Science in Information Technology Letters
Vol. 6., No. 1, May 2025, pp. 44-54
ISSN 2722-4139
http://pubs2.ascee.org/index.php/sitech

# A comparative study on SMOTE, CTGAN, and hybrid SMOTE-CTGAN for medical data augmentation

Ninda Khoirunnisa [a,1,*], Miftahurrahma Rosyda [a,2]

[a] Informatics Department, Universitas Ahmad Dahlan, Indonesia
[1] ninda@tif.uad.ac.id; [2] miftahurrahma.rosyda@tif.uad.ac.id
* Corresponding Author

ARTICLE INFO

**Article history**
Received April 29, 2025
Revised May 22, 2025
Accepted May 29, 2025

**Keywords**
Medical tabular data
Imbalance data
SMOTE
CTGAN
Data augmentation

ABSTRACT

The imbalance of clinical datasets remains a challenge in medical data mining, often resulting in models biased toward majority outcomes and reduced sensitivity to rare but clinically critical cases. This study presents a comparative evaluation of three augmentation strategies—Synthetic Minority Oversampling Technique (SMOTE), Conditional Tabular GAN (CTGAN), and a hybrid SMOTE+CTGAN—on the Framingham Heart Study dataset for cardiovascular disease prediction. Augmented datasets were evaluated using Decision Tree, Random Forest, and XGBoost classifiers across multiple metrics, including accuracy, precision, recall, and F1-score. Results demonstrate that classifiers trained on imbalanced data achieved high accuracy but poor minority recall (<0.40), confirming model's bias toward majority class. SMOTE yielded the strongest improvements in minority recall (up to 0.88 with XGBoost) and balanced F1 across classes, though at the cost of reduced majority recall. CTGAN and SMOTE+CTGAN delivered more moderate improvements in minority recall (0.66–0.77) while preserving higher majority recall (>0.86), providing a gentler trade-off. These findings indicate that while SMOTE remains a robust baseline for addressing imbalance, hybrid and GAN-based approaches offer practical alternatives for preserving majority performance. The results highlight that augmentation choice should be informed by clinical context.



## 1. Introduction

The rapid growth of electronic health records (HER) and other clinical datasets has opened new opportunities for data-driven decision-making in healthcare. However, one of the most commonly found challenges in medical data mining is class imbalance or unequal distribution. The difference in prevalence between classes in medical dataset often results in reduced model's sensitivity and ureliable outcomes due to bias toward majority cases in diagnosis because minority cases are usually underrepresented compared to the common ones [1], [2]. Research by Yuda et al. [3] has shown that early diagnosis of Alzheimer's requires highlt sensitive classifiers, yet traditional supervised learning approach struggle when faced with limited samples of positive cases.

Conventional solutions to class imbalance include resampling strategies such as undersampling the majority class and oversampling the minority class. Among these, the Synthetic Minority Over-sampling Technique (SMOTE) [4] has been widely used due to its ability to generate synthetic data by interpolating between k minority class nearest neighbours. While this method is effective in specific domains, SMOTE suffers from key limitations in overgeneralization and increased overlapping especially between samples that lie on the border between classes [5]. To address these challenges, researchers have increasingly turned toward generative models, particularly Generative Adversarial Networks (GANs) [6], which have demonstrated the ability to capture complex distribution and generate realistic synthetic samples, despite that it is originally used to generate image data.

Recent advances in GAN-based data augmentation have shown promising results for diverse healthcare applications. Zhang et al. [2] applied a Wasserstein GAN with Gradient Penalty (WGAN-GP) to one-dimensional clinical radiomics data, demonstrating superior performance ccompared to SMOTE and baseline GANs in terms of AUC, sensitivity, accuracy, and specificity. Similarly, Sharma et al. [1] introduced SMOTified-GAN, a hybrid model with two phases where the first phase include generating sample with SMOTE that further refined using GAN, resulting an increase 9% in F1-scores across benchmark datasets. This result indicate better balance of sensitivity and specificity on the proposed method. Other works have adapten GAN architectures to specific healthcare domains, namely an enhances conditional GAN was shown to preserve cardiovascular data distribution better than CTGAN [7], while WGAN-GP combined with dimensionality reduction via UMAP improved the separability of Alzheimer's diagnostic imaging data [3].

