study used two main metrics to evaluate the performance of the GAN methods on machine learning; the evaluation metrics are accuracy and F1-score. Accuracy quantifies the proportion of correctly predicted instances out of the total instances in the dataset. To evaluate the model's performance on imbalanced data, the F1-score is commonly used, where accurately predicting minority class instances is crucial. It penalizes extreme trade-offs between precision and recall, ensuring a balanced evaluation. Accuracy and F1-score are measured by Equations (2)–(5) as follows:

$$ACC = (TP + TN) / (TP + TN + FP + FN) \quad (2)$$

$$RC = TP / (TP + FN) \quad (3)$$

$$PR = TP / (TP + FP) \quad (4)$$

$$F1 - Score = 2 * (PR * RC) / (PR + RC) \quad (5)$$

where TP is True Positives (correctly predicted positive class), TN is True Negatives (correctly predicted negative class), FP is False Positives (incorrectly predicted as positive), and FN is False Negatives (incorrectly predicted as negative).

4. Results and Experiments

4.1. Dataset Analysis

The dataset in this study was collected from a public source, the Kaggle platform. The dataset is a credit score multi-class classification problem with 99,960 rows of data, 21 features, and 3 labeled classes: good (0), poor (1), and standard (2) [50]. Figure 3 visualizes the class distribution; the distribution of class is imbalanced, and the standard class is the majority class with 53.2% numbers.

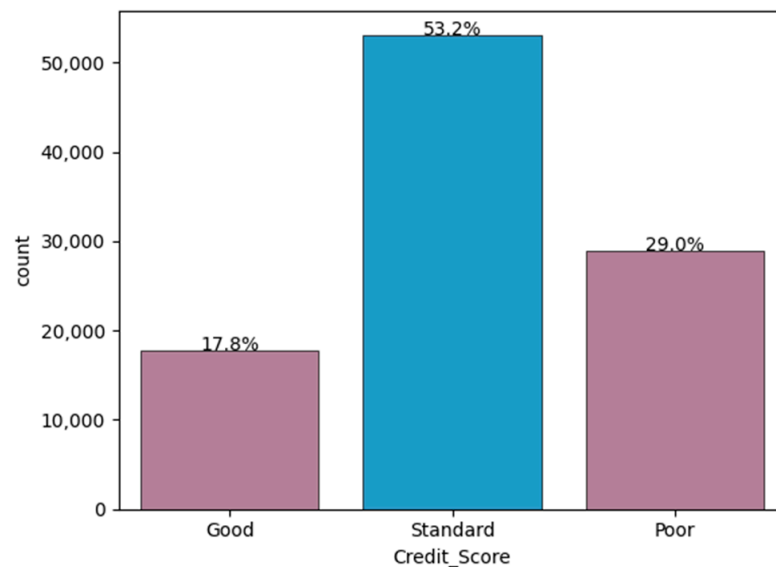


Figure 3. Credit score class distribution.

According to the correlation matrix in Figure 4, features such as “Delay_from_due_date” and “Num_of_Delayed_Payment” appear to be positively correlated, suggesting that individuals with more delayed payments also have a longer delay in making payments; “Credit_Utilization_Ratio” and “Monthly_Balance” show a negative correlation, meaning that as credit utilization increases, monthly balance tends to decrease. “Outstanding_Debt” and “Total_EMI_per_month” exhibit a positive correlation, indicating that higher outstanding debt is associated with higher monthly EMI payments. Table 2 describes the data features and a short description of each feature. There are 20 features as a dependent variable that contain continuous values, and one independent variable labeled “Credit_score” that contains three categorical values.

