Journal of Informatics and Web Engineering
Vol. 4 No. 3 (October 2025) eISSN: 2821-370X

Enhancing Imbalanced Data Augmentation: A Comparative Study of GANified-SMOTE and Latent Factor Integration

Rusma Anieza Ruslan[1], Nureize Arbaiy[2], Pei-Chun Lin[3*]
[1,2]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia
[3]Department of Information Engineering and Computer Science, Feng Chia University, Taichung, Taiwan
**corresponding author: (peiclin@fcu.edu.tw; ORCiD: 0000-0003-0735-2693)*

Abstract - One such serious problem in machine learning (ML) is imbalanced datasets. Minority class samples are usually sparse but hold significant meaning. The model can become biased toward the majority class due to unbalanced class distribution. This results in fraudulently high accuracy without being able to detect minority cases. This bias is also most perilous in critical applications, where ignoring minority cases can be highly destructive. To overcome this problem, the Synthetic Minority Oversampling Technique (SMOTE) is one of the most widely used. SMOTE creates balanced class distribution by interpolating between existing minority samples. It creates samples that are too close to one another and can lead to overfitting and limit the generalization of the model. Recent advancements in generative modeling, especially Generative Adversarial Networks (GANs), offer a more effective solution to handle class imbalance. GANs utilizes a generative discriminator structure to produce synthetic data highly similar to real data. A hybrid technique called GANified-SMOTE combines the power of SMOTE with the generation power of GANs to produce more diverse and realistic minority class samples. The technique improves the model strength and eliminates the limitations of traditional oversampling. This paper presents the incorporation of latent factors into the architecture of GANified-SMOTE framework. Latent variables reveal hidden structures and relations in the data, leading to a closer synthetic sample and improving classification accuracy. By incorporating latent factors, this research aims to build a better oversampling method for imbalanced classification sets.

Keywords—Imbalance Dataset, SMOTE, Generative Adversarial Networks, Latent Factor, Accuracy, Classification

Received: 14 June 2025; Accepted: 29 August 2025; Published: 16 October 2025

1. INTRODUCTION

Fraud detection [1], [2], diagnosis in healthcare [3] and natural language processing are only a few of the areas which are faced with imbalanced dataset problems in ML. The very crucial minority class, made up of infrequent occurrences is typically underrepresented [4]. The resulting models will be skewed towards the majority class, yielding high accuracy outcomes from the algorithm but poor performance for the minority class [5]. In dire cases, bias can result

in undetected illness or missed samples of manipulative behavior, which can lead to significant damage. Several approaches have been given to address the problem caused by biased datasets. One such widely used oversampling technique is known as SMOTE. This method enhances minority class coverage by oversampling, making synthetic samples by interpolating present samples [6], [7]. Although this form of oversampling is advantageous, it also has the potential to overfit depending on the type of sample that results [8]. Furthermore, regular oversampling methods may also fail to capture minority class distribution and adversely affect model's outcomes.

GANs, are a new approach in generative modelling. A GAN has two neural networks which are a generator and a discriminator that experience a competition [9]. The generator produces imitated data, and the discriminator examines whether it is original or not in comparison to real data [10], [11]. Based on this architecture, GANs can generate realistic samples with close identity to the statistical attributes of the training set. The power of GANs lies in generating varied and intricate data, which can be utilized to enhance imbalanced datasets. GANified-SMOTE merges the power of SMOTE with GANs' generation property and thus provide various advantages over more traditional approaches [12]. Since GANs are employed, GANified-SMOTE can produce better quality synthetic samples that better capture the minority class distribution. Additionally, this step adds diversity to the data and decreases overfitting and consequently enhances the generalization of models trained on this to new data.