GANs have also been evaluated for their utility in generating synthetic medical tabular data. Ahmed et al. [8] evaluated six variants of GANs, namelu GAN, CGAN, CTGAN, CRAMER GAN, DRAGAN, and WGAN across multiple healthcare datasets such as Breast Cancer Wisconsin, Lung Cander, and Fetal Cardiocography. Their study found that advanced architectures like CGAN and CTGAN not only increase classification performance in terms of accuracy when combined with classifiers such as XGBoost and SVM, but also maintained statistical fidelity and correlation structures within the generated tabular data. These findings highlight the dual role of GAN-based augmentation in both addressing imbalance dataset and supporting data privacy in medical data analytics. Additionally, other approaches have explored the integration of GANs with other oversampling strategies. For example, an enhanced GAN (E-GAN) that utilize deep convolutional GAN and modified convolutional neural network (DCG-MCNN) combined with RSMOTE preprocessing improved the classification of imbalanced medical disease datasets [9], while conditional WGAN-GP has been successfully used to augment small clinical audio datasets, leading to measurable increment in F1-score [10].

Taken together, the growing number in research demonstrates that hybrid approaches integrating traditional oversampling method such as SMOTE with GAN-based models can effectively mitigate the limitations of each individual technique and enhance predictive performance in medical data mining. However, most prior studies have focused on domain-specific datasets such as radiomics, imaging, or small-scaled clinical datasets, leaving a gap in the systematic comparison of these approaches on large, population-based tabular datasets.

In this paper, we aim to address this gap by conducting a comparative analysis of three data augmentation methods, namely SMOTE, GAN, and a hybrid SMOTE+GAN on the Framingham Heart Study dataset. This dataset has been widely used in cardiovascular risk prediction, providing a representative case of real-world clinical imbalance, particularly between patients who develop

cardiovascular disease and those who do not. By evaluating the performance of these augmentation strategies across multiple classifiers, we investigate the effectiveness in improving sensitivity, specificity, and overall predictive robustness. Our findings contribute to ongoing efforts to identify practical and scalable solutions for handling class imbalance in healthcare analytics.

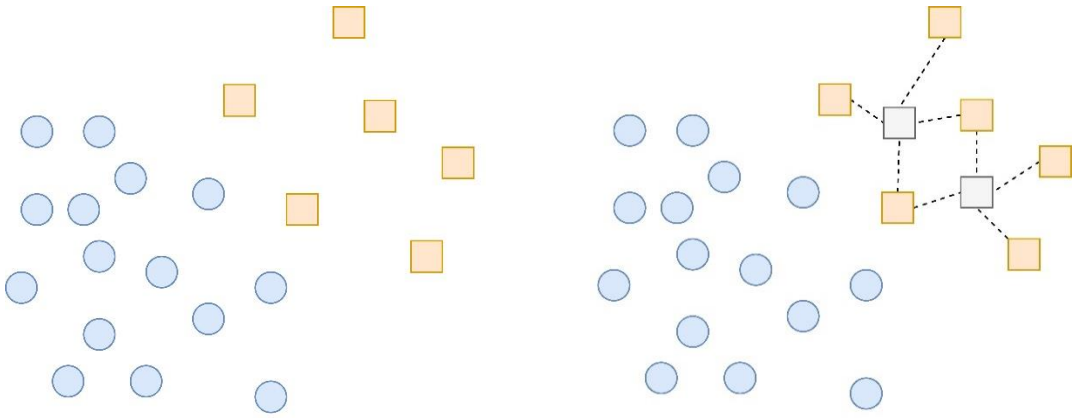
## 2. The Proposed Method/Algorithm

### 2.1. Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE was first introduced by Chawla et al. [4] as a data-level strategy to mitigate class imbalance problem in a dataset. Instead of simply duplicating samples from minority class, SMOTE generates synthetic samples by interpolating between existing k-nearest neighbours instances. Specifically, for each minority instance x, one of its neighbours xnn is selected, and a new sample is generated as follows:

$$X_{new} = X + \lambda \times (X_{nn} - X), \lambda \epsilon [0,1] \quad (1)$$

where $X$ is the original minority sample, $X_{nn}$ is one of its nearest neighbours, and λ is a random number between 0 and 1. This interpolation ensures that the new minority samples lies somewhere along the line segment between x and xnn. This process is illustrated in Fig 1.