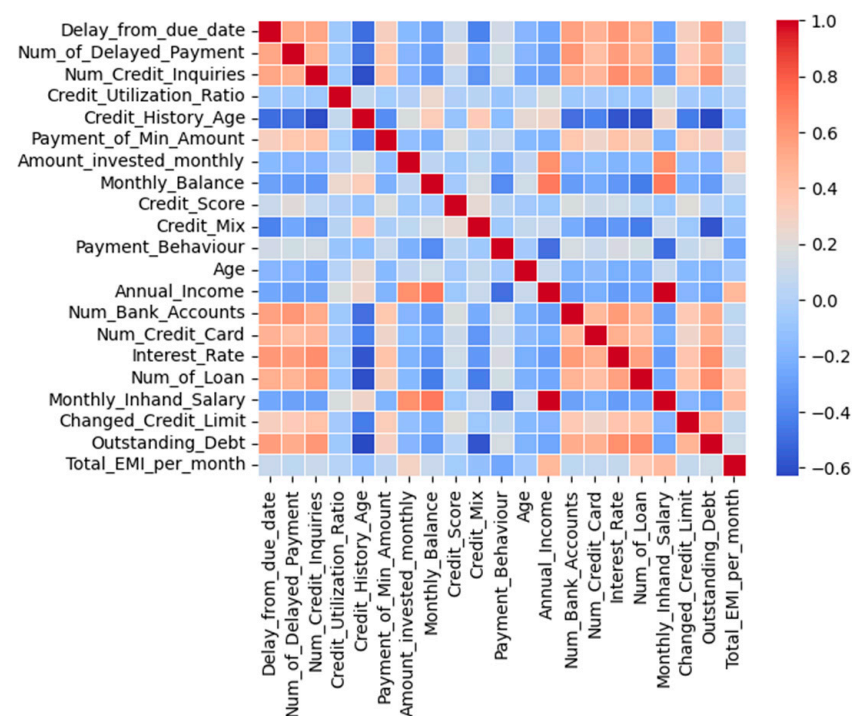


Figure 4. Features of the correlation matrix.

Table 2. Dataset features [47].

No	Features	Description
1	Age (5 cat, 5 cont)	Age of customer
2	Annual_Income (4 cat, 58 cont)	Annual income of the customer
3	Monthly_Inhand_Salary	Monthly base salary of customer
4	Num_Bank_Accounts (3 cat, 15 cont)	Number of bank accounts that the customer holds
5	Num_Credit_Card	Number of other credit cards held by a customer
6	Interest_Rate	Interest rate on credit card
7	Num_of_Loan	Number of loans taken from the bank
8	Delay_from_due_date	Average number of days delayed from the payment date
9	Num_of_Delayed_Payment	Average number of payments delayed by a customer
10	Changed_Credit_Limit	Percentage change in credit card limit
11	Num_Credit_Inquiries	Number of credit card inquiries
12	Credit_Mix	Classification of the mix of credits
13	Outstanding_Debt	Remaining debt to be paid
14	Credit_Utilization_Ratio	Utilization ratio of credit card
15	Credit_History_Age	Age of credit history of the customer
16	Payment_of_Min_Amount	The minimum amount was paid by the customer
17	Payment_Behavior	High or low value that customer spent
18	Total_EMI_per_month	Monthly EMI payments
19	Amount_invested_monthly	Monthly amount invested by the customer
20	Monthly_Balance	Monthly balance amount of the customer
21	Credit_Score	Target label (Poor, Standard, Good)

4.2. Experimental Settings

The experiments in this research were conducted in Python using the Google Collab environment. The configurations of four Generative Adversarial Network (GAN) methods used in synthetic data generation, each associated with a specific package and parameter settings provided in Table 3. The value of the hyperparameter was chosen based on the best performance of GAN by conducting several experiments on different values. This study used various types of machine learning to conduct a multi-class classification after GAN oversampling sampling methods. An overview of various machine learning algorithms, their corresponding packages, and the parameter configurations is shown in Table 4.

Table 3. Packages and hyperparameter settings in experiments.

GAN Method	Package	Parameter
CTGAN	CTGAN	Epoch = 100
CopulaGAN	CopulaGANSynthesizer	Epoch = 100
DraGAN	DRAGAN	Epoch = 100, batch size = 500, epochs = 100, learning rate = $[5 \times 10^4, 3 \times 10^3]$
WGAN-GP	WGANGP	Epoch = 100, batch size = 500, epochs = 100, learning rate = $[5 \times 10^4, 3 \times 10^3]$

Table 4. Packages, functions, and parameters for machine learning method.

Algorithm	Package	Parameter
KNN	KNeighborsClassifier	n_neighbors = 5
DT	DecisionTreeClassifier	default
LR	LogisticRegression	max_iter = 1000
XGB	XGBClassifier	default
RF	RandomForestClassifier	n_estimators = 100, random_state = 42
LGB	LGBMClassifier	n_estimators = 500, learning_rate = 0.1, max_depth = 6

The experimental procedure of this study divided the dataset into two subsets: 75% of the data were allocated to the training set, while the remaining 25% were reserved for testing. This training set was used to fit the models and enable them to learn underlying patterns within the data, while the test set provided an independent evaluation to measure model performance and assess generalizability on unseen data.

