Journal of Information Systems Engineering and Management
2025, 10(45s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Enhanced GANified-SMOTE with Latent Factor for Improved Classifier Performance on Imbalanced Datasets

Rusma Anieza Ruslan[1], Nureize Arbaiy[2], Pei-Chun Lin[3]
[1]*Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, 86400, Malaysia*
[2] *Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, 86400, Malaysia*
[3]*Department of Information Engineering and Computer Science, Feng Chia University, No. 100 Wenhwa Rd., Taichung 40724, Taiwan*

ARTICLE INFO

Received: 28 Dec 2024

Revised: 18 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

Introduction: Imbalanced datasets cause significant issues in classification tasks that might have a negative impact on the model's performance. It frequently results in minority classes having worse predictive accuracy. This leads to lower accuracy for minority classes. This issue affects model performance and risks missing crucial insights that inform decision-making.

Objectives: This study presents a novel methodology combining the Generative Adversarial Network-Based Synthetic Minority Oversampling Technique (GANified-SMOTE) with latent factor approaches to enhance classifier performance on imbalanced datasets.

Methods: We evaluate the effectiveness of this framework across various datasets, demonstrating its ability to generate high-quality synthetic samples that accurately reflect the underlying data distribution.

Results: Our experimental results show that the Enhanced GANified-SMOTE significantly improves accuracy when integrated with classifiers like Random Forest (RF). Specifically, our method achieves an outstanding accuracy of 0.999971 in the Credit Card Fraud Detection task, along with near-perfect precision and recall metrics.

Conclusions: These results underscore the potential of our approach to improve classification reliability and reduce false negatives in critical applications, addressing the limitations of traditional classification techniques in imbalanced contexts.

Keywords: class imbalance, classification tasks, GANified-SMOTE, generative adversarial network, machine learning.

INTRODUCTION

Imbalanced datasets are a collection of datasets that have a different number of minority classes and majority classes [1, 2]. Usually, the majority class is the most frequent class, and the minority class is the least frequent class. Example in real-world machine learning (ML) applications such as fraudulent transactions [3, 4]. In fraud detection, the number of fraudulent transactions is usually less than the number of non-fraudulent transactions, with the number of samples belonging to the minority class being significantly smaller compared to the majority class. Therefore, these rare fraudulent activities must be identified more accurately to avoid financial losses. If imbalances in the dataset are not considered, this can lead to adverse effects such as losses on fraudulent transactions.

However, if the dataset contains fewer samples of the minority class than the majority class, then the ML model will favor the majority class. This leads to poor performance of the minority class and inaccurate or biased predictions [5]. This happens because the model preferentially learns patterns and makes predictions for the majority class because it is more represented than the minority class. Eliminating imbalances in the dataset is crucial for accurate predictions and decisions. Therefore, there are various methods to solve this problem, including SMOTE.

SMOTE is a widely used method for addressing class imbalance in datasets by generating synthetic samples of the minority class [6, 7]. However, SMOTE can have several limitations that can hinder its effectiveness, such as overlapping samples between classes [8]. This issue arises when synthetic samples generated from existing minority class samples closely resemble the training data, leading models to memorize specific points rather than learning

generalizable patterns. As a result, while a model may perform well on the training set, its ability to generalize to unseen data can be compromised, resulting in poor performance in real-world applications. This study proposes GANified-SMOTE with Latent Factor to overcome these limitations.

GANified-SMOTE with Latent Factor is proposed in this study. GANified-SMOTE combines GAN and SMOTE to create synthetic samples using random noise and a latent factor. The existing SMOTE-GAN uses only random noise, a starting point for generating synthetic samples. It helps the generator create diverse and varied outputs. This new method uses random noise to add variation to sample production, and the latent factor identifies crucial underlying patterns that correspond to the properties of actual data. Combining these components improves minority class representation, increases the model's accuracy, and improves the development of realistic and diverse synthetic samples. This enhances overall performance and reduces the likelihood of overfitting and class overlap.

The structure of this document is as follows: Section I gives research backgrounds. Section II provides a literature review related to the research study. Section III explains the methods used in the study. Section IV discuss the results obtained in the study. Section V concludes the overall research study.

LITERATURE REVIEW

Literature reviews are presented in this section. This review synthesises existing research on imbalanced datasets, highlighting key findings, theoretical frameworks and areas where further investigation is needed.

A. Resampling

Resampling is a technique that changes or adjusts the composition of a dataset. These methods are beneficial when the original dataset is imbalanced. For example, one class or group is underrepresented compared to others. To address this class imbalance, resampling techniques can either oversample the minority class, undersample the majority class or combine both [9, 10]. The two main categories of resampling techniques are undersampling and oversampling [11, 12].

1. Oversampling

Oversampling is one of the techniques used in ML to avoid class imbalance. Oversampling increases the minority class by creating a synthetic sample nearly identical to the original sample [13, 14]. In his method, the number of events from the minority class is randomly doubled until the number is equal to or greater than the number of events in the majority class [15]. This method helps ML models to learn more effectively and without bias. No information is lost in this method [16] as it is a process of duplication rather than discarding. Various oversampling techniques are used to address the imbalance in the dataset.

Random oversampling (ROS) is a technique used to address class imbalance by duplicating samples from the minority class at random. This approach aims to balance the distribution of courses within the dataset, making it more equitable for model training. As noted by [17] and [18], ROS can help improve the performance of ML algorithms by ensuring that the minority class is adequately represented, thus enhancing the model's ability to learn from all classes effectively.

The SMOTE method addresses class imbalance by generating new samples for the minority class. It interpolates between existing minority class samples and their nearest neighbours [6]. According to [19], SMOTE enhances the representation of the minority class, allowing ML models to learn more effectively from the data. This technique not only increases the number of minority class samples but also helps in creating a more informative and diverse dataset, ultimately improving model performance.

ADASYN is a technique designed to enhance class balance by generating synthetic samples, focusing on minority samples that are more challenging for models to learn [20]. According to [21], ADASYN increases the number of synthetic samples based on the density of minority class instances, prioritizing the creation of samples in areas where the minority class is underrepresented. This targeted approach helps improve model performance by enabling better learning from complex cases, thereby addressing the challenges of class imbalance more effectively.

Journal of Information Systems Engineering and Management
2025, 10(45s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

An example of the application of the oversampling method in addressing the imbalance of the dataset is explained in Table 1. It highlights the use of techniques such as ROS and RUS. SMOTE and its variants like Borderline-SMOTE, RU-SMOTE, RU-ADASYN, and Deep Smote. The studies aim to evaluate the effectiveness of these techniques in addressing data imbalance problems. The findings indicate that ROS often performs well in accuracy and efficiency. Meanwhile, techniques like RUS and ROS significantly improve the classification performance of ANN on imbalanced datasets. DeepSMOTE is noted for its superior performance on small and imbalanced datasets. The table also suggests that the oversampling ratio can negatively impact precision. It implies a need for careful parameter tuning when applying these techniques.