(a) Original Samples (b) Original and Synthetic Samples

**Fig. 1.** Illustration of the SMOTE Procedure

By construcing new data points that belong to minority class rather than simply duplicating existing minority samples, SMOTE proved its ability in tackling overfitting problem and improving classifier robustness. Numerous studies in healthcare and clinical domains have demonstrated the effectiveness of SMOTE in handling skewed data distributions, including diabetes prediction [11], [12], obesity risk classification [13], body mass index (BMI) risk stratification [14]. While effective, SMOTE may also generate borderline or noisy samples when the minority and majority classes overlap significantly [15], [16].

### 2.2. CTGAN

In contrast to SMOTE, Conditional Tabular GAN (CTGAN) [17] learns and reproduces the distribution of tabular data which often includes both continuous and categorical features, handling mixed data types, and imbalanced categories more effectively than interpolation-based methods such as SMOTE. By introducing mode-specific normalization where each continuous feature is modeled using

a variational Gaussian mixture model, CTGAN can generate more realistic and diverse synthetic records that are easier for neural networks to learn.

In addition, CTGAN employs a conditional generator that explicitly incorporates categorical variables into the training process. By conditioning the generator and discriminator on selected categorical values, and by using a log-frequency sampling strategy, the model ensures that minority categories are presented more frequently during training. This conditional sampling is combined with a cross-entropy loss that penalizes the generator when it fails to produce samples consistent with the conditioned category. Through this mechanism, CTGAN mitigates mode collapse and balances representation across rare classes, allowing it to generate synthetic data that reflects both continuous and categorical distributions faithfully. Empirical results demonstrate that CTGAN produces higher-quality tabular data than traditional GANs or Bayesian network–based models, particularly in cases with strong class imbalance.

As shown in Fig. 2, CTGAN model consists of three main parts, namely conditional vector, generator loss, and training-by-sampling. CTGAN introduces conditional vector (cond) which enables the model to explicitly condition on discrete variables during generation process. Each categorical variable $D_1$, ..., $D_{Nd}$ is transformed into a one-hot encoded vector $d_i = [d_i^{(k)}]$ for k = 1, ..., $|D_i|$. Alongside this representation, a corresponding mask vector $m_i^{(k)} = 1$ if i = i* and k = k* and $m_i^{(k)} = 0$ otherwise. The complete conditional vector is then constructed as:

$$cond = m_1 \oplus \ldots \oplus m_{Nd} \tag{2}$$

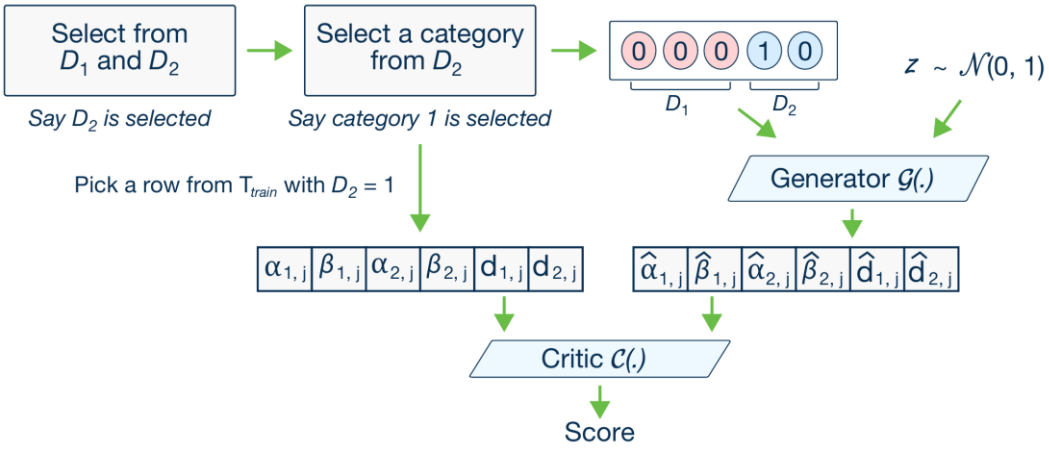



**Fig. 2.** CTGAN model [17]

During training, the conditional generator receives this cond vector together with noise input and attempts to produce synthetic samples that is consistent with the imposed condition. However, nothing in the forward pass prevents the generator from violating the condition. To ensure compliance, CTGAN introduces an auxiliary cross-entropy loss between the target mask vector and the generated categorical output, averaged across the batch. This penalization forces the generator to replicate the conditioning vector faithfully. Furthermore, CTGAN implements a training-by-sampling strategy, where conditional vectors are sampled according to the empirical distribution of categorical values in the real dataset. This ensures that both frequent and rare categories are adequately represented during training, allowing the discriminator to evaluate the divergence between the conditional distribution of generated samples. By integrating conditional vectors, cross-entropy regularization, and balanced sampling, CTGAN achieves high-quality generation of categorical data even in the presence of strong class imbalance.