4.3. Results and Discussion

Based on our experimental results, each GAN-based sampling method improves performance across multiple classes and classifiers compared to using no sampling method (NONE). This improvement indicates that GAN can effectively balance multi-class datasets and enhance classifier performance, as shown in Table 5. The answer to RQ1, “Are GAN-based oversampling methods improving the performance of multi-class credit score classification?” based on the experimental results, the GAN method improves the performance of multi-class on minority class data but does not improve in majority class data; the worst case, the results of majority class data were decreasing.

Table 5. Experimental results comparison on F1-score.

Alg/Class	NONE	CTGAN	CopulaGAN	DraGAN	WGAN-GP
DT/					
0	0.681	0.875	0.785	0.882	0.928
1	0.732	0.827	0.754	0.810	0.735
2	0.763	0.757	0.709	0.751	0.763
KNN/					
0	0.646	0.760	0.760	0.842	0.920
1	0.761	0.744	0.748	0.792	0.764
2	0.768	0.742	0.729	0.748	0.768
LR/					
0	0.337	0.714	0.617	0.741	0.843
1	0.538	0.725	0.671	0.711	0.515
2	0.689	0.461	0.503	0.585	0.640
XGB/					
0	0.673	0.857	0.828	0.870	0.909
1	0.758	0.857	0.820	0.831	0.753
2	0.838	0.773	0.757	0.775	0.784
RF/					
0	0.735	0.901	0.849	0.904	0.936
1	0.804	0.891	0.839	0.867	0.806
2	0.816	0.813	0.797	0.813	0.814
LGB/					
0	0.736	0.905	0.875	0.906	0.938
1	0.776	0.875	0.852	0.849	0.783
2	0.800	0.798	0.785	0.798	0.802

WGAN-GP yields the most consistent and highest performance across nearly all classifiers and classes, highlighting its strength as a GAN-based sampling method for enhancing class-specific accuracy in multi-class classification. DraGAN performs comparably well to WGAN-GP across most classifiers, suggesting it as a viable alternative in cases where WGAN-GP may not be optimal. CTGAN and CopulaGAN, although these methods improve performance over no sampling, other methods generally lag WGAN-GP and DraGAN in boosting classifier accuracy. WGAN-GP and DraGAN not only improve overall accuracy but also achieve balanced gains across classes, which is particularly valuable for multi-class classification tasks, where class imbalance can degrade model performance. However, the performance of models in class 2 decreased from NONE sampling, and some methods achieved the same score as NONE sampling.

Table 6 shows the mean rank score on the F1-score. WGAN-GP consistently achieves the lowest mean rank across classifiers, indicating that it generally outperforms other methods in enhancing classifier performance. Table 7 shows the overall accuracy results. WGAN-GP yields the highest overall accuracy across almost all classifiers. WGAN-GP with RF achieved the highest accuracy result, indicating the combination of the methods.

Table 6. Mean rank score on F1-score.

Alg	NONE	CTGAN	CopulaGAN	DraGAN	WGAN-GP
DT	3.83	2.33	4.00	2.67	2.17
KNN	3.17	4.17	4.17	2.00	1.50
LR	4.00	2.50	3.67	2.83	2.00
XGB	3.33	3.00	3.67	2.33	2.67
RF	3.67	2.50	4.00	2.50	2.33
LGB	3.33	2.67	4.00	2.33	2.67

Table 7. Overall accuracy results.

Alg	NONE	CTGAN	CopulaGAN	DraGAN	WGAN-GP
DT	0.739	0.819	0.748	0.817	0.838
KNN	0.734	0.748	0.745	0.796	0.841
LR	0.604	0.642	0.599	0.671	0.718
XGB	0.831	0.831	0.803	0.829	0.843
RF	0.797	0.869	0.829	0.863	0.873
LGB	0.781	0.859	0.837	0.854	0.865

To answer RQ2, “How effective are GAN-based oversampling methods on credit score classification performance?”. As visualized in Figure 5, the GAN method applied to each machine learning algorithm was improving accuracy. However, CTGAN and CopulaGAN did not significantly improve the accuracy performance, while DraGAN and WGAN-GP consistently performed well across all algorithms, frequently achieving the highest or near-highest accuracy scores. This indicates that WGAN-GP is effective in generating realistic synthetic data that enhances machine learning performance.