Table 1. Application of Oversampling

Author	Technique	Contribution	Findings
[22]	ROS	Compare the effectiveness of ROS with more advanced oversampling	ROS performs better in terms of robust accuracy and computational efficiency
[17]	RUS ROS RURO RU-SMOTE RU-ADASYN	Influence of resampling techniques on the performance of ANN classifiers in cybersecurity	RUS and ROS techniques significantly improve the classification performance of ANN on imbalanced cybersecurity datasets
[23]	ROS SMOTE Borderline-SMOTE ADASYN Deep SMOTE	Evaluate the effectiveness of oversampling strategies that are intended to address data imbalance.	DeepSMOTE performed better on small and imbalanced datasets than any other oversampling method.
[24]	ROS SMOTE	Explores how different oversampling algorithms and imbalance ratios affect the performance of classification algorithms	The oversampling ratio has a negative impact on precision but a significant positive impact on AUC and recall rate.
[25]	SMOTE-Oversampling (SOS) ROS	Address class imbalance by increasing the number of minority class data through random and synthetic methods	SOS and ROS methods successfully increased the number of minority class samples.

2. Undersampling

In undersampling, the majority class sample is deleted and eliminated [13, 14] until the number of samples equals the minority class [26]. Undersampling can also be done by randomly removing samples from the majority class until the desired balance is reached [27]. Since the method works by reducing the size of the dataset, it can help to reduce the training time. However, this method has some notable drawbacks, such as potentially losing important information [28, 29]. This is because it involves deleting and eliminating samples. Various undersampling techniques are used to address the imbalance in the dataset.

Random undersampling (RUS) is a technique that helps balance datasets by randomly removing samples from the majority class. As noted by [30] and [17], this method effectively addresses class imbalance, which can significantly impact the performance of ML models. By reducing the number of instances in the majority class, RUS facilitates a more equitable representation of the classes in the dataset.

TomekLinks is a technique that improves dataset quality by removing samples from the majority class identified as TomekLinks. These samples are considered noisy and are often located at the boundaries between classes. According to [31], this method enhances class separation and can effectively address issues related to class imbalance, thereby improving the performance of ML models.

NearMiss is a sampling technique that aims to balance datasets by selecting samples from the majority class that are most similar to those in the minority class. This method effectively enhances the representation of the minority class

Journal of Information Systems Engineering and Management
2025, 10(45s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

while maintaining the essential characteristics of the majority class. [31] and [32] highlight that NearMiss can significantly improve model performance by fostering better class distinction and reducing the impact of class imbalance in ML applications.

ClusterC, also known as Cluster Centroids, is a technique that addresses class imbalance by dividing the majority class into clusters using k-means clustering [33]. The centroids of these clusters are then utilized to represent the majority class in the dataset, effectively balancing it with the minority class. [34] and [35] emphasize that this method not only reduces the number of samples in the majority class but also preserves the essential characteristics of the data, leading to improved performance in ML models.

Edited Nearest neighbours (ENN) is a technique that utilizes k-nearest neighbours (KNN) to improve dataset quality by removing most samples incorrectly classified. This method enhances the integrity of the dataset by focusing on the accuracy of class labels. According to [36] and [19], ENN effectively reduces noise in the majority class, leading to better classification performance in ML models by ensuring that only reliable samples are retained.

An example of the application of the undersampling method in addressing the dataset's imbalance is explained in Table 2. It summarises various undersampling techniques used in research to address class imbalance problems. The table highlights the specific methods employed, contributions and the key findings. Notably, RUS is frequently used. Studies demonstrate its effectiveness in achieving comparable performance to a balanced dataset, even with significant class imbalances. Furthermore, the research explores using RUS in conjunction with Ensemble of Classifier Chains (ECC) to improve performance in multilabel data classification. The table also introduces novel undersampling methods like RFCL and RBU. This method aims to reduce overlapping degrees and offer faster alternatives. The table underscores the diverse approaches and their impact on improving classification performance in an imbalanced dataset.

Table 2. Application of Undersampling

Author	Technique	Contribution	Findings
[37]	RUS	Demonstrates that effective classification performance on big data can be achieved with minimal alterations to the original dataset even in the presence of significant class imbalance	Maintaining a minority class ratio between 0.1% and 1.0% can yield performance comparable to using a fully balanced dataset with similar results observed when applying RUS to achieve a 50:50 class ratio.
[38]	Ensemble of Classifier Chains (ECC) with RUS	Improves ECC for imbalanced multilabel data by finding better ways to use the majority class examples without needing more computing power	Proposed approaches significantly improve performance across various evaluation metrics.
[39]	RUS	Provides a comprehensive overview of the class imbalance problem	RUS approaches can effectively mitigate this issue by improving the balance between class distribution.
[40]	Random Forest Clearing Rule (RFCL)	Proposes the RFCL, a novel undersampling method aimed at reducing the overlapping degree rather than merely resampling	RFCL outperforms seven classic and two recent under-sampling algorithms regarding F1-score and area under the curve (AUC).
[41]	Radial-Based Undersampling (RBU)	Proposes a novel Radial-Based Undersampling (RBU) algorithm, which serves as a faster alternative to the Radial-Based Oversampling (RBO) algorithm	RBU performs comparably to the original Radial-Based Oversampling (RBO) algorithm.

Journal of Information Systems Engineering and Management
2025, 10(45s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3. SMOTE

SMOTE is one of the techniques used in resampling to solve the problem of data imbalance [42]. The SMOTE approach was inspired by a technique that proved successful in recognizing handwritten characters and was applied by previous researchers in 1997 [26]. SMOTE works by identifying samples from the minority class and generating a synthetic sample that is identical to the actual sample [8].

Synthetic sample generation is based on nearest neighbour selection based on Euclidean distance [43], as shown in Figure 1. The new synthetic sample is generated by interpolating between the minority sample and its neighbours [44]. By increasing the number of samples in the minority class, SMOTE can balance the dataset to avoid bias, and the classifier can learn a better representation for both classes.

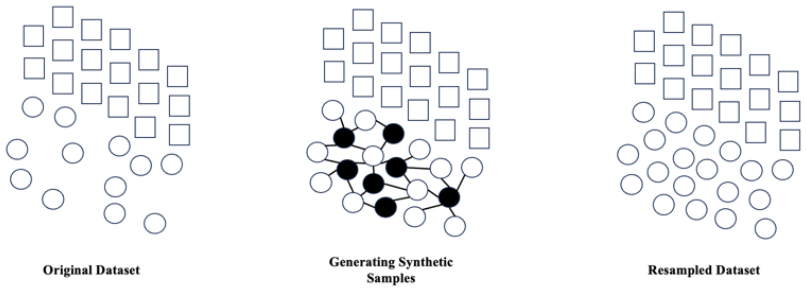

Figure 1. SMOTE Method

Various SMOTE variants are used to address the imbalance in the dataset. The SMOTE, as introduced by [26], is an advanced method for addressing class imbalance in datasets. Unlike traditional oversampling techniques that duplicate existing samples, SMOTE generates synthetic samples. This approach creates new, unique instances based on the characteristics of the minority class, thereby enhancing the representation of these classes in the training data while avoiding the pitfalls of overfitting associated with mere duplication.