In the case of GANified-SMOTE, this model seems to suffer from generating good quality samples for the minority class. This results in overfitting with poor performance. Additionally, the absence of dimensionality reduction results in the inclusion of noise or irrelevant characteristics that can weaken the quality of synthetic samples and fail strong classification. Thus, in the present study, the latent factors are included in GANified-SMOTE, which is a positive step towards addressing the problem of synthetic sample generation in the context of imbalance. By utilizing hidden factors like merging, there is significant potential for ongoing advancements in the field, which could greatly enhance classification performance. The key concept behind this consolidation is to augment the functionality of the framework by improving the complexities in data handled by GANified-SMOTE.

This paper is structured as follows. Section 1 presents the background of the study. Section 2 presents related literature under Related Work. Under Methodology, Section 3 explains the research methods utilized in the study. In Section 4, the author states the findings under Results and Discussion. Finally, Section 5 provides the overall findings in the Conclusion.

2. RELATED WORKS

This section presents a literature review in this research area consisting of the SMOTE resampling technique, GANs for data augmentation and a hybrid approach of GANified-SMOTE.

2.1. Resampling Technique: SMOTE

Resampling technique constitutes a pillar of the approach to handling imbalanced datasets. Among them, SMOTE has been leading due to its innovative approach to generating synthetic samples of the minority class. SMOTE was proposed by Chawla et al. [13] to avoid the limitations of random oversampling, which leads to overfitting by merely duplicating the existing minority samples. Yet, SMOTE creates new samples through interpolating between the existing minority class samples, basically increasing the diversity and number of the minority class samples [14]. SMOTE includes several key steps in its algorithm. First, for each sample of the minority class, the algorithm chooses its k-nearest neighbours [15]. Then, it selects one or more of them at random and generates synthetic samples by adding points along all line segments connecting the original sample to its selected neighbours. By doing so, SMOTE generates new samples that are not mere replicas of existing samples but inject variety in the minority class with data points that reflect the underlying distribution.

Apart from its effectiveness, SMOTE also presents its drawbacks. One of the major problems is that it creates representative synthetic samples of the true data distribution when the minority class is sparse or has intricate boundaries. Also, SMOTE may have the tendency to introduce noise by interpolating distant samples in the feature space and generating samples that do not reflect the actual attributes of the minority class. Various variants of SMOTE have been introduced that address their issues. Table 1 provides a summary of SMOTE variants.

Table 1. Variants of SMOTE

Variants	Mechanism
SMOTE: Synthetic Minority Oversampling Technique [13]	Creates synthetic samples instead of using replacement for oversampling
Borderline-SMOTE: Borderline Synthetic Minority Oversampling Technique [15]	Objective was to generate synthetic samples for minority class samples that are near the decision boundary
Safe-level SMOTE: Safe-level Synthetic Minority Oversampling Technique [16]	Used a safe level constraint to determine suitable samples for generating synthetic samples and removing noise
SVM SMOTE: Support Vector Machine Synthetic Minority Oversampling Technique [17]	Used SVM to determine and generate synthetic samples with the support vectors
CDSMOTE: Clustered Synthetic Minority Oversampling Technique [18]	Uses a cluster-based approach to generate synthetic samples based on minority sample distribution and density
Deep SMOTE: Deep Learning-based Synthetic Minority Oversampling Technique [19]	Applied Deep Learning to generate synthetic samples by learning the minority class sample distribution

2.2. GANs in Data Augmentation

GANs have revolutionized the field of generative modelling since their introduction in 2014 [9]. GANs consist of two neural networks: a generator and a discriminator. The primary function of the generator is to generate synthetic data, and the discriminator determines whether the data it receives is real or artificial [10], [11]. The system provides a competitive environment that requires both networks to continuously enhance their performance. During training, the generator attempts to produce synthetic samples that are indistinguishable from real data.[10]. Meanwhile, the discriminator enhances its ability to distinguish between synthetic and real samples [11]. This adversarial training approach results in synthetic data of high quality that effectively captures the underlying distribution of the original dataset.