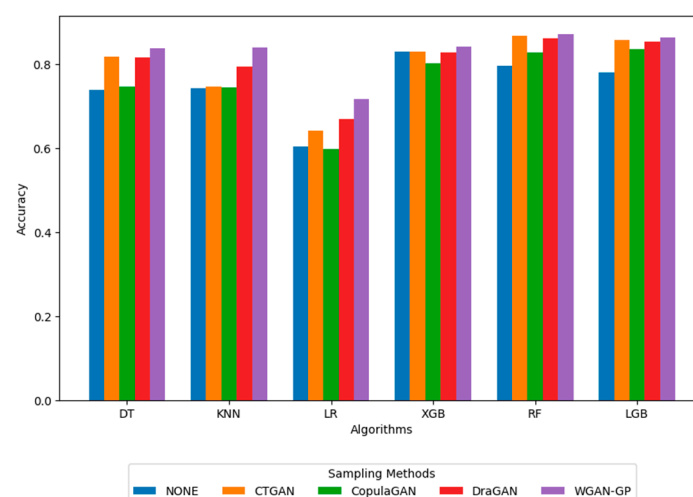


Figure 5. Accuracy comparison across sampling methods and machine learning algorithms.

To answer RQ3, “What is the most effective combination of GAN-based oversampling methods and machine learning algorithms for credit score classification?”. As shown in Table 7, WGAN-GP + RF achieved the best performance of accuracy with 0.873 and the best

performance of the F1-score on class 0: 0.936 and class 1: 0.806; unfortunately, class 2 did not perform well on the F1-score.

Table 8 shows the statistical test based on the mean rank score using the Tukey HSD post hoc test. The result shows that p -values for all comparisons are very low (≤ 0.0018), meaning WGAN-GP results differences are statistically significant. The confidence intervals (CIs) further confirm that the true mean differences are likely positive, reinforcing the reliability of the experiment results.

Table 8. Tukey HSD post hoc test.

Group 1	Group 2	Mean Diff	p -Value	Lower CI	Upper CI	Significant
CTGAN	WGANP	1.0567	0.0018	0.3399	1.7735	TRUE
CopulaGAN	WGANP	1.475	0	0.7582	2.1918	TRUE
DraGAN	WGANP	1.695	0	0.9782	2.4118	TRUE

Figure 6 illustrates how closely the synthetic data generated by WGAN-GP matches the real data in terms of feature distributions. In the plots, the density curves of real data (in green) and synthetic data (in red) are overlaid for each variable. Although WGAN-GP shows effectiveness in replicating the general patterns of many features, there are still discrepancies for certain variables, particularly where the real data distributions have more complexity or irregularity.