Borderline-SMOTE, described by [45], is a variant of the SMOTE that specifically targets the generation of synthetic samples for minority class instances located near the decision boundary. This approach emphasizes the most critical samples (those at the edge of the minority class), enhancing the model's ability to distinguish between classes. Borderline-SMOTE aims to improve classification performance and robustness against misclassifications by focusing on these boundary instances.

SVM SMOTE, introduced by [46], is a technique that combines the principles of Support Vector Machines (SVM) with SMOTE. This method integrates SVM to identify support vectors and critical data points that define the decision boundary. By generating synthetic samples based on these support vectors, SVM SMOTE enhances the representation of the minority class while maintaining the integrity of the decision boundary. This approach not only improves the model's performance in classifying minority instances but also contributes to a more robust learning process. Table 3 discusses SMOTE's existing applications in addressing the dataset's imbalance.

Table 3. Application of SMOTE

Technique	Contribution	Findings
SMOTE [47]	Highlighting the effectiveness of SMOTE and its extensions in improving classification performance	SMOTE has been influenced in addressing class imbalance
Improved SMOTE [48]	Proposes an improved SMOTE algorithm based on the normal distribution to address the class-imbalance problem in data classification	The improved SMOTE algorithm outperforms the original SMOTE in classification accuracy.
SMOTE [49]	Introduces the application of the SMOTE repeatedly to balance multi-class datasets effectively	Iterative SMOTE processes yielded better results compared to a single application of SMOTE.

Journal of Information Systems Engineering and Management
2025, 10(45s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

SMOTE ROS [50]	Proposes an efficient Fraud Detection Framework specifically designed for credit card transactions with imbalanced data	Indicate that the proposed fraud detection framework effectively addresses the imbalanced dataset problem in credit card transactions.
SMOTE [9]	Examines the impact of class imbalance on classification performance and employs a SMOTE	Both the resampling method and normalization techniques positively influence the performance of classification models.

B. Generative Adversarial Network

GANs are generative ML techniques proposed a few years ago [51]. Data augmentation with GAN is a technique that utilizes the capabilities of GAN models to create synthetic data patterns that can be used to augment an existing dataset. The two main components of generative models or GANs are the discriminator and generator networks, as shown in Figure. 2.

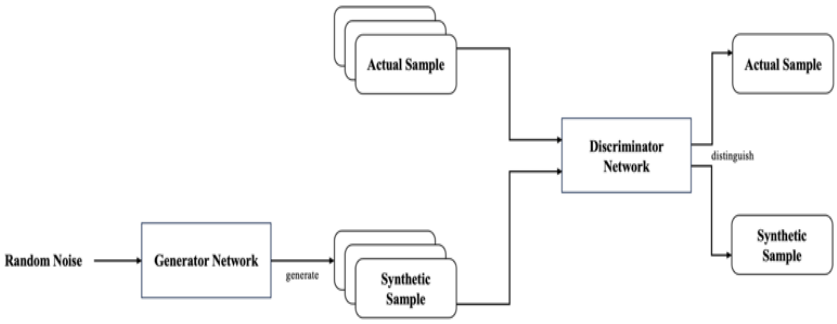

Figure 2. Basic Structure of GAN

The generator network generates more realistic synthetic samples that resemble the real sample based on random noise to deceive the discriminator network [52]. At the same time, the discriminator networks try to improve their ability to distinguish between real and synthetic samples [53]. This process is repeated until the discriminator network can no longer distinguish between them. The real sample refers to the sample obtained from the original dataset [54]. In contrast, the synthetic sample refers to the sample that the GAN generates based on the sample learned from the real sample [54].

Since GAN was introduced in 2014 by researchers [51], many new variants of GAN have been proposed. Wasserstein Generative Adversarial Networks (WGAN) present an innovative approach to training generative models, offering an alternative to traditional GAN methods. Introduced by [55], WGAN utilizes the Wasserstein distance to measure convergence between the generated and real data distributions.

Conditional Generative Adversarial Nets (cGANs) introduce a novel framework for training generative models that allows data generation to be conditioned on specific inputs. As proposed by [56], cGANs extend the traditional GAN architecture by incorporating additional information, such as class labels or other data features, into the generator and discriminator models.

Duo-GAN is a framework designed to tackle the challenges associated with heavily imbalanced datasets by generating synthetic data. Duo-GAN [57], leverages the strengths of generative adversarial networks to create balanced datasets that enhance model training.

Synthetic Data Generation GAN (SDG-GAN) is a novel system introduced to generate synthetic data specifically for training supervised classifiers. As outlined by [58], SDG-GAN employs the principles of GANs to create realistic synthetic datasets that help mitigate the issues related to limited or imbalanced training data. Table 4 discusses existing applications of GANs. It shows how each technique aims to generalize the class balance by increasing minority classes in small and imbalanced datasets.

Journal of Information Systems Engineering and Management
2025, 10(45s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 4. Application of GANs

Technique	Contribution	Findings
GAN [59]	Presents a novel data augmentation method using a GAN to address the challenge of imbalanced spectral data in the classification of material characteristics	The proposed method enhances classifier performance compared to existing data augmentation methods.
GAN Neural Transfer Style [60]	Highlights the significant role of data augmentation in improving the generalization ability of DL models	Neural Augmentation techniques outperform traditional methods in specific tasks.
GAN [61]	Presents a DL methodology for binary classification of network traffic by representing network flows as 2D images, leveraging a GAN and a CNN	The proposed methodology improves predictive accuracy compared to existing intrusion detection architectures.
GAN [62]	Discusses various techniques used to improve the performance and stability of GANs, for instance	GANs achieve this by deriving back-propagation signals using a method with a pair of networks.
CNN GAN [45]	Implement a code visualization method and utilize GAN to generate more samples of malicious code variants.	CNNs plus the GAN model can achieve a higher classification accuracy than related work.

C. Hybrid Approach

The hybrid approach techniques combine the strengths of SMOTE and GANs to address class imbalance in ML datasets. By integrating the ability of SMOTE to create synthetic samples with the generative capabilities of GANS, this hybrid approach not only increases minority class samples but also ensures that they are more representative of the underlying data distribution.

1. SMOTified-GAN

According to [63], SMOTified-GAN is an innovative technique that combines the strengths of SMOTE and GANs, as shown in Figure 3, to enhance the generation of synthetic samples for the minority class. This two-phase approach first utilizes SMOTE to create initial synthetic samples, which are then refined using a GAN to produce more realistic data distributions.