In medical domain, CTGAN has been effectively used to balance skewed dataset such as generating generating synthetic cardiovasculat data to enhance prodictive models [7]. Additionally, CTGAN was also used to improve model accuracy and interpretability in cell signaling tasks with significant class imbalance [18]. In Parkinson's disease dementia studies, CTGAN outperformed traditional resampling methods like SMOTE in both AUC and F1 metrics, especially under extreme imbalance ratios [19]. A mobile healthcare study used CTGAN to synthesize cardiovascular tabular data, successfully preserving feature distributions while managing discrete imbalanced classes [7]. A 2025 benchmarking study across healthcare datasets including Breast Cancer Wisconsin, Lung Cancer, and CTG, confirming CTGAN's superior impact on classifier accuracy when compared with other GAN variants [8]. Furthermore, a combined CTGAN and decision classifier model (CTGAN-DC) significantly enhanced sensitivity and specificity in Kawasaki Disease diagnosis, addressing imbalance directly in a clinical setting [20].

### 2.3. SMOTE+GAN

Hybrid model SMOTE+GAN such as SMOTified-GAN is a two-phase oversampling model that utilise SMOTE and GAN where overgeneralized samples produced by SMOTE are transformed into more realistic distribution of data by GAN [1]. SMOTified-GAN applied transfer learning approach where GAN works on the output generated by SMOTE rather than generating the sample itself. The process of sample generation with SMOTified-GAN is shown in Fig 3.

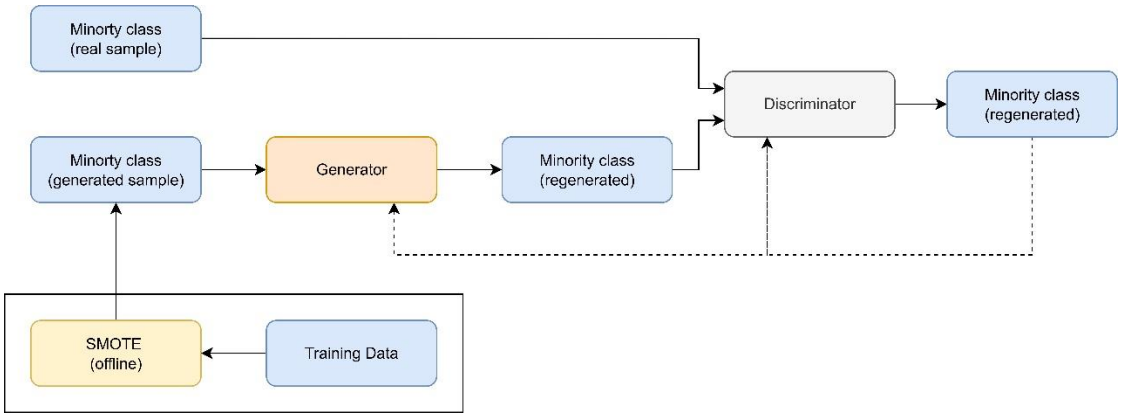



**Fig. 3.** Process of generating sample with SMOTified-GAN

### 2.4. Machine Learning

Supervised learning task such as classification plays a crucial role in evaluating the fidelity of synthetic or oversampled dataset. Decision tree (DT) which has interpretable structure is a classifier algorithm that create partitions of data from bigger set of data [21]. Decision tree's partition the feature space based on simple threshold rules, discrepancies in synthetic data quality often manifest as unstable or overly complex trees, which can indicate low fidelity in oversampling approaches [22]. Therefore, this method serve as a useful baseline to assess whether the oversampled data preserves the fundamental relationships within the original dataset. Apart from this, ensemble methods such as Random Forests (RF) provide a more robust way to test fidelity as it aggregates predictions from multiple decision trees trained on the subset of dataset. This approach reduce variance and highlights whether generated data preserve generalized patterns [23]. On the other hand, Extreme Gradient Boosting (XGBoost) is a classifier that excels in supervised learning tasks such as classification and regression that implements gradient-boosted decision tree which allow missing values handling automatically [8], [24]. This method offers a more

fine-grained evaluation since its sequential boosting mechanism is highly sensitive to distributional shifts.