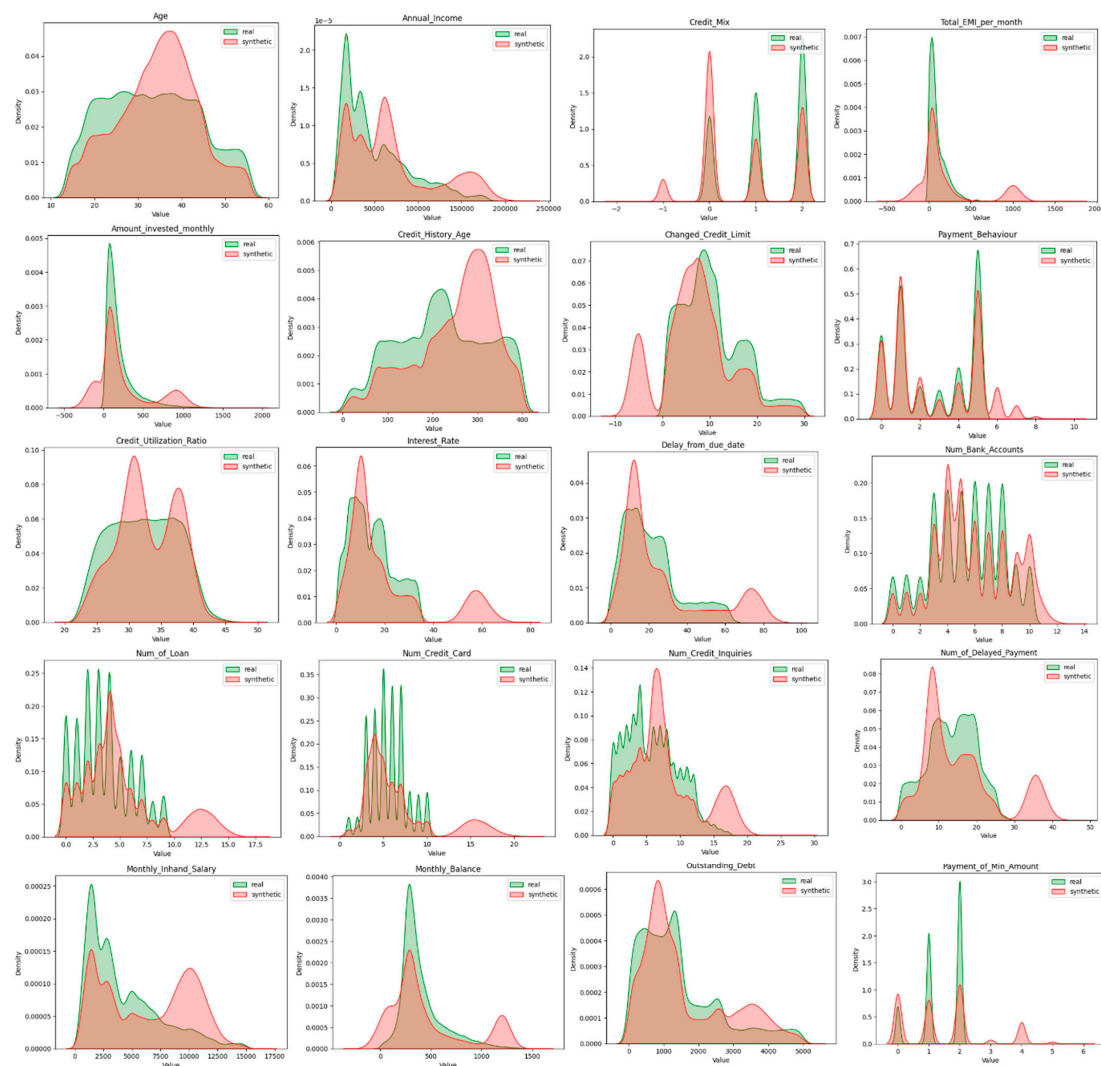


Figure 6. WGAN-GP synthetic data vs. real data distribution.

5. Conclusions

Credit score is a crucial decision-making process; wrong decisions could lead to a loss for financial institutions. Classification algorithms can help automate credit score prediction; but commonly, there is an imbalanced problem with credit score data. This study presents one of the solutions to the imbalanced multi-class problem of credit score classification data. GAN-based oversampling methods were conducted and applied to several machine learning algorithms with one public dataset.

The result of this study shows the most effective GAN-based oversampling method is WGAN-GP compared to another GAN-based method. Furthermore, the results of oversampling and real data merge to be applied with several machine learning models. Based on the experimental results, WGAN-GP and RF achieved the best performance according to the F1-score and accuracy metrics. The limitation of this study is the classification of the majority of data experiencing performance degradation. In future work, there are several possible solutions to this problem, such as a hybrid sampling method combining GAN-based oversampling with undersampling to balance the learning process between majority and minority classes. Cost-sensitive learning assigning more cost to the weak performer class could make the classifier more robust. Feature selection technique choosing the best feature that correlates to credit score decision.

Another aspect that could be further explored is the scalability of GAN-based methods when applied to larger datasets. As dataset size increases, challenges such as higher computational costs, mode collapse, and training instability may arise, potentially impacting the overall performance of GAN-generated synthetic data. Future research could investigate optimization strategies, such as improved training techniques or model architecture, to enhance the scalability and efficiency of GAN-based oversampling in large-scale credit score applications.

Author Contributions: Conceptualization, I.N.M.A.; methodology, P.-C.L. and P.W.; software, I.N.M.A.; validation, I.N.M.A. and P.-C.L.; formal analysis, I.N.M.A.; investigation, P.W.; resources, I.N.M.A.; data curation, I.N.M.A.; writing—original draft preparation, I.N.M.A.; writing—review and editing, P.-C.L. and P.W.; visualization, P.-C.L.; supervision, P.W.; project administration, P.-C.L.; funding acquisition, P.-C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science and Technology Council (NSTC), Taiwan, under grant number NSTC 113-2221-E-035-072.