This method significantly improves the quality and diversity of minority class samples, achieving performance gains of up to 9% in F1-score compared to other algorithms. SMOTified-GAN is particularly effective in scenarios where GANs alone may struggle due to limited minority class data. Its successful application across benchmark datasets highlights its versatility and effectiveness in improving classification performance in imbalanced datasets.

Journal of Information Systems Engineering and Management
2025, 10(45s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

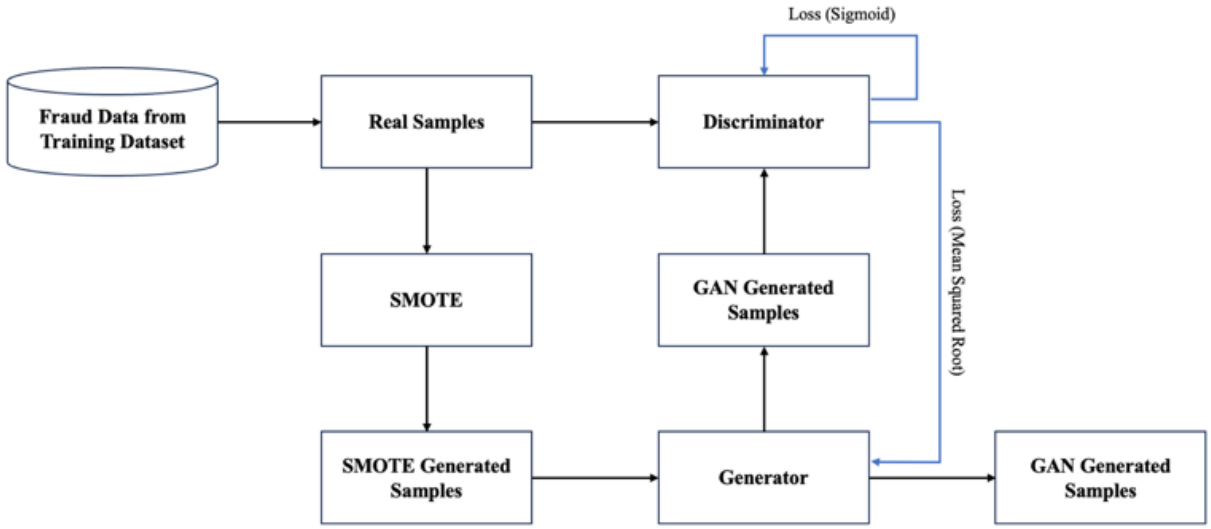

Figure 3. SMOTified-GAN Architecture

2. GANified-SMOTE

According to [7], GANified-SMOTE is a method that integrates GANS into the SMOTE process, as shown in Figure 4, to enhance the generation of synthetic samples. In this approach, GANs are employed to create additional synthetic samples that complement those generated by SMOTE, ultimately aiming to produce a more balanced dataset.

This technique has demonstrated strong performance across various datasets, particularly in financial fraud detection, where it effectively mitigates bias towards the majority class. GANified-SMOTE excels at handling varying amounts of generated samples, making it adaptable to different scenarios. Its ability to improve the accuracy of minority class predictions is especially valuable in domains like financial fraud detection, where class imbalance poses significant challenges.

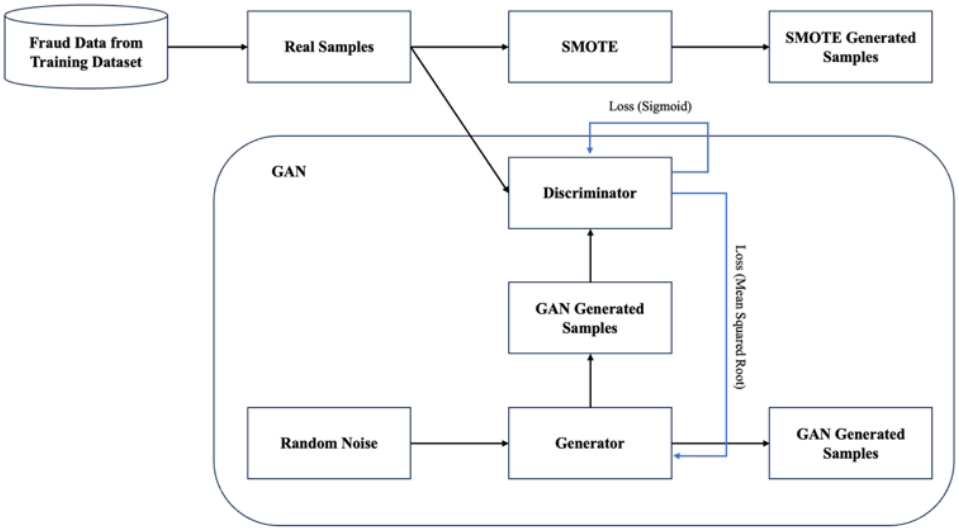

Figure 4. GANified-SMOTE Architecture

METHODS

The methods used are presented in this section. We outline the research design, data collection techniques and analytical strategies employed to investigate our research questions.

Journal of Information Systems Engineering and Management
2025, 10(45s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A. Data Preparation

Data preparation is a crucial phase in the process of data analysis. This process involves converting the raw data into a format suitable for analysis [64]. This process involves two main tasks: data collection and data preprocessing.

1. Data Collection

Data collection involves gathering data from the Kaggle platform. Kaggle provides an extensive collection of datasets from various fields that allow users to explore, analyze and build models while engaging with a community of data scientists. Three publicly available datasets were used in this study, which are listed:

- Credit Card Fraud Detection Dataset [65]
- Titanic Dataset [66]
- Pima Indians Diabetes Database [67]

2. Data Preprocessing

Data preprocessing is essential to transforming raw data into a clean and usable format before analysis or modelling [68]. This phase involves several steps:

- Feature and Target Separation: Involves splitting the dataset into two distinct components, the features and the target variables
- Feature Scaling: Involves adjusting the range and distribution of the input features to ensure that they are on a similar scale
- Class Separation: Identifying the different classes in the target variable and splitting the dataset into various subsets based on these classes
- Data Partitioning: Involves the separation of subsets for training and test sets

B. Proposed Research Model

The study aims to implement a hybrid approach that combines GAN with SMOTE, which is called GANified-SMOTE with Latent Factor. Classification tasks are performed by generating realistic synthetic samples for the minority class. By using GANs to create high-quality synthetic data and SMOTE to enhance this data further, the study aims to improve the representation of underrepresented classes.