### 2.5. Performance Metrics

To assess the performance of classifiers trained on either original or oversampled dataset, a comprehensive set of evaluation metrics was employed. Accuracy measures the overall proportion of correctly classified samples among all samples, providing a general sense of classifier performance across both majority and minority classes [25]. However, employing this evaluation metrics in imbalanced dataset can be misleading as it may be dominated by majority class. To address this, precision, recall, and F1-score are reported for each class separately. Precision quantifies the proportion of correctly predictive positive samples out of all predicted positives, whereas recall measures the proportion of correctly predicted positives out of all actual positives. Additionally, F1-score is defined as a metric that balances the trade-off between precision and recall and is particularly informative in scenarios with class imbalance [4], [26]. TO capture overall performance across classes, macro-averaged metrics compute the unweighted mean of precision, recall, and F1-score for all classes, treating each class equally regardless of its prevalence. In contrast, weighted-averaged metrics compute the number of samples in each class, providing a performance measure that reflects class imbalance.

## 3. Method

### 3.1. Dataset

We used the Framingham Heart Study dataset, a widely used clinical dataset for cardiovascular disease prediction. Several columns unrelated to the modeling task, such as patient IDs and time-to-event columns, were removed to focus on predictive features. Missing numerical values were imputed using the median of each column, and extreme outliers were filtered based on clinically reasonable thresholds for variables such as BMI, blood pressure, and glucose levels. The final dataset contains only relevant features for predicting cardiovascular disease (CVD) incidence.

### 3.2. Data Preprocessing

Prior to modeling, continuous features were standardised using z-score normalization. The dataset was stratified into training and testing sets with a 75%-25% split to preserve class distribution.

### 3.3. Augmentation Method

To address class imbalance in the dataset, we implemented three approaches: SMOTE, CTGAN, and a hybrid method called SMOTE+CTGAN.

- SMOTE

  SMOTE generates synthetic minority-class samples by interpolating between existing minority instances. For each minority sample, a set of k nearest neighbors is selected, and new samples are created along the line segments connecting the sample to randomly chosen neighbors. In our implementation, five nearest neighbors were used. SMOTE increases the decision region of the minority class, reducing bias toward the majority while maintaining sample diversity [4]. However, it may generate borderline or noisy samples if minority and majority classes overlap.

- CTGAN

CTGAN is a generative adversarial network designed for tabular data with mixed data types and imbalanced categories. The generator learns the conditional distribution of each feature given a specified discrete variable, producing realistic synthetic samples that preserve the correlation and distribution of original features. In our implementation, CTGAN was applied only to the minority class to generate additional synthetic instances, with parameters set to 50 epochs and batch size 500. This approach allows the augmentation of minority-class data while retaining statistical fidelity [17].

- SMOTE+CTGAN

  SMOTE+CTGAN is a hybrid approach that combines the strengths of SMOTE and CTGAN. First, SMOTE is applied to the minority class to provide an initial "jump start" by generating a small set of synthetic minority samples. This expanded minority set is then used to train a CTGAN model, which produces additional synthetic samples based on the enhanced minority distribution. The resulting synthetic data are merged with the original dataset, producing a more balanced dataset for training classifiers. This approach aims to reduce the generation of unrealistic samples while leveraging GAN's ability to capture complex distributions.

### 3.4. Classifiers

To evaluate the effectiveness of the proposed augmentation methods, three supervised classifiers were trained on the augmented datasets: Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). Decision Trees create partitions of the feature space and assign class labels to leaf nodes [22], Random Forests aggregate multiple decision trees to reduce variance and improve generalization, and XGBoost implements gradient-boosted decision trees with automated handling of missing values [24].

### 3.5. Evaluation Metrics

Classification performance was evaluated using a comprehensive set of metrics: accuracy, precision, recall, F1-score, and support for each class, as well as macro-averaged and weighted-averaged metrics to account for class imbalance. These metrics allow assessment of both overall performance and minority-class fidelity, which is critical when evaluating oversampling and synthetic data generation methods [4], [26], [27].