Data Availability Statement: Data and code can be accessed on GitHub: <https://github.com/mahayasa/gan-multi-class-credit-score> (accessed on 10 November 2024).

Acknowledgments: The authors would also like to extend their appreciation to the National Science and Technology Council (NSTC) for the International Internship Pilot Program (IIPP) conducted in Rebecca Lab, Feng Chia University, Taiwan.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

References

1. Tripathi, D.; Edla, D.R.; Kuppili, V.; Bablani, A. Evolutionary Extreme Learning Machine with novel activation function for credit scoring. *Eng. Appl. Artif. Intell.* **2020**, *96*, 103980. [CrossRef]
2. Paleologo, G.; Elisseeff, A.; Antonini, G. Subagging for credit scoring models. *Eur. J. Oper. Res.* **2010**, *201*, 490–499. [CrossRef]
3. Van Sang, H.; Nam, N.H.; Nhan, N.D. A novel credit scoring prediction model based on feature selection approach and parallel random forest. *Indian J. Sci. Technol.* **2016**, *9*, 1–6. [CrossRef]
4. Junior, L.M.; Nardini, F.M.; Renso, C.; Trani, R.; Macedo, J.A. A novel approach to define the local region of dynamic selection techniques in imbalanced credit scoring problems. *Expert Syst. Appl.* **2020**, *152*, 113351. [CrossRef]