Figure 5 illustrates the integration of GAN with SMOTE to generate synthetic samples. It starts with actual samples extracted from a training dataset, which the discriminator evaluates to distinguish between actual and generated data. The generator creates synthetic samples using random noise with latent factors to replicate the characteristics of real samples. The discriminator uses a sigmoid loss function to measure its ability to classify the samples accurately. In contrast, the generator uses a mean squared error loss to evaluate its success in deceiving the discriminator. After generating these samples, SMOTE is applied to the GAN-generated data to increase the diversity of the synthetic samples and reduce reliance on dominant patterns, helping to avoid overfitting. The results of SMOTE are then combined with the original training dataset to create a balanced and comprehensive dataset for training.

Journal of Information Systems Engineering and Management
2025, 10(45s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

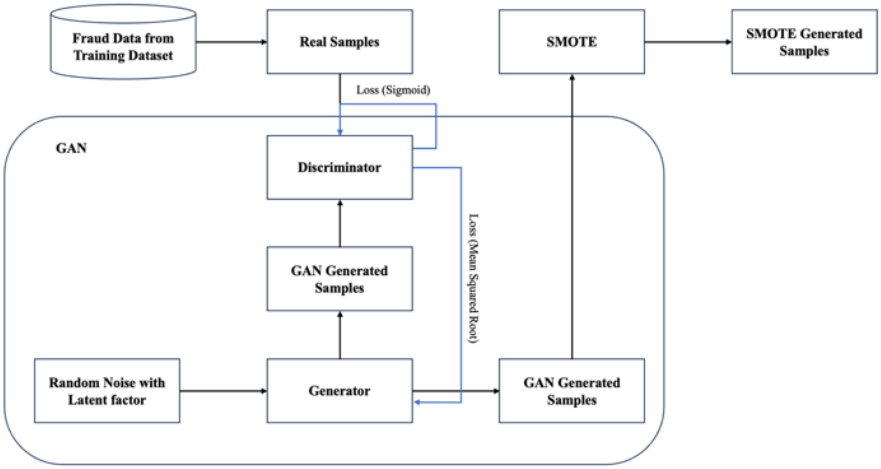

Figure 5. GANified-SMOTE with Latent Factor

C. Model Performance and Evaluation

It systematically evaluates the proposed research model's effectiveness and accuracy after its implementation. This phase includes various metrics and techniques to determine how well the model performs in classifying the data, especially identifying minority classes in imbalanced datasets.

1. Experimental setup

The experiment uses Python Jupyter Notebook version 7.2.0, which provides a flexible and interactive environment for performing data analysis and developing ML models.

2. Parameter Settings

Parameter setting is a crucial aspect in the development of ML models, as it involves the selection of values for various hyperparameters that significantly affect the model's performance, stability during training and speed of convergence. In this study, the parameters used are:

- Epochs: 10
- Generator Learning Rate: 0.0002
- Discriminator Learning Rate: 0.0001
- Optimizer: Adam

3. Performance Analysis

Based on their training and assessment results, the performance analysis evaluates three classifiers (Random Forest, Gradient Boosting and Decision Tree). Each classifier is trained on a balanced dataset and tested on a corresponding set of performance metrics such as accuracy, recall, precision and F1-score as in (1) until (4).

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

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

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

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

RESULTS

Experimental results are presented in this section. It shows how different approaches to dealing with imbalanced datasets and preventing overfitting affected model performance.

Journal of Information Systems Engineering and Management
2025, 10(45s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 5. Classification Results with SMOTE for Imbalanced Datasets

Dataset	Classifier	Accuracy	Precision	Recall	F1-Score	Time (s)
Credit Card Fraud Detection	RF	**0.999454**	0.890244	0.768421	0.824859	236.60
	GB	0.993973	0.195062	0.831579	0.316000	470.11
	DT	0.998132	0.463576	0.736842	0.569106	31.94
Titanic	RF	**0.488095**	0.357143	0.483871	0.410959	0.07
	GB	0.440476	0.309524	0.419355	0.356165	0.04
	DT	**0.488095**	0.342105	0.419355	0.376811	0.00
Diabetes	RF	0.746753	0.627119	0.685185	0.654867	0.09
	GB	**0.798701**	0.682540	0.796296	0.735043	0.13
	DT	0.733766	0.658537	0.500000	0.568421	0.00

Table 5 presents the performance metrics of the classifiers (RF, GB and DT) across three different datasets. For the Credit Card Fraud Detection datasets, RF achieves the highest accuracy at 0.999454, indicating excellent model performance. In contrast, the Titanic dataset shows much lower accuracy across all classifiers with both RF and DT at 0.488095, suggesting that these models struggle to predict outcomes effectively for this dataset. The Diabetes dataset exhibits moderate accuracy, with GB performing the best at 0.798701. Accuracy highlights the classifiers' varying effectiveness depending on the dataset, with RF consistently performing well in fraud detection tasks.

Table 6. Classification Results with GAN for Imbalanced Datasets

Dataset	Classifier	Accuracy	Precision	Recall	F1-Score	Time (s)
Credit Card Fraud Detection	RF	0.999772	1.0	0.845238	0.916129	36.05
	GB	**0.999877**	0.987342	0.928571	0.957055	243.09
	DT	0.998982	0.632653	0.738095	0.681319	4.70
Titanic	RF	0.726190	0.714286	0.571429	0.634921	0.06
	GB	**0.797619**	0.781250	0.714286	0.746269	0.04
	DT	0.726190	0.714286	0.571429	0.634921	0.00
Diabetes	RF	0.987013	1.0	0.960000	0.979592	0.09
	GB	**0.993506**	1.0	0.980000	0.989899	0.12
	DT	0.980519	1.0	0.940000	0.969072	0.00

Table 6 compares the performance of three classifiers (RF, GB and DT) across three different datasets. In the Credit Card Fraud Detection dataset, GB outperforms the others with the highest accuracy (0.999877), excellent precision (0.987342), and the best recall (0.928571), making it the most effective choice. For the Titanic dataset, GB also leads with an accuracy of 0.797619, indicating its robustness in identifying positive cases. In the Diabetes dataset, all classifiers performed well, but GB again achieved the highest accuracy (0.993506) and intense precision and recall. GB consistently shows superior performance across the datasets, underscoring its effectiveness in various classification tasks.