GANs are seen as effective and versatile across a variety of applications, including synthesis and text-to-image synthesis, as well as speech synthesis. Their successful applications highlight their capability to revolutionize a variety of industries by generating realistic data and improving model performance. In the scenario of imbalanced datasets, GANs offer an effective solution compared to traditional resampling techniques. One of the key advantages of using GANs for data augmentation is that they can learn complex data distributions. Unlike techniques like SMOTE, which perform linear interpolation between existing samples, GANs can generate diverse and realistic synthetic samples that represent complex patterns in the data. This is particularly applicable to minority classes in imbalanced data, where the distribution may be non-linear and complex. It has been shown that classifiers trained on imbalanced data. Through the generation of synthetic samples that are characteristic of the minority class, GANs can enhance the model's generalization and accurate prediction on unseen data. This approach not only addresses the issue of class imbalance but also minimizes the risk of overfitting, since the new samples provide a larger variety of training samples. Several variants of GANs have been explored to further enhance their performance in data augmentation, as outlined in Table 2.

2.3. GANified-SMOTE: A Hybrid Approach

GANified-SMOTE is a crucial step towards seeking the solution to the problem that occurs when handling imbalanced datasets with the utilities that are being integrated between GANs and SMOTE [12]. The hybrid approach is supposed to leverage the sample generation capabilities of GANs while maintaining the fundamental principles of SMOTE, which deals with having the right kind of meaningful and representative samples being generated for the minority class. In its algorithm, SMOTE starts the process of generating artificial samples of the minority class. The standard

SMOTE generates samples by interpolating the minority samples. While GANified-SMOTE allows the generator to learn a more complex distribution of the minority class. The GAN can generate diverse synthetic samples capable of capturing subtle patterns and variations through training on the minority data. Therefore, it can generate a more complex representation of the minority class. This is useful in scenarios where the minority class is sparse or has non-linear dependencies in the feature space. Another key advantage of GANified-SMOTE is that it can create high-quality generated samples that are not copies or slightly edited versions of existing samples. With the generation of more diverse and realistic samples, GANified-SMOTE prevents the risk of overfitting. The synthesized data not only amounts to additional minority samples but also enhances the quality overall, providing a more robust training set for classifiers.

Table 2. Variants of GANs

Variants	Mechanism
GANs: Generative Adversarial Networks [9]	A new model framework for generative model estimation with an adversarial process through the joint training of a generative model (G) and discriminative model (D)
WGAN: Wasserstein Generative Adversarial Network [20]	An adaption of traditional GAN training
cGAN: Conditional Generative Adversarial Network [21]	A technique for training generative models to facilitate conditioning data on specific inputs
Duo-GAN: Dual Generative Adversarial Network [22]	Generates synthetic datasets to address the problems of highly imbalanced data
Majority-Minority GAN Transfer: Majority-Minority Generative Adversarial Network Transfer [23]	Harness the power of GANs and transfer learning to create more effective solutions for synthetic data generation
CTAB-GAN: Categorical and Tabular Generative Adversarial Network [24]	A specially designed framework particularly for generating synthetic data for tabular datasets with categorical and continuous features
SDG-GAN: Stochastic Distributional Generative Adversarial Network [25]	One of the modifications of the traditional GAN and one that is highly interested in modelling and synthesizing data that possesses inherent stochasticity and distributional characteristics
cWGAN: Conditional Wasserstein Generative Adversarial Network [26]	A variant of the WGAN with the addition of conditional data to serve as a guide

3. RESEARCH METHODOLOGY

The proposed methodology integrates GANified-SMOTE [12] and latent factor. The approach is aimed at enhancing the minority class representation of imbalanced datasets and thus improving classifier performance.

3.1. Data Collection and Preprocessing

Data collection is accomplished through obtaining datasets from the Kaggle platform that offers a wide range of publicly accessible datasets in different domains [27]. Two specific datasets are employed to examine the performance of the proposed methodology in this study, as listed in Table 3.

Preprocessing data plays a key role in transforming raw data into a neat form to be utilized in analysis or modelling [30], as listed in Table 4.