5. Moscato, V.; Picariello, A.; Sperlí, G. A benchmark of machine learning approaches for credit score prediction. *Expert Syst. Appl.* **2020**, *165*, 113986. [[CrossRef](#)]
6. Gicić, A.; Subasi, A. Credit scoring for a microcredit data set using the synthetic minority oversampling technique and ensemble classifiers. *Expert Syst.* **2018**, *36*, e12363. [[CrossRef](#)]
7. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *arXiv* **2014**, arXiv:abs/1406.2661. [[CrossRef](#)]
8. Dong, Y.; Xiao, H.; Dong, Y. SA-CGAN: An oversampling method based on single attribute guided conditional GAN for multi-class imbalanced learning. *Neurocomputing* **2022**, *472*, 326–337. [[CrossRef](#)]
9. Huang, C.-L.; Chen, M.-C.; Wang, C.-J. Credit Scoring with a Data Mining Approach Based on Support Vector Machines. *Expert Syst. Appl.* **2007**, *33*, 847–856. [[CrossRef](#)]
10. Mohammadi, N.; Zangeneh, M. Customer Credit Risk Assessment using Artificial Neural Networks. *Int. J. Inf. Technol. Comput. Sci.* **2016**, *8*, 58–66. [[CrossRef](#)]
11. Imtiaz, S.; Brimicombe, A.J. A Better Comparison Summary of Credit Scoring Classification. *Int. J. Adv. Comput. Sci.* **2017**, *8*, 1–4. [[CrossRef](#)]
12. Tripathi, D.; Edla, D.R.; Kuppili, V.; Dharavath, R. Binary BAT Algorithm and RBFN Based Hybrid Credit Scoring Model. *Multimed. Tools Appl.* **2020**, *79*, 31889–31912. [[CrossRef](#)]
13. Kuppili, V.; Tripathi, D.; Edla, D.R. Credit score classification using spiking extreme learning machine. *Comput. Intell.* **2019**, *36*, 402–426. [[CrossRef](#)]
14. Parvin, A.S.; Saleena, B. An Ensemble Classifier Model to Predict Credit Scoring—Comparative Analysis. In Proceedings of the 2020 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS), Chennai, India, 14–16 December 2020; pp. 27–30.
15. Veeramanikandan, V.; Jeyakarthic, M. Parameter Tuned Deep Learning Model for Credit Risk Assessment and Scoring Applications. *Recent Adv. Comput. Sci. Commun.* **2020**, *14*, 2958–2968. [[CrossRef](#)]
16. Ahmed, F.; Shamsuddin, R. A Comparative Study of Credit Card Fraud Detection Using the Combination of Machine Learning Techniques with Data Imbalance Solution. In Proceedings of the 2021 2nd International Conference on Computing and Data Science (CDS), Stanford, CA, USA, 28–29 January 2021; pp. 112–118.
17. Maurya, A.; Gaur, S. A Decision Tree Classifier Based Ensemble Approach to Credit Score Classification. In Proceedings of the 2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India, 3–4 November 2023; pp. 620–624.
18. Ziemba, P.; Becker, J.; Becker, A.; Radomska-Zalas, A.; Pawluk, M.; Wierzba, D. Credit decision support based on real set of cash loans using integrated machine learning algorithms. *Electronics* **2021**, *10*, 2099. [[CrossRef](#)]
19. Sotiropoulos, D.N.; Koronakos, G.; Solanakis, S.V. Evolving Transparent Credit Risk Models: A Symbolic Regression Approach Using Genetic Programming. *Electronics* **2024**, *13*, 4324. [[CrossRef](#)]
20. Hayashi, Y. Emerging Trends in Deep Learning for Credit Scoring: A Review. *Electronics* **2022**, *11*, 3181. [[CrossRef](#)]
21. Sun, J.; Li, J.; Fujita, H. Multi-class imbalanced enterprise credit evaluation based on asymmetric bagging combined with light gradient boosting machine. *Appl. Soft Comput.* **2022**, *130*, 109637. [[CrossRef](#)]
22. Xiao, H.; Xiao, Z.; Wang, Y. Ensemble classification based on supervised clustering for credit scoring. *Appl. Soft Comput.* **2016**, *43*, 73–86. [[CrossRef](#)]
23. Singh, I.; Kumar, N.; Srinivasa, K.G.; Maini, S.; Ahuja, U.; Jain, S. A multi-level classification and modified PSO clustering based ensemble approach for credit scoring. *Appl. Soft Comput.* **2021**, *111*, 107687. [[CrossRef](#)]
24. Tanha, J.; Abdi, Y.; Samadi, N.; Razzaghi, N.; Asadpour, M. Boosting methods for multi-class imbalanced data classification: An experimental review. *J. Big Data* **2020**, *7*, 1–47. [[CrossRef](#)]
25. Dong, X.; Yu, Z.; Cao, W.; Shi, Y.; Ma, Q. A survey on ensemble learning. *Front. Comput. Sci.* **2019**, *14*, 241–258. [[CrossRef](#)]
26. Arora, N.; Kaur, P.D. A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. *Appl. Soft Comput.* **2019**, *86*, 105936. [[CrossRef](#)]
27. Nekooimehr, I.; Lai-Yuen, S.K. Adaptive Semi-Unsupervised Weighted Oversampling (A-Suwo) for Imbalanced Datasets. *Expert Syst. Appl.* **2016**, *46*, 405–416. [[CrossRef](#)]
28. Douzas, G.; Bacao, F.; Last, F. Improving Imbalanced Learning Through a Heuristic Oversampling Method Based on K-Means And SMOTE. *Inf. Sci.* **2018**, *465*, 1–20. [[CrossRef](#)]
29. Gangwar, A.K.; Ravi, V. WiP: Generative Adversarial Network for Oversampling Data in Credit Card Fraud Detection. In Proceedings of the 2019 International Conference, Hyderabad, India, 16–20 December 2019.
30. Salazar, A.; Vergara, L.; Safont, G. New Applications of An Oversampling Method Based on Generative Adversarial Networks. In Proceedings of the 2020 International Conference on Computational Science, Las Vegas, NV, USA, 16–18 December 2020.
31. Engelmann, J.; Lessmann, S. Conditional Wasserstein GAN-based Oversampling Of Tabular Data For Imbalanced Learning. *arXiv* **2020**, arXiv:2006.00842. [[CrossRef](#)]