Table 7. Classification Results with SMOTified-GAN for Imbalanced Datasets

Dataset	Classifier	Accuracy	Precision	Recall	F1-Score	Time (s)
Credit Card Fraud Detection	RF	**0.999894**	0.999789	1.0	0.999894	240.02
	GB	0.977912	0.986025	0.969654	0.977771	489.30
	DT	0.998215	0.997459	0.998982	0.998220	26.20
Titanic	RF	0.467290	0.492308	0.571429	0.528926	0.07
	GB	**0.504673**	0.525424	0.553571	0.539130	0.03
	DT	0.467290	0.492308	0.571429	0.528926	0.00
Diabetes	RF	**0.790000**	0.756522	0.861386	0.805556	0.10

Journal of Information Systems Engineering and Management
2025, 10(45s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

	GB	0.785000	0.758929	0.841584	0.798122	0.12
	DT	0.735000	0.744898	0.722772	0.733668	0.00

Table 7 presents the performance metrics of three classifiers (RF, GB and DT) across three different datasets. For the Credit Card Fraud Detection dataset, RF achieves the highest accuracy at 0.999894, closely followed by DT at 0.998215, while GB lags at 0.977912. RF also exhibits remarkable precision (0.999789) and perfect recall (1.0), leading to a very high F1-score of 0.999894. In contrast, the Titanic dataset shows significantly lower performance across all classifiers, with GB performing slightly better with an accuracy of 0.504673, while both RF and DT score the same at 0.467290. Lastly, the Diabetes dataset reveals that RF again outperforms the others with an accuracy of 0.79 while both GB and DT show similar performance, around 0.785 and 0.735, respectively. These results illustrate RF's strong capabilities in the Credit Card Fraud Detection and Diabetes datasets, whereas all classifiers struggle with the Titanic dataset.

Table 8. Classification Results with GANified-SMOTE for Imbalanced Datasets

Dataset	Classifier	Accuracy	Precision	Recall	F1-Score	Time (s)
Credit Card Fraud Detection	RF	**0.999965**	0.999895	1.0	0.999947	379.45
	GB	0.986951	0.994527	0.966252	0.980186	746.66
	DT	0.999009	0.998701	0.998333	0.998517	34.41
Titanic	RF	**0.760000**	0.693878	0.557377	0.618182	0.09
	GB	**0.760000**	0.679245	0.590164	0.631579	0.05
	DT	**0.760000**	0.693878	0.557377	0.618182	0.00
Diabetes	RF	0.879365	0.877778	0.745283	0.806123	0.12
	GB	0.869841	0.849462	0.745283	0.793970	0.18
	DT	**0.888889**	0.851485	0.811321	0.830918	0.00

Table 8 compares the performance of various classifiers on three datasets. For the Credit Card Fraud Detection dataset, the RF model excels with an impressive accuracy of 0.999965 and perfect recall, indicating its effectiveness in identifying fraudulent transactions with minimal false positives. GB and DT models also perform well, but with lower accuracy and recall than RF. In the Titanic dataset, all classifiers achieve the same accuracy of 0.76, suggesting that none are particularly strong in predicting survival outcomes, although precision and recall differ slightly. Finally, in the Diabetes dataset, the RF model again demonstrates robust performance with an accuracy of 0.879365 and a high F1-Score of 0.806123. The Decision Tree model also shows promise with the highest accuracy of 0.888889, but has slightly lower precision and recall. These results highlight the importance of selecting appropriate classifiers based on each dataset's specific characteristics and requirements to achieve optimal performance.

Table 9. Classification Results with GANified-SMOTE with Latent Factor for Imbalanced Datasets

Dataset	Classifier	Accuracy	Precision	Recall	F1-Score	Time (s)
Credit Card Fraud Detection	RF	**0.999971**	0.999930	0.999983	0.999956	355.99
	GB	0.987971	0.994972	0.968884	0.981755	726.03
	DT	0.999015	0.998806	0.998245	0.998525	35.52
Titanic	RF	**0.754286**	0.680000	0.557377	0.612613	0.08
	GB	0.702857	0.591837	0.475410	0.527273	0.05
	DT	**0.754286**	0.680000	0.557377	0.612613	0.00
Diabetes	RF	**0.882540**	0.879121	0.754717	0.812183	0.12
	GB	0.857143	0.814433	0.745283	0.778325	0.18
	DT	0.866667	0.813725	0.783019	0.798077	0.00

Table 9 presents the performance metrics of classifiers on the Credit Card Fraud Detection, Titanic, and Diabetes datasets, highlighting their varying effectiveness. In the Credit Card Fraud Detection dataset, the RF model excels

with an accuracy of 0.999971 and a recall of 0.999983, demonstrating its ability to identify fraudulent transactions accurately. The GB and DT models also perform well, but with lower recall for GB. For the Titanic dataset, both RF and DT achieve the same accuracy of 0.754286 but struggle with precision and recall, indicating challenges in predicting survival outcomes. The GB model performs even less effectively. In the Diabetes dataset, RF again shows strong performance with an accuracy of 0.88254 and a solid F1-Score of 0.812183, while GB and DT also maintain good accuracy. These results underscore the need to choose classifiers carefully based on dataset characteristics and the required balance of precision and recall for optimal prediction outcomes.

In conclusion, the RF classifier, when utilized with the GANified-SMOTE with Latent Factor technique, demonstrates unparalleled effectiveness in managing imbalanced datasets, particularly excelling in the Credit Card Fraud Detection task with an outstanding accuracy of 0.999971 and near-perfect recall. This combination enhances the model's predictive capabilities and significantly reduces false positives, making it an optimal choice for critical applications such as fraud detection. The exceptional performance of RF paired with GANified-SMOTE with Latent Factor underscores the importance of selecting advanced classifiers and innovative techniques in machine learning, ensuring the highest levels of accuracy and reliability in classification tasks.

CONCLUSION

In addressing the challenges posed by imbalanced datasets, it is evident that traditional classification techniques often struggle to predict minority classes, leading to suboptimal performance in accurately predicting and increased false negatives. The results from the various classifiers applied across different datasets highlight the necessity of employing advanced strategies to mitigate these issues. Models can be significantly enhanced by incorporating methods such as SMOTE and its variations, improving their ability to recognize and correctly classify underrepresented instances. This not only clusters predictive accuracy but also elevates the model's overall reliability, particularly in critical applications where the cost of misclassification is high.

Among the various approaches tested, combining the RF classifier with the GANified-SMOTE with the Latent Factor technique emerges as the most effective solution for handling imbalanced datasets. This integration achieves an exceptional accuracy of 0.999971 in the Credit Card Fraud Detection task, alongside near-perfect precision and recall metrics. The ability of this technique to generate synthetic instances while preserving the underlying data distribution allows the RF classifier to make more informed decisions, thereby minimizing false positives and enhancing overall performance. Such results underscore the effectiveness of combining robust classifiers with innovative sampling methods to achieve superior classification outcomes.

Future research could explore several avenues to enhance the performance of classifiers on imbalanced datasets. One potential direction is the investigation of hybrid approaches that combine multiple sampling techniques, such as integrating GANified-SMOTE with other oversampling and undersampling methods to optimize data representation. Additionally, exploring the application of deep learning architectures in conjunction with these techniques may yield promising results, particularly in more complex datasets. Lastly, extending this research to real-world scenarios, where data may evolve, could provide insights into these models' adaptability and long-term effectiveness, paving the way for more resilient and accurate predictive systems in various critical domains.