## 4. Results and Discussion

Table 1 presents the macro averaged metrics for three augmentation methods (SMOTE, CTGAN, SMOTE+CTGAN) and the imbalanced baseline, tested with three classifiers: Decision Tree (DT), Random Forest (RF), and XGBoost (XGB). Metrics reported include Accuracy, Macro Precision, Macro Recall, and Macro F1-score. The imbalanced baselines consistently achieved relatively high accuracy, ranging from 0.72 to 0.8. However, their macro recall and F1 values remained low (0.65–0.67), reflecting strong imbalance in class sensitivity. Among the augmentation methods, SMOTE delivered the best overall macro performance. For example, Random Forest and XGBoost with SMOTE achieved a macro F1 of 0.8484. In comparison, CTGAN and SMOTE+CTGAN produced consistent improvements over the imbalanced baseline but its score falls slightly below SMOTE by approximately 1-2%. These results indicate that while GAN-based augmentations enhanced balance, they did not surpass SMOTE in macro-level performance.

**Table 1.** Performance of different augmentation methods across classifiers

| Classifier | Augmentation | Accuracy | Macro F1 | Macro Precision | Macro Recall |
|---|---|---|---|---|---|
| Decision Tree | No | 0.7289 | 0.6489 | 0.6458 | 0.6529 |
| | SMOTE | 0.7790 | 0.7790 | 0.7790 | 0.7790 |
| | CTGAN | 0.7809 | 0.7749 | 0.7738 | 0.7764 |
| | SMOTE+CTGAN | 0.7880 | 0.7820 | 0.7811 | 0.7832 |
| Random Forest | No | 0.8072 | 0.6978 | 0.7631 | 0.6748 |
| | SMOTE | 0.8484 | 0.8484 | 0.8484 | 0.8484 |
| | CTGAN | 0.8424 | 0.8316 | 0.8515 | 0.8234 |
| | SMOTE+CTGAN | 0.8401 | 0.8290 | 0.8493 | 0.8208 |
| XGBoost | No | 0.7991 | 0.6975 | 0.7400 | 0.6786 |
| | SMOTE | 0.8484 | 0.8482 | 0.8505 | 0.8484 |
| | CTGAN | 0.8280 | 0.8175 | 0.8317 | 0.8110 |
| | SMOTE+CTGAN | 0.8303 | 0.8190 | 0.8317 | 0.8114 |

Table 2 shows the detail of class-specific performance, explicitly showing recall, precision, and F1-score for both the majority class (0) and minority class (1). In the imbalanced setting, all classifiers displayed extreme bias toward the majority class. For instance, Random Forest achieved a recall of 0.9413 for class 0 but only 0.4084 for class 1. Decision Tree and XGBoost showed similar patterns, with recall for class 1 falling below 0.5 despite majority recall higher than 0.80. This confirms that classifiers trained on imbalanced data highly biased toward the majority class.

Oversampling approaches markedly shifted the balance between classes. SMOTE consistently yielded the highest recall for the minority class, raising values to 0.7828 with Decision Tree, 0. 8528 with Random Forest, and 0. 8097 with XGBoost. This improvement, however, came at the cost of reduced majority recall, which dropped to the 0.77–0.88 range. In contrast, CTGAN and SMOTE+CTGAN achieved moderate improvements for the minority class (recall around 0.71–0.75) while maintaining relatively higher majority recall (0.80–0.90). This indicates that GAN-based augmentation provided a lighter trade-off, with smaller gains for the minority but less loss for the majority.

**Table 2.** Class specific performance metrics

| Classifier | Augmentation | Recall-0 | Recall-1 | Precision-0 | Precision-1 | F1-0 | F1-1 |
|---|---|---|---|---|---|---|---|
| Decision Tree | No | 0.8059 | 0.5000 | 0.8273 | 0.4642 | 0.8165 | 0.4814 |
| | SMOTE | 0.7751 | 0.7828 | 0.7812 | 0.7768 | 0.7782 | 0.7798 |
| | CTGAN | 0.8015 | 0.7513 | 0.8216 | 0.7260 | 0.8114 | 0.7385 |
| | SMOTE+CTGAN | 0.8108 | 0.7555 | 0.8257 | 0.7365 | 0.8182 | 0.7459 |
| Random Forest | No | 0.9413 | 0.4084 | 0.8255 | 0.7007 | 0.8796 | 0.5160 |
| | SMOTE | 0.8450 | 0.8528 | 0.8508 | 0.8460 | 0.8479 | 0.8489 |
| | CTGAN | 0.9310 | 0.7157 | 0.8239 | 0.8790 | 0.8742 | 0.7890 |
| | SMOTE+CTGAN | 0.9301 | 0.7115 | 0.8216 | 0.8769 | 0.8725 | 0.7856 |
| XGBoost | No | 0.9213 | 0.4360 | 0.8293 | 0.6507 | 0.8728 | 0.5221 |
| | SMOTE | 0.8870 | 0.8097 | 0.8235 | 0.8775 | 0.8541 | 0.8423 |
| | CTGAN | 0.9076 | 0.7143 | 0.8195 | 0.8440 | 0.8613 | 0.7738 |
| | SMOTE+CTGAN | 0.9183 | 0.7046 | 0.8162 | 0.8579 | 0.8643 | 0.7737 |

