International Journal of Computer Techniques–IJCT Volume 13 Issue 2, March-April - 2026
Open Access and Peer Review Journal ISSN 2394-2231 https://ijctjournal.org/

An Analysis of Synthetic Data Generation Using GANs for Tabular Data

Sooraj S, Dr. Amrita Parashar

School of Computing Science and Engineering, VIT Bhopal University, Madhya Pradesh– 466114

Abstract - Class imbalance is a common problem in many real-world classification tasks where the minority classes contain important information but are underrepresented in the data. Classical ML methods are often prone to bias towards the majority class, and this leads to poor prediction results on the minority classes. The aim of this study is to assess the effectiveness of synthetic data generation in addressing the issue of imbalance between the two methods.Three different datasets of varying sizes were used: a small dataset on customer churn, a medium-sized Adult Census Income dataset, and a large dataset on financial transaction data. Classical machine learning approaches were first trained on the original imbalanced datasets.Subsequently, SMOTE and the three GAN-based models for synthetic tabular data generation, namely CTGAN, TVAE, and CopulaGAN, were used for synthetic minority sample generation. Precision, recall, F1-score, accuracy, and ROC AUC were used for evaluating the performance of the generated datasets. From the experiments, it was observed that SMOTE has the ability to improve minority class performance for small and medium-sized datasets. Although GAN-based models show promise for performance improvement, it was observed that such models are highly sensitive to dataset characteristics, such as dataset size and feature dimensionality, and preprocessing techniques. For large financial datasets, where feature variables are pre-scaled and anonymous, the performance of the GAN models was found to be poor in learning the underlying distribution of the dataset. Overall, it was observed that although GAN-based models show promise for performance improvement, it should be noted that traditional oversampling techniques are always reliable, and the performance of GAN-based models should be evaluated based on dataset characteristics.

Keywords: Class Imbalance, Oversampling, SMOTE, GAN-based Data Augmentation, Tabular Data, Minority Class Prediction.

1. Introduction

In supervised machine learning for applications in fraud detection, disease diagnosis, credit risk estimation, and customer behavior studies, managing unbalanced datasets has consistently been a challenge. In each of these applications, there is a majority class of events that outnumber high-value events and a minority class of events that are high-value but rare. As a result, supervised machine learning algorithms for unbalanced data frequently produce biased decision boundaries that predict that a data instance will belong to the majority class rather than the high-value minority class.

Since nearly all machine learning algorithms rely on the availability of a balanced class distribution, class imbalance significantly impairs the performance of conventional classifiers. The aforementioned problem is made even more severe in critical domains, where misclassifying members of the minority class can lead to major risk and financial issues. Therefore, in order to accurately estimate the performance level, the evaluation criteria in the context of classifiers should take into account the use of the recall, F1 measure, and ROC-AUC measure.

Various strategies have been documented in the body of existing literature to address the challenges associated with the issues of class imbalance.Some of these approaches fall under the category of data level techniques, which

carry out data level sampling strategies. Others fall under algorithm-level methods that carry out weighting at the class level. Others fall under the category of cost-sensitive education. The SMOTE technique is one of these approaches and is frequently employed because of its effectiveness. However, these SMOTE-based techniques use linear interpolation in the characteristic space, which could result in erroneous data samples.

By leveraging the power of synthetic data, new avenues for resolving issues related to class imbalance have been created using techniques in generative modeling. Generative Adversarial Networks and Variational Autoencoders are capable of producing synthetic data that resembles real data and also learns the actual distribution of the data. However, new variations like CTGAN, TVAE, and CopulaGAN have been proposed to effectively address such issues with increased success in imbalance learning tasks because tabular datasets present special challenges for conventional generating models.

For the task of tabular classification, the contribution of the present work is the methodical evaluation of the effectiveness of the data generation approaches with the use of GAN compared to the traditional approaches for handling the imbalance problem. For the purpose of the experiment, several real-world datasets with different sizes and features are considered, and the performance of the approaches has been evaluated with the use of the traditional performance metrics from the perspective of the classification task, particularly with the emphasis on the detection of the minority class. The main contribution of the present work is the useful information regarding the dataset dependency and the advantages and limitations of the approaches with the use of the GAN model.

2. Literature Review:

In the context of tabular classification problems, the current research has proposed a methodical assessment of GAN-based synthetic data generation approaches compared to other conventional imbalance handling strategies. In particular, the performance of CTGAN, TVAE, and CopulaGAN is evaluated along with some conventional strategies such as SMOTE and class-weighted learning. A number of real datasets with different sizes and features are used for experimentation, and performance is evaluated using conventional metrics from a classification perspective. The goal of this paper is to offer useful information about the dataset dependency, strengths, and limitations of GAN-based techniques for tabular data imbalance.