ACKNOWLEDGMENT

This research was supported by Universiti Tun Hussein Onn Malaysia (UTHM) through GPPS (Vot Q662). This research was funded by the National Science and Technology Council (NSTC), Taiwan, under grant number NSTC 113-2221-E-035-072.

REFERENCES

[1] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. Expert systems with applications, 73, 220-239.

[2] Felix, E. A., & Lee, S. P. (2019). Systematic literature review of preprocessing techniques for imbalanced data. IET Software, 13(6), 479-496. Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, M. S., & Zeineddine, H. (2019). An experimental study with imbalanced classification approaches for credit card fraud detection. IEEE Access, 7, 93010-93022.

[3] Zhang, Y. F., Lu, H. L., Lin, H. F., Qiao, X. C., & Zheng, H. (2022). The Optimized Anomaly Detection Models Based on an Approach of Dealing with Imbalanced Dataset for Credit Card Fraud Detection. Mobile Information Systems, 2022.
[4] Paulus, J. K., & Kent, D. M. (2020). Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. NPJ digital medicine, 3(1), 99.
[5] Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. 2020 11th International Conference on Information and Communication Systems (ICICS), 243-248.
[6] Brandt, J., & Lanzén, E. (2021). A comparative review of SMOTE and ADASYN in imbalanced data classification.
[7] Sharma, A., Singh, P. K., & Chandra, R. (2022). SMOTified-GAN for class imbalanced pattern classification problems. Ieee Access, 10, 30655-30665.
[8] Sun, L., Li, M., Ding, W., Zhang, E., Mu, X., & Xu, J. (2022). AFNFS: Adaptive fuzzy neighborhood-based feature selection with adaptive synthetic over-sampling for imbalanced data. Information Sciences, 612, 724-744.
[9] Rubaidi, Z., Ammar, B., & Benaouicha, M. (2022). Fraud Detection Using Large-scale Imbalance Dataset. Int. J. Artif. Intell. Tools, 31, 2250037:1-2250037:23.
[10] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.
[11] Gnip, P., Vokorokos, L., & Drotár, P. (2021). Selective oversampling approach for strongly imbalanced data. PeerJ Computer Science, 7, e604.
[12] Ratnasari, A. P. (2024). Performance of Random Oversampling, Random Undersampling, and SMOTE-NC Methods in Handling Imbalanced Class in Classification Models. Valley International Journal Digital Library, 494-501.
[13] García, V., Sánchez, J. S., Marqués, A. I., Florencia, R., & Rivera, G. (2020). Understanding the apparent superiority of over-sampling through an analysis of local information for class-imbalanced data. Expert Systems with Applications, 158, 113026.
[14] Dablain, D., Krawczyk, B., & Chawla, N. V. (2022). DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data. IEEE Transactions on Neural Networks and Learning Systems, 34(9), 6390-6404.
[15] Hordri, N. F., Yuhaniz, S. S., Azmi, N. F. M., & Shamsuddin, S. M. (2018). Handling class imbalance in credit card fraud using resampling methods. Int. J. Adv. Comput. Sci. Appl, 9(11), 390-396.
[16] Zheng, Z., Cai, Y., & Li, Y. (2015). Oversampling method for imbalanced classification. Computing and Informatics, 34(5), 1017-1037.
[17] Khushi, M., Shaukat, K., Alam, T. M., Hameed, I. A., Uddin, S., Luo, S., ... & Reyes, M. C. (2021). A comparative performance analysis of data resampling methods on imbalance medical data. IEEE Access, 9, 109960-109975.
[18] Gu, X., Angelov, P. P., & Soares, E. A. (2020). A self-adaptive synthetic over-sampling technique for imbalanced classification. International Journal of Intelligent Systems, 35(6), 923-943.
[19] Rahmi, N. S., Wardhani, N. W. S., Mitakda, M. B., Fauztina, R. S., & Salsabila, I. (2022, November). SMOTE Classification and Random Oversampling Naive Bayes in Imbalanced Data:(Case Study of Early Detection of Cervical Cancer in Indonesia). In 2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA) (pp. 1-6). IEEE.
[20] Rendon, E., Alejo, R., Castorena, C., Isidro-Ortega, F. J., & Granda-Gutierrez, E. E. (2020). Data sampling methods to deal with the big data multi-class imbalance problem. Applied Sciences, 10(4), 1276.
[21] Bobadilla, J., Gutiérrez, A., Yera, R., & Martínez, L. (2023). Creating synthetic datasets for collaborative filtering recommender systems using generative adversarial networks. Knowledge-Based Systems, 280, 111016.
[22] Sabha, S., Assad, A., Din, N., & Bhat, M. (2023). Comparative Analysis of Oversampling Techniques on Small and Imbalanced Datasets Using Deep Learning. 2023 3rd International conference on Artificial Intelligence and Signal Processing (AISP), 1-5.
[23] Xiang, Z., Xu, Y., & Tang, Z. (2023). How Does Oversampling Affect the Performance of Classification Algorithms?. 2023 IEEE Symposium on Computers and Communications (ISCC).