Table 3. Lists of Datasets

Datasets	Explanation
Credit Card Fraud Detection [28]	These are credit card transactions, and the aim is detection of fraudulent activities
Pima Indians Diabetes Database [29]	These are Pima Indians' medical records and are used to predict diabetes onset from diagnostic tests

Table 4. Preprocessing Steps

Steps	Explanation
Feature and Target Separation	The information is divided into two categories: feature (input variables) and target variables (result to be predicted). This separation makes data manipulation and analysis more straightforward
Feature Scaling	To maintain all the input features on the same scale, feature scaling techniques such as normalization or standardization are employed. The step is necessary for scale-sensitive algorithms such as gradient descent-based ones
Class Separation	The target variable is analysed to identify different classes. Data is split into various subsets based on these classes, which is particularly important in addressing class imbalance
Data Partitioning	The dataset is divided into the train and test sets. The training set is used to train the models, while the test set is kept aside to evaluate the performance of models. Partitioning is performed to ensure that the performance of the model is evaluated against unseen data, providing a better estimate of its generalization ability

3.2. Integration of GAN and SMOTE

The study tries to utilize a hybrid approach known as GANified-SMOTE with Latent Factor, wherein SMOTE. The approach focuses on generating real-like synthetic samples of the minority class for classification. By using GANs for generating high-quality synthetic data and then further augmenting it with SMOTE, the study is intended to increase the representation of minority classes. Synthetic samples are initiated from real samples of a training set, which are evaluated by the discriminator to separate true and generated data. The generator generates synthetic samples from random noise and latent variables to mimic the features of real samples. The discriminator employs a sigmoid loss function to calculate its accuracy in relation to how well it classifies the samples, and the generator employs a mean squared error loss to calculate its capability to mislead the discriminator. Synthetic samples are then produced, and SMOTE is performed on the data generated by GAN to enhance diversity and decrease reliance on dominant patterns to avoid overfitting. The result from SMOTE is then merged with the original training set to provide a balanced and full dataset for model training. Figure 1 shows where GAN is embedded in SMOTE to generate synthetic samples.

3.3. Latent Factor Integration

The integration of latent factors in the GANified-SMOTE framework reveals hidden structures within the data. This then enhances synthetic sample quality as well as classification performance. The incorporation is necessary to construct more robust models capable of handling complex data distributions. The latent variables are incorporated into the generator's architecture as additional input dimensions. This allows the generator to learn a representation that captures more intricate patterns in the data. Figure 2 presents the pseudocode for this process.