32. Li, H.; Fan, R.; Shi, Q.; Du, Z. Class Imbalanced Fault Diagnosis Via Combining K-Means Clustering Algorithm with Generative Adversarial Networks. *J. Adv. Comput. Intell. Inform.* **2021**, *25*, 346–355. [\[CrossRef\]](#)
33. Dablain, D.; Krawczyk, B.; Chawla, N.V. DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 6390–6404. [\[CrossRef\]](#)
34. Wang, Y.; Zhang, X.; Liu, H.; Li, B.; Yu, J.; Liu, K.; Qin, L. Intelligent Identification of the Line-Transformer Relationship in Distribution Networks Based on GAN Processing Unbalanced Data. *Sustainability* **2022**, *14*, 8611. [\[CrossRef\]](#)
35. García-Vicente, C.; Chushig-Muzo, D.; Mora-Jiménez, I.; Fabelo, H.; Gram, I.T.; Løchen, M.-L.; Granja, C.; Soguero-Ruiz, C. Evaluation of Synthetic Categorical Data Generation Techniques for Predicting Cardiovascular Diseases and Post-Hoc Interpretability of the Risk Factors. *Appl. Sci.* **2023**, *13*, 4119. [\[CrossRef\]](#)
36. Liu, Y.; Wang, S.; Sui, H.; Zhu, L. An Ensemble Learning Method with GAN-Based Sampling and Consistency Check for Anomaly Detection of Imbalanced Data Streams with Concept Drift. *PLoS ONE* **2024**, *19*, e0292140. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Cai, Z.; Wang, X.; Zhou, M.; Xu, J.; Jing, L. Supervised Class Distribution Learning For GANs-Based Imbalanced Classification. In Proceedings of the IEEE International Conference on Data Mining (ICDM), Beijing, China, 8–11 November 2019.
38. Li, X.; Luo, M.; Ji, S.; Zhang, L.; Lu, M. Evaluating Generative Adversarial Networks Based Image-Level Domain Transfer for Multi-Source Remote Sensing Image Segmentation and Object Detection. *Int. J. Remote Sens.* **2020**, *41*, 7343–7367. [\[CrossRef\]](#)
39. Li, D.; Kotani, D.; Okabe, Y. Improving Attack Detection Performance in NIDS Using GAN. In Proceedings of the 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC), Madrid, Spain, 13–17 July 2020; pp. 817–825.
40. Al-Najjar, H.A.H.; Pradhan, B.; Sarkar, R.; Beydoun, G.; Alamri, A.M. A New Integrated Approach for Landslide Data Bal-ancing and Spatial Prediction Based on Generative Adversarial Networks (GAN). *Remote Sens.* **2021**, *13*, 4011. [\[CrossRef\]](#)
41. Abedi, M.; Hempel, L.; Sadeghi, S.; Kirsten, T. GAN-Based Approaches for Generating Structured Data in The Medical Do-main. *Appl. Sci.* **2022**, *12*, 7075. [\[CrossRef\]](#)
42. Duan, S.; Liu, C.; Han, P.; Jin, X.; Zhang, X.; He, T.; Pan, H.; Xiang, X. HT-Fed-GAN: Federated Generative Model for Decentralized Tabular Data Synthesis. *Entropy* **2022**, *25*, 88. [\[CrossRef\]](#) [\[PubMed\]](#)
43. Serte, S.; Dirik, M.A.; Al-Turjman, F. Deep Learning Models for COVID-19 Detection. *Sustainability* **2022**, *14*, 5820. [\[CrossRef\]](#)
44. Rafiei, A.; Rad, M.G.; Sikora, A.; Kamaleswaran, R. Improving Irregular Temporal Modeling By Integrating Synthetic Data to The Electronic Medical Record Using Conditional GANs: A Case Study of Fluid Overload Prediction in The Intensive Care Unit. *medRxiv* **2023**. [\[CrossRef\]](#)
45. Kaabachi, B.; Briki, F.; Kulynych, B.; Despraz, J.; Raisaro, J.L. Tunable Privacy Risk Evaluation of Generative Adversarial Networks. *Stud. Health Technol. Inform.* **2024**.
46. Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; Veeramachaneni, K. Modeling Tabular data using Conditional GAN. *arXiv* **2019**, arXiv:1907.00503.
47. Pathare, A.; Mangrulkar, R.; Suvarna, K.; Parekh, A.; Thakur, G.; Gawade, A. Comparison of tabular synthetic data generation techniques using propensity and cluster log metric. *Int. J. Inf. Manag. Data Insights* **2023**, *3*, 100177. [\[CrossRef\]](#)
48. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A. Improved Training of Wasserstein GANs. *arXiv* **2017**, arXiv:1704.00028.
49. Nejad, F.S.; Ebadzadeh, M.M. Stock market forecasting using DRAGAN and feature matching. *Expert Syst. Appl.* **2023**, *244*, 122952. [\[CrossRef\]](#)
50. Multi-Class Classification Problem. Available online: <https://www.kaggle.com/datasets/sudhanshu2198/processed-data-credit-score> (accessed on 11 November 2024).

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.