The results clearly demonstrate that classifiers trained on imbalanced medical data are heavily biased toward the majority class. Despite achieving high overall accuracy, their recall for minority outcomes

remains poor, which undermines their clinical usefulness. Oversampling methods effectively mitigated this problem in different ways. SMOTE provided the strongest gains in minority sensitivity, achieving near-balanced recall for both classes, but at the expense of majority performance. CTGAN and SMOTE+CTGAN improved minority recall while preserving higher majority recall, offering a more balanced trade-off.

In practical terms, if the goal of an application is to maximize detection of rare but critical outcomes, SMOTE appears to be the most effective strategy. However, when the preservation of majority class accuracy is also important, GAN-based augmentation offers a viable alternative. SMOTE+CTGAN further highlights the potential of hybrid approaches to combine the strengths of both interpolation-based and generative strategies. Overall, these findings align with the broader literature suggesting that while generative approaches hold promise for privacy-preserving and more realistic data augmentation, classical oversampling remains competitive and in some cases superior for improving classification performance in highly imbalanced medical datasets.

## 5. Conclusion

This study provides a systematic comparison of SMOTE, CTGAN, and SMOTE+CTGAN as augmentation strategies for imbalanced medical tabular data. Using the Framingham Heart Study dataset, we show that classifiers trained on imbalanced data exhibit strong bias toward the majority class, achieving recall above 0.94 for majority outcomes but failing to exceed 0.50 for minority outcomes. Augmentation with SMOTE produced the largest gains in minority recall, raising sensitivity to 0.78–0.85 across classifiers, and achieved the highest macro F1-scores. However, these gains were accompanied by reductions in majority recall, indicating a trade-off between minority sensitivity and overall balance. CTGAN and SMOTE+CTGAN produced more modest gains for minority recall while maintaining relatively higher majority recall, suggesting that generative methods may be preferable when preserving majority performance is clinically important. Taken together, our findings highlight that no single augmentation method universally outperforms the others. Instead, the choice of strategy should be tailored to the clinical application. For tasks prioritizing early detection of rare outcomes, SMOTE remains highly effective. For scenarios where both majority and minority predictions must remain stable, GAN-based or hybrid approaches offer more balanced trade-offs.

## References

[1] 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: https://doi.org/10.1109/ACCESS.2022.3158977.

[2] Y. Zhang *et al.*, "GAN-based one dimensional medical data augmentation," *Soft Comput.*, vol. 27, no. 15, pp. 10481–10491, Aug. 2023, doi: 10.1007/s00500-023-08345-z.

[3] E. Yuda, T. Ando, I. Kaneko, Y. Yoshida, and D. Hirahara, "Comprehensive Data Augmentation Approach Using WGAN-GP and UMAP for Enhancing Alzheimer's Disease Diagnosis," *Electronics*, vol. 13, no. 18, p. 3671, Sep. 2024, doi: 10.3390/electronics13183671.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/JAIR.953.

[5] H. Hairani, T. Widiyaningtyas, and D. Dwi Prasetya, "Addressing Class Imbalance of Health Data: A Systematic Literature Review on Modified Synthetic Minority Oversampling Technique (SMOTE) Strategies," *JOIV Int. J. Informatics Vis.*, vol. 8, no. 3, p. 1310, Sep. 2024, doi: 10.62527/joiv.8.3.2283.

[6] I. Goodfellow *et al.*, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020, doi: 10.1145/3422622.

[7] M. Alqulaity and P. Yang, "Enhanced Conditional GAN for High-Quality Synthetic Tabular Data Generation in Mobile-Based Cardiovascular Healthcare," *Sensors*, vol. 24, no. 23, p. 7673, Nov. 2024, doi: 10.3390/s24237673.

[8] H. A. Ahmed, J. A. Nepomuceno, B. Vega-Márquez, and I. A. Nepomuceno-Chamorro, "Synthetic Data Generation for Healthcare: Exploring Generative Adversarial Networks Variants for Medical Tabular Data," *Int. J. Data Sci. Anal.*, pp. 1–16, May 2025, doi: 10.1007/s41060-025-00816-w.