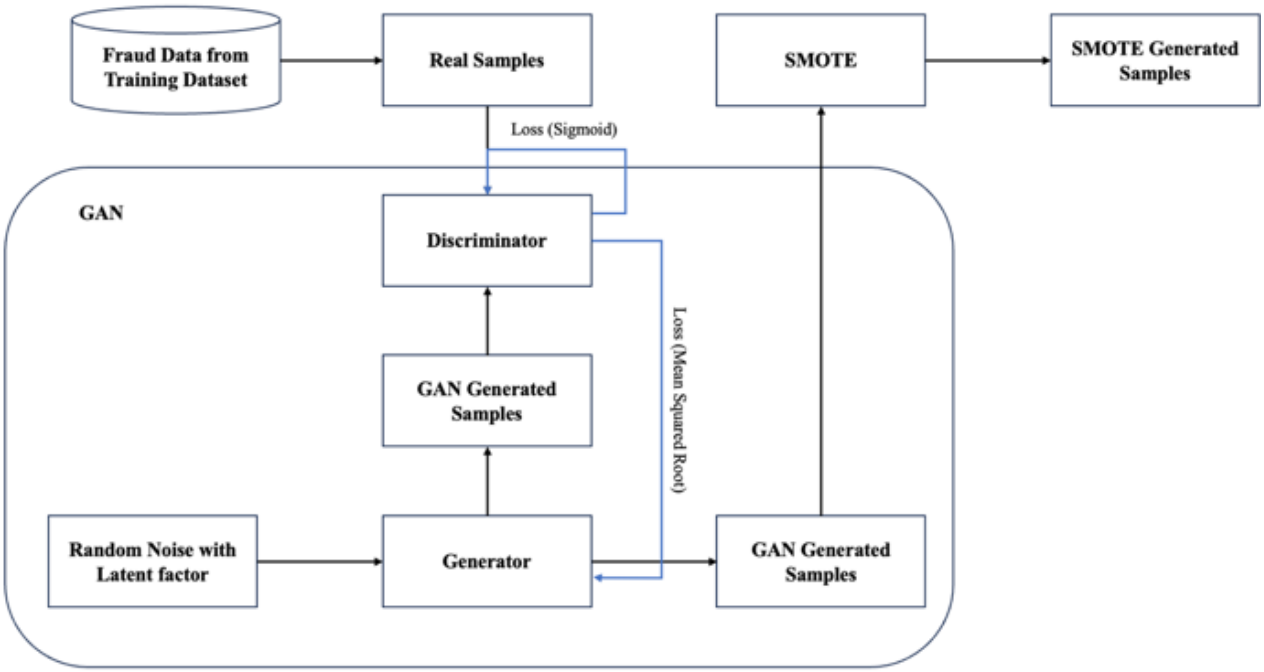

Figure 1. Proposed GANified-SMOTE with Latent Factor

Algorithm 1: GANified-SMOTE with Latent Factor

```
FUNCTION generate_random_noise (latent_dim, num_samples);

#Initialize parameters
        SET mean = 0
        SET std_dev = 1

#Generate random noise
        noise = np.random.normal(mean, std_dev, (num_samples, latent_dim))

#Return the generated noise
        RETURN noise
```

Figure 2. Pseudocode of Latent Factor Integration

3.4. Model Training and Evaluation

The study validates the efficacy and accuracy of the proposed research model after systematically applying it. The process of evaluation is performed based on various measures and techniques to quantify the performance of the model. This process is known as data classification and minority class detection in imbalanced datasets. The performance is focused on three classifiers: Random Forest (RF), Gradient Boosting (GB) and Decision Tree (DT). Each classifier is trained on a balanced dataset and tested on a corresponding set of performance metrics, including accuracy, recall, precision and F1-score. The respective formulas are as follows.

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

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

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

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

4. RESULTS AND DISCUSSIONS

Experimental results are presented in this section. It shows the effect that various strategies to deal with imbalanced datasets, as well as overfitting, have on model performance.

4.1. GANified-SMOTE

Table 5 lists the experimental results of GANified-SMOTE methods on two datasets, the Credit Card Fraud Detection and the Pima Indian Diabetes. In the case of the Credit Card Fraud Detection, the RF was phenomenal in achieving 99.9965% and 99.9895% precision and 100% perfect recall. This indicated its better performance in accurately detecting fraudulent transactions with minimized false positives. GB classifiers were also satisfactory with an accuracy of 98.6951%. However, it had a slightly lower recall, indicating that it had failed to detect some cases of fraud. DT classifier also performed well by achieving an accuracy of 99.9009%. This verifies how these classifiers perform well in detecting fraud in this significant application.

However, for the Pima Indians Diabetes dataset, all the classifiers were reasonably good but could not classify all the positive cases. The RF classifier was the best with 87.9365% accuracy and a precision of 87.7778%. But its recall of 74.5283% resulted in some of the true cases of diabetes not being classified. The GB classifier also suffered from the same drawback, with 86.9841% accuracy and a precision of 84.9462. The DT classifier was also fine with an accuracy value of 88.88889% but still performed badly in terms of recall. This situation indicates a need to improve model training further so that the model detection rate for diabetes samples improves.