[24] Hayaty, M., Muthmainah, S., & Ghufran, S. M. (2020). Random and synthetic over-sampling approach to resolve data imbalance in classification. International Journal of Artificial Intelligence Research, 4(2), 86-94.
[25] Hasanin, T., & Khoshgoftaar, T. (2018, July). The effects of random undersampling with simulated class imbalance for big data. In 2018 IEEE international conference on information reuse and integration (IRI) (pp. 70-79). IEEE.
[26] Hoyos-Osorio, J., Alvarez-Meza, A., Daza-Santacoloma, G., Orozco-Gutierrez, A., & Castellanos-Dominguez, G. (2021). Relevant information undersampling to support imbalanced data classification. Neurocomputing, 436, 136-146.
[27] Chakraborty, S., Kumar, K., Tadepalli, K., Pailla, B. R., & Roy, S. (2024). Unleashing the power of explainable AI: sepsis sentinel's clinical assistant for early sepsis identification. Multimedia Tools and Applications, 83(19), 57613-57641.
[28] Chen, Y., & Zhang, Z. (2022). An easy numeric data augmentation method for early-stage COVID-19 tweets exploration of participatory dynamics of public attention and news coverage. Information Processing & Management, 59(6), 103073.
[29] Khan, A. A., Chaudhari, O., & Chandra, R. (2023). A review of ensemble learning and data augmentation models for class imbalanced problems: combination, implementation and evaluation. Expert Systems with Applications, 122778.
[30] Bagui, S., & Li, K. (2021). Resampling imbalanced data for network intrusion detection datasets. Journal of Big Data, 8(1), 1-41.
[31] Bao, L., Juan, C., Li, J., & Zhang, Y. (2016). Boosted near-miss under-sampling on SVM ensembles for concept detection in large-scale imbalanced datasets. Neurocomputing, 172, 198-206.
[32] Palli, A. S., Jaafar, J., Hashmani, M. A., Gomes, H. M., & Gilal, A. R. (2022). A hybrid sampling approach for imbalanced binary and multi-class data using clustering analysis. IEEE Access, 10, 118639-118653.
[33] Khushi, M., Shaukat, K., Alam, T. M., Hameed, I. A., Uddin, S., Luo, S., ... & Reyes, M. C. (2021). A comparative performance analysis of data resampling methods on imbalance medical data. IEEE Access, 9, 109960-109975.
[34] Lin, W. C., Tsai, C. F., Hu, Y. H., & Jhang, J. S. (2017). Clustering-based undersampling in class-imbalanced data. Information Sciences, 409, 17-26.
[35] Guan, D., Yuan, W., Lee, Y. K., & Lee, S. (2009). Nearest neighbor editing aided by unlabeled data. Information Sciences, 179(13), 2273-2282.
[36] Kraiem, M. S., Sánchez-Hernández, F., & Moreno-García, M. N. (2021). Selecting the suitable resampling strategy for imbalanced data classification regarding dataset properties. An approach based on association models. Applied Sciences, 11(18), 8546.
[37] Liu, B., & Tsoumakas, G. (2020). Dealing with class imbalance in classifier chains via random undersampling. Knowledge-Based Systems, 192, 105292.
[38] Devi, D., Biswas, S. K., & Purkayastha, B. (2020, July). A review on solution to class imbalance problem: Undersampling approaches. In 2020 international conference on computational performance evaluation (ComPE) (pp. 626-631). IEEE.
[39] Zhang, R., Zhang, Z., & Wang, D. (2021). RFCL: A new under-sampling method of reducing the degree of imbalance and overlap. Pattern Analysis and Applications, 24, 641-654.
[40] Koziarski, M. (2020). Radial-based undersampling for imbalanced data classification. Pattern Recognition, 102, 107262.
[41] Chung, J., Zhang, J., Saimon, A. I., Liu, Y., Johnson, B. N., & Kong, Z. (2024). Imbalanced spectral data analysis using data augmentation based on the generative adversarial network. Scientific Reports, 14(1), 13230.
[42] Elreedy, D., Atiya, A. F., & Kamalov, F. (2024). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. Machine Learning, 113(7), 4903-4923.
[43] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.
[44] Pradipta, G. A., Wardoyo, R., Musdholifah, A., Sanjaya, I. N. H., & Ismail, M. (2021, November). SMOTE for handling imbalanced data problem: A review. In 2021 sixth international conference on informatics and computing (ICIC) (pp. 1-8). IEEE.

[45] Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-smote: A new over-sampling method in imbalanced data sets learning. In International Conference on Intelligent Computing (pp. 878–887). Springer.
[46] Nguyen, H. M., Cooper, E. W., & Kamei, K. (2011). Borderline over-sampling for imbalanced data classification. International Journal of Knowledge Engineering and Soft Data Paradigms, 3(1), 4–21.
[47] Wang, S., Dai, Y., Shen, J., & Xuan, J. (2021). Research on expansion and classification of imbalanced data based on SMOTE algorithm. Scientific reports, 11(1), 24039.
[48] Rachmatullah, M. I. C. (2022). The Application of Repeated SMOTE for Multi Class Classification on Imbalanced Data. MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, 22(1), 13-24.
[49] Abd El-Naby, A., Hemdan, E. E. D., & El-Sayed, A. (2023). An efficient fraud detection framework with credit card imbalanced data in financial services. Multimedia tools and applications, 82(3), 4139-4160.
[50] Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative adversarial networks: An overview. IEEE signal processing magazine, 35(1), 53-65.
[51] Sahal, N., Krishnamoorthy, R., & Singh, N. (2024, January). Generating Synthetic Text using Generative Adversarial Networks. In 2024 International Conference on Optimization Computing and Wireless Communication (ICOCWC) (pp. 1-7). IEEE.
[52] Petzka, H., Kronvall, T., & Sminchisescu, C. (2022). Discriminating Against Unrealistic Interpolations in Generative Adversarial Networks. arXiv preprint arXiv:2203.01035.
[53] Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. 2020 11th International Conference on Information and Communication Systems (ICICS), 243-248.
[54] Arjovsky, M., Chintala, S., & Bottou, L. (2017, July). Wasserstein generative adversarial networks. In International conference on machine learning (pp. 214-223). PMLR.
[55] Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.
[56] Ferreira, F., Lourenço, N., Cabral, B., & Fernandes, J. P. (2021). When two are better than one: Synthesizing heavily unbalanced data. IEEE Access, 9, 150459-150469.
[57] Charitou, C., Dragicevic, S., & Garcez, A. D. A. (2021). Synthetic data generation for fraud detection using gans. arXiv preprint arXiv:2109.12546.
[58] Cheah, P. C. Y., Yang, Y., & Lee, B. G. (2023). Enhancing financial fraud detection through addressing class imbalance using hybrid SMOTE-GAN techniques. International Journal of Financial Studies, 11(3), 110.
[59] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of big data, 6(1), 1-48.
[60] Andresini, G., Appice, A., De Rose, L., & Malerba, D. (2021). GAN augmentation to deal with imbalance in imaging-based intrusion detection. Future Generation Computer Systems, 123, 108-127.
[61] Tarawneh, A. S., Hassanat, A. B., Almohammadi, K., Chetverikov, D., & Bellinger, C. (2020). Smotefuna: Synthetic minority over-sampling technique based on furthest neighbour algorithm. IEEE Access, 8, 59069-59082.
[62] Wang, Z., Wang, W., Yang, Y., Han, Z., Xu, D., & Su, C. (2022). CNN-and GAN-based classification of malicious code families: A code visualization approach. International Journal of Intelligent Systems, 37(12), 12472-12489.
[63] Sharma, A., Singh, P. K., & Chandra, R. (2022). SMOTified-GAN for class imbalanced pattern classification problems. Ieee Access, 10, 30655-30665.
[64] Li, J., Fu, H., Hu, K., & Chen, W. (2023). Data Preprocessing and Machine Learning Modeling for Rockburst Assessment. Sustainability, 15(18), 13282.
[65] Janiobachmann. (2019b, July 3). Credit Fraud || Dealing with Imbalanced Datasets. Kaggle. https://www.kaggle.com/code/janiobachmann/credit-fraud-dealing-with-imbalanced-datasets/input
[66] Subinium. (2021, March 25). *Awesome Visualization with Titanic Dataset*. Kaggle. https://www.kaggle.com/code/subinium/awesome-visualization-with-titanic-dataset
[67] Mragpavank. (2021b, March 24). PIMA Indians Diabetes Database. Kaggle. https://www.kaggle.com/code/mragpavank/pima-indians-diabetes-database/input
[68] Wongvorachan, T., He, S., & Bulut, O. (2023). A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. Information, 14(1), 54.