2.1 Classical Imbalanced Learning Approaches

Early research on unbalanced learning focused on resampling methods and cost-sensitive learning. By adding class weights or misclassification costs to the learning objective, cost-sensitive approaches penalize errors on minority classes more harshly. Despite their relative success, these techniques don't change the distribution of the underlying data and might not be effective when there is a lot of class overlap. While data-level techniques such as random undersampling and oversampling have also been widely used, they often suffer from overfitting or information loss, respectively.

The Synthetic Minority Oversampling Technique and its variants have gained popularity as synthetic oversampling methods. SMOTE avoids the creation of duplicate data by interpolating between existing instances in feature space to create new instances of the minority class, thus improving class balance. Extensions such as ADASYN and Borderline-SMOTE aim to improve sample generation near decision limits. While these methods are effective, they rely on local neighborhood structures and linear assumptions, which limits their ability to model complex feature dependencies and can result in noisy or unreal samples, especially on high-dimensional tabular datasets.

2.2 Generative Models for Tabular Data

Researchers are now trying to focus more on generative models to develop synthetic data in tables to overcome the disadvantages of interpolation-based methods. A special kind of generative models designed to learn from data distributions themselves is called Variational Autoencoder (VAE). TVAE extends the VAE

model to tabular data by specifically modeling mixed data types, and its training behavior is consistent. However, it has often been seen that VAE-based models create data samples that are too smooth, causing difficulty in distinguishing between classes during classification.

The ability of Generative Adversarial Networks (GANs) to effectively learn complex high-dimensional probability distributions has further made them a more effective alternative. Although various difficulties were discovered in the use of GANs, including training instability as well as the handling of categorical variables, the initial application of GANs-based approaches for tabular data revealed the viability of adversarial learning, beyond the image domain. Various GAN-based designs, such as the use of the CTGAN model, which relies on the use of conditional generation to assist in the effective simulation of discrete variables, have since been presented as having the ability to resolve the previously highlighted difficulties.

Another major direction in this domain is carried out by copula-based generative models. CopulaGAN is specially designed to handle feature dependencies while preserving marginal distributions with the use of adversarial learning and copulas. This sort of approach is particularly suited to financial and transactional data, given the effectiveness of this approach demonstrated by earlier research on the ability to maintain correlation conditions in numerical data. However, the availability of data and modeling dependencies are the key factors here.

2.3 GANs for Imbalanced Classification

The application of GAN-generated synthetic data, especially in imbalanced classification scenarios, has been studied extensively. When compared to traditional resampling strategies, it has been reported that GAN-based oversampling strategies have better recall values and F1-score values for minority classes in various domains, including intrusion detection systems and health analytics. It has been highlighted in some literature that classifiers, when trained on a union of synthetic and real data, often perform better compared to classifiers that use only real data. However, a significant portion of recent literature has concentrated on deep learning classifiers or dataset-specific classifiers, while few attempts have been made to comparatively analyze traditional oversampling strategies.

The rapid progress of synthetic tabular data creation techniques like GANs, VAEs, diffusion models, and combinations of these models is also reflected in the recent techniques. Although the generators like big language model-based generators and diffusion-based generators show promise with respect to data integrity and stability, their usage is mostly limited due to the complexities associated with the implementation process as well as the resources needed for the implementation process itself. Due to the excellent downstream utility of GAN-based models, the generators remain a popular choice.

2.4 Positioning of the Present Work

Even though substantial studies have reported notable achievements, it is still essential to carry out comprehensive empirical studies to compare balanced management strategies with GAN-based synthetic data generation methods with a comprehensive framework. It has been found that there is limited research that has assessed GAN-based oversampling strategies that make use of traditional machine learning methods for imbalanced datasets of varying sizes with complex features.

By systematically comparing class-weighted learning, SMOTE, and many GAN-based approaches, like CTGAN, TVAE, and CopulaGAN, on two different imbalanced data sets, this paper fills up those gaps. This paper succeeds in providing valuable insights into the pros and cons of generative oversampling methods for imbalanced tabular classification through an evaluation of classification performance via metrics like recall, F1 score, and ROC-AUC.