Table 5. Performance for GANified-SMOTE

Datasets	**Classifier**	**Performance Metrics**			
		Accuracy	**Precision**	**Recall**	**F1-score**
Credit Card Fraud Detection	Random Forest	**0.999965**	0.999895	1.0	0.999947
	Gradient Boosting	0.986951	0.994527	0.966252	0.980186
	Decision Tree	0.999009	0.998701	0.998333	0.998517
Pima Indians Diabetes	Random Forest	0.879365	0.877778	0.745283	0.806123
	Gradient Boosting	0.869841	0.849462	0.745283	0.793970
	Decision Tree	**0.888889**	0.851485	0.811321	0.830918

4.2. GANified-SMOTE with Latent Factor

Table 6 shows the GANified-SMOTE with Latent Factor results. RF classifier works brilliantly in the Credit Card Fraud Detection with a 99.9971% accuracy level, indicating its ability to detect most of the fraud samples accurately. Its precision and recall rates are also extremely high at 99.9930% and 99.9983% levels, respectively. It means that the model not only predicts correctly but also minimises false positives. The DT classifier performs excellently with 99.9015% accuracy. The GB model lags with 98.7971% accuracy. It indicates its poorer performance in this specific application.

For the Diabetes dataset, the RF classifier was again found to perform well with 88.2540% accuracy, indicating its consistent performance in distinguishing diabetic from non-diabetic patients. The precision ratio of 87.9121% and recall ratio of 75.4717% suggest it accurately identifies true positive samples but can be optimised further in reducing false positives. The GB classifier also displays marginally lower performance figures at 85.7143% accuracy, again highlighting the RF's advantages here. Compared to the credit card dataset, the diabetes performance metrics indicate a more challenging classification task, which mirrors the intricacy of medical data analysis.

Table 6. Performance for GANified-SMOTE with Latent Factor

Datasets	**Classifier**	**Performance Metrics**			
		Accuracy	**Precision**	**Recall**	**F1-score**
Credit Card Fraud Detection	Random Forest	**0.999971**	0.999930	0.999983	0.999956
	Gradient Boosting	0.987971	0.994972	0.968884	0.981755
	Decision Tree	0.999015	0.998806	0.998245	0.998525
Pima Indians Diabetes	Random Forest	0.882540	0.879121	0.754717	0.812183
	Gradient Boosting	0.857143	0.814433	0.745283	0.778325
	Decision Tree	**0.866667**	0.813725	0.783019	0.798077

5. CONCLUSION

In conclusion, this comparative analysis of GANified-SMOTE and GANified-SMOTE with Latent Factor demonstrates significant advancements in solving the issues related to imbalanced data. With the integration of GANs and the traditional augmentation techniques like SMOTE, the models diversify the data and improve model performance. However, there are challenges to be met, particularly regarding scalability as well as computational cost. As the size of datasets grows, the resources required to train GANs may become so costly that real-time applications in fields like medicine and fraud prevention might be limited.

To address these challenges, future research should focus on enhancing training algorithms for GANs to scale better, possibly employing mini-batch training or distributed computing. Examining hybrid approaches where GANified-SMOTE is integrated with other augmentation methods might go further toward building model robustness. Further, investigation into other latent factor extraction protocols may yield even better representations for complex relationships within data. Finally, using these techniques on multiple domains and developing new metrics for measurement will yield comprehensive model performance analysis, paving the way for further innovation in imbalanced learning and data augmentation.

ACKNOWLEDGEMENT

This research was supported by Universiti Tun Hussein Onn Malaysia (UTHM) through GPPS (Vot Q662).

FUNDING STATEMENT

The authors received no funding from any party for the research and publication of this article.

AUTHOR CONTRIBUTIONS

Rusma Anieza Ruslan: Study Conception and Design, Data Collection, Analysis and Interpretation of Results;
Nureize Arbaiy: Study Conception and Design & Manuscript Preparation;
Pei-Chun Lin: study Conception and Design, Analysis and Interpretation of Results.

CONFLICT OF INTERESTS

No conflict of interests were disclosed.