[9] T. Suresh, Z. Brijet, and T. D. Subha, "Imbalanced medical disease dataset classification using enhanced generative adversarial network," *Comput. Methods Biomech. Biomed. Engin.*, vol. 26, no. 14, pp. 1702–1718, Oct. 2023, doi: 10.1080/10255842.2022.2134729.

[10] M. Seibold, A. Hoch, M. Farshad, N. Navab, and P. Fürnstahl, "Conditional Generative Data Augmentation for Clinical Audio Datasets," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13437 LNCS, Springer, Cham, 2022, pp. 345–354, doi: 10.1007/978-3-031-16449-1_33.

[11] Bakti Putra Pamungkas, Muhammad Jauhar Vikri, and Ita Aristia Sa'ida, "Application of SMOTE-ENN Method in Data Balancing for Classification of Diabetes Health Indicators with C4.5 Algorithm," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 14, no. 2, pp. 183–188, May 2025, doi: 10.32736/sisfokom.v14i2.2350.

[12] M. Khairul Rezki, M. I. Mazdadi, F. Indriani, M. Muliadi, T. H. Saragih, and V. A. Athavale, "Application Of SMOTE To Address Class Imbalance In Diabetes Disease Classification Utilizing C5.0, Random Forest, And SVM," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 4, pp. 343–354, Aug. 2024, doi: 10.35882/jeeemi.v6i4.434.

[13] M. Syofian and I. Maulana, "Enhancing Obesity Risk Classification: Tackling Data Imbalance with SMOTE and Deep Learning," *J. Ris. Inform.*, vol. 6, no. 4, pp. 231–236, Sep. 2024, doi: 10.34288/jri.v6i4.349.

[14] Selly Anastassia Amellia Kharis, Melisa Arisanty, and Arman Haqqi Anna Zili, "Application of SMOTE in Multiclass Body Mass Index Classification:," *Proceeding Int. Semin. Sci. Technol.*, vol. 4, pp. 37–48, Apr. 2025, doi: 10.33830/isst.v4i1.5229.

[15] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," in *Lecture Notes in Computer Science*, vol. 3644, no. PART I, Springer, Berlin, Heidelberg, 2005, pp. 878–887, doi: 10.1007/11538059_91.

[16] S. Gholampour, "Impact of Nature of Medical Data on Machine and Deep Learning for Imbalanced Datasets: Clinical Validity of SMOTE Is Questionable," *Mach. Learn. Knowl. Extr.*, vol. 6, no. 2, pp. 827–841, Apr. 2024, doi: 10.3390/make6020039.

[17] L. Xu, M. Skoularidou, L. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 7335–7345. [Online]. Available at: https://dl.acm.org/doi/10.5555/3454287.3454946.

[18] M. E. Sánchez-Gutiérrez and P. P. González-Pérez, "Addressing the class imbalance in tabular datasets from a generative adversarial network approach in supervised machine learning," *J. Algorithm. Comput. Technol.*, vol. 17, Jan. 2023, doi: 10.1177/17483026231215186.

[19] G. Eom and H. Byeon, "Searching for Optimal Oversampling to Process Imbalanced Data: Generative Adversarial Networks and Synthetic Minority Over-Sampling Technique," *Mathematics*, vol. 11, no. 16, p. 3605, Aug. 2023, doi: 10.3390/math11163605.

[20] C.-S. Hung, C.-H. R. Lin, J.-S. Liu, S.-H. Chen, T.-C. Hung, and C.-M. Tsai, "Enhancing generalization in a Kawasaki Disease prediction model using data augmentation: Cross-validation of patients from two major hospitals in Taiwan," *PLoS One*, vol. 19, no. 12, p. e0314995, Dec. 2024, doi: 10.1371/journal.pone.0314995.

[21] M. A. Friedl and C. E. Brodley, "Decision tree classification of land cover from remotely sensed data," *Remote Sens. Environ.*, vol. 61, no. 3, pp. 399–409, Sep. 1997, doi: 10.1016/S0034-4257(97)00049-7.

[22] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/BF00116251.

[23] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[24] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-August-2016, pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785.

[25] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.

[26] Haibo He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[27] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," *PLoS One*, vol. 10, no. 3, p. e0118432, Mar. 2015, doi: 10.1371/journal.pone.0118432.