ETHICS STATEMENTS

Our publication ethics follow The Committee of Publication Ethics (COPE) guideline. https://publicationethics.org/.

REFERENCES

[1] Y. F. Zhang, H. L. Lu, H. F. Lin, X. C. Qiao, and H. Zheng, "The optimized anomaly detection models based on an approach of dealing with imbalanced dataset for credit card fraud detection," *Mobile Information Systems*, vol. 2022, 2022, doi: 10.1155/2022/8027903.

[2] E. A. Felix, and S. P. Lee, "Systematic literature review of preprocessing techniques for imbalanced data," *IET Software*, vol. 13, no. 6, pp. 479–496, 2019, doi: 10.1049/iet-sen.2018.5211.

[3] S. Makki, Z. Assaghir, Y. Taher, R. Haque, M. S. Hacid, and H. Zeineddine, "An experimental study with imbalanced classification approaches for credit card fraud detection," *IEEE Access*, vol. 7, pp. 93010–93022, 2019, doi: 10.1109/ACCESS.2019.2927257.

[4] J. K. Paulus, and D. M. Kent, "Predictably unequal: Understanding and addressing concerns that algorithmic clinical prediction may increase health disparities," *NPJ Digital Medicine*, vol. 3, no. 1, p. 99, 2020, doi: 10.1038/s41746-020-0290-y.

[5] E. A. Anaam, S. C. Haw, and P. Naveen, "Applied fuzzy and analytic hierarchy process in hybrid recommendation approaches for E-CRM," *International Journal on Informatics Visualization*, vol. 6, no. 2, 2022, doi: 10.30630/joiv.6.2-2.1043.

[6] H. Haixiang, Y. Yijing, Z. Shang, G. Mingyun, Y. Yuanyue, and F. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Systems with Applications*, vol. 73, pp. 220–239, 2017, doi: 10.1016/j.eswa.2016.12.035.

[7] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: Overview study and experimental results," in *2020 11th International Conference on Information and Communication Systems (ICICS)*, pp. 243–248, 2020, doi: 10.1109/ICICS49469.2020.239556.

[8] J. Brandt, and E. Lanzén, "A comparative review of SMOTE and ADASYN in imbalanced data classification," 2021.

[9] A. Sharma, P. K. Singh, and R. Chandra, "SMOTified-GAN for class imbalanced pattern classification problems," *IEEE Access*, vol. 10, pp. 30655–30665, 2022, doi: 10.1109/ACCESS.2022.3152607.

[10] N. S. Rahmi, M. F. M. Fudholi, R. Hidayat, and R. R. Isnanto, "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, 2022, doi: 10.1109/ICITDA56564.2022.10058378.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[12] H. Petzka, T. Kronvall, and C. Sminchisescu, "Discriminating against unrealistic interpolations in generative adversarial networks," *arXiv:*2203.01035, 2022.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[14] G. A. Pradipta, R. A. R. Hidayat, and S. A. Kusumawardani, "SMOTE for handling imbalanced data problem: A review," in *2021 Sixth International Conference on Informatics and Computing (ICIC)*, pp. 1–8, 2021, doi: 10.1109/ICIC54025.2021.9673451.

[15] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *International Conference on Intelligent Computing*, pp. 878–887, Springer, 2005, doi: 10.1007/11538059_91.

[16] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 475–482, Springer, 2009, doi: 10.1007/978-3-642-01307-2_43.

[17] H. M. Nguyen, E. W. Cooper, and K. Kamei, "Borderline over-sampling for imbalanced data classification," *International Journal of Knowledge Engineering and Soft Data Paradigms*, vol. 3, no. 1, pp. 4–21, 2011, doi: 10.1504/IJKESDP.2011.039875.

[18] E. Elyan, C. F. Moreno-Garcia, and C. Jayne, "Cdsmote: Class decomposition and synthetic minority class oversampling technique for imbalanced-data classification," *Neural Computing and Applications*, vol. 33, no. 7, pp. 2839–2851, 2021, doi: 10.1007/s00521-020-05122-1.

[19] D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 9, pp. 6390–6404, 2022, doi: 10.1109/TNNLS.2021.3074578.

[20] A. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning*, pp. 214–223, PMLR, 2017.

[21] M. Mirza, and S. Osindero, "Conditional generative adversarial nets," *arXiv*:1411.1784, 2014.

[22] F. Ferreira, A. Soares, and P. Cortez, "When two are better than one: Synthesizing heavily unbalanced data," *IEEE Access*, vol. 9, pp. 150459–150469, 2021, doi: 10.1109/ACCESS.2021.3125685.

[23] A. Langevin, L. O. Hall, and R. Woods, "Synthetic data augmentation of imbalanced datasets with generative adversarial networks under varying distributional assumptions: A case study in credit card fraud detection," *Journal of the Operational Research Society*, 2021, doi: 10.1080/01605682.2021.2006124.

[24] Z. Zhao, R. K. K. Lau, S. Singh, and Y. Wang, "CTAB-GAN: Effective table data synthesizing," in *Proceedings of the Asian Conference on Machine Learning*, pp. 97-112. PMLR, 2021.

[25] C. Charitou, S. Dragicevic, and A. D. A. Garcez, "Synthetic data generation for fraud detection using GANs," *arXiv*:2109.12546, 2021.

[26] J. Engelmann, and S. Lessmann, "Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning," *Expert Systems with Applications*, vol. 174, pp. 114582, 2021, doi: 10.1016/j.eswa.2021.114582.

[27] J. Li, X. Liu, Y. Xu, and Y. Zhang, "Data preprocessing and machine learning modeling for rockburst assessment," *Sustainability*, vol. 15, no. 18, pp. 13282, 2023, doi: 10.3390/su151813282.

[28] Janiobachmann, "Credit Fraud || Dealing with Imbalanced Datasets," *Kaggle*, Jul. 3, 2019. [Online]. Available: https://www.kaggle.com/code/janiobachmann/credit-fraud-dealing-with-imbalanced-datasets/input

[29] Mragpavank, "PIMA Indians Diabetes Database," *Kaggle*, Mar. 24, 2021. [Online]. Available: https://www.kaggle.com/code/mragpavank/pima-indians-diabetes-database/input

[30] T. Wongvorachan, S. He, and O. Bulut, "A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining," *Information*, vol. 14, no. 1, pp. 54, 2023, doi: 10.3390/info14010054.

BIOGRAPHIES OF AUTHORS

Rusma Anieza Ruslan holds a B.S. Degree in Information Technology from Universiti Tun Hussein Onn Malaysia (UTHM), Johor, in 2024. She is now continuing her studies for an M.S. in Information Technology at UTHM with research on "Addressing Imbalance Datasets in Machine Learning." She can be contacted at email: rusmaanieza17@gmail.com.

Nureize Arbaiy currently in the Department of Software Engineering at the Faculty of Computer Science and Information Technology, UTHM. Her research interests are multi-criteria decision-making, fuzzy regression, fuzzy auto-regression, information details, and probability theory. She has supervised several PhD and master's students and published articles in various international journals and conference proceedings. She has acted as a reviewer for multiple journals and conferences. She can be contacted at email: nureize@uthm.edu.my.

Pei-Chun Lin received a B.Sc. degree in the Department of Mathematics from National Kaohsiung Normal University, Kaohsiung, Taiwan and received an M.Sc. degree in the Department of Mathematical Sciences from National Chengchi University, Taipei, Taiwan. Moreover, she received a Ph.D. degree from Graduate School of Information, Production and systems, Waseda University, Fukuoka, Japan. She is currently an assistant professor in the Department of Information Engineering and Computer Science, Feng Chia University. Her research interests include soft computing, artificial intelligent, robot computing, statistical modeling and big data analysis. She can be contacted at email: peiclin@fcu.edu.tw.