3. Problem Definition & Objectives:

While performing classification on real-world tabular data, it is quite common to observe extreme class imbalance problems. Here, we

usually observe occurrences of rare but significant events such as customer churn, low-income individuals, fraudulent transactions, or credit defaults belonging to the minority class. In such conditions, different types of machine learning models are found to improve accuracy on the bulk of the majority class transactions, thereby failing to detect minority class instances.

Traditional resampling techniques, e.g., Synthetic Minority Oversampling Technique (SMOTE), have been proposed to address this problem to some extent by creating synthetic minority samples through interpolation in the feature space. However, it is generally difficult for SMOTE, which has been found effective with moderate-sized datasets, to maintain the relationships between features, especially in large datasets with large dimensions, and sometimes the samples created by SMOTE can be noisy, which may result in low precision, especially when class boundaries are not clear.

Recent advances in deep generative models, specifically Generative Adversarial Networks (GANs) for tabular data such as CT-GAN, TVAE, and CopulaGAN, offer an alternative solution that learns the joint distribution of features and synthesizes more realistic samples. Several studies have been conducted on GAN-based resampling methods, yet there is a lack of empirical knowledge regarding the efficacy of the techniques in various datasets with different sizes and pre-processing conditions.

This research seeks to address the problem of evaluating the efficiency of both traditional and generative resampling methods in handling the problem of imbalance in tabular data classification at varying scales. To achieve this, this study aims at evaluating the effect of both dataset size (small, medium, and large) and feature preprocessing (unscaled features vs. scaled features) on the performance of both the SMOTE and GAN resampling techniques, which can be effective in conditional circumstances where they generalize well.

In this paper, the suitability, limitations, and stability of GAN-based resampling methods are investigated in-depth, through a systematic evaluation on various real-world datasets, by analyzing classification performance in terms of precision, recall, F1-score, accuracy, and ROC-AUC.

4. Dataset Description:

Four tabular real-world datasets of different sizes, feature compositions, and levels of class imbalance were selected to judge the effectiveness of traditional oversampling techniques and GAN-based data generation synthetically using these varying data scales and characteristics.

4.1 Dataset 1: Customer Churn Dataset

The Customer Churn dataset is used for building a predictor to determine the likelihood of a consumer discontinuing a service. This dataset contains 7,043 instances with 20 input features and one binary target feature.

The feature set includes a mixture of:

- **Demographic attributes** such as gender, senior citizen status, partner, and dependents,
- **Service-related attributes** including phone service, internet service, online security, streaming services, and technical support,
- **Account and billing attributes** such as tenure, contract type, payment method, monthly charges, and total charges.

The dataset has moderate class imbalance, containing 73.5% non-churners and 26.5% churners. The dataset, with its size and diversity in the types of features, represents the baseline scenario, which demonstrates the effects of the sampling/resampling strategies and the use of generative methods on imbalanced classification problems.

4.2 Dataset 2: Adult Census Income Dataset

The widely used socio-economic data, Adult Census Income, has been designed to predict if an individual's income is above a certain threshold for the year. It comprises 48,842 examples along with 14 input attributes and

one target variable.

The dataset comprises:

- **Numerical features** such as age, education number, hours per week, capital gain, and capital loss,
- **Categorical features** including work class, education, marital status, occupation, relationship, race, sex, and native country.

The data is moderately imbalanced, with the minority class having 24% of the instances (income > 50K), whereas the rest of the classes comprise 76% of the data. The current data set is of medium scale and can be used to observe the performance of resampling methods and GANs under moderate class imbalance with numerical and categorical data.

4.3 Dataset 3: Santander Customer Transaction Prediction Dataset

The dataset, Santander Customer Transaction, is a large transactional dataset that has been used as a benchmark to make predictions for rare customer events. It contains 200,000 examples, with each having 200 numerical features to be used as input variables, as well as a single binary output feature, making a total of 202 features.

All input features are:

- **Continuous numerical variables**,
- **Pre-scaled and anonymized**,
- Free from missing values,
- Abstractly named (e.g., var_0 to var_199), with no semantic interpretation.

The dataset is highly imbalanced, where the minority class in the dataset consists of approximately 10% of the total samples. The high dimensionality, anonymization, and pre-scaling of the data are challenging for the generative models, which can be considered a stress test for the approach.

4.4 Dataset 4: Home Credit Default Risk Dataset

The Home Credit Default Risk dataset is a practical finance dataset made publicly available by Home Credit Group. It comprises anonymized application data used to predict the default probability of the applicants. The dataset contains 307,511 instances with 120 input features and one output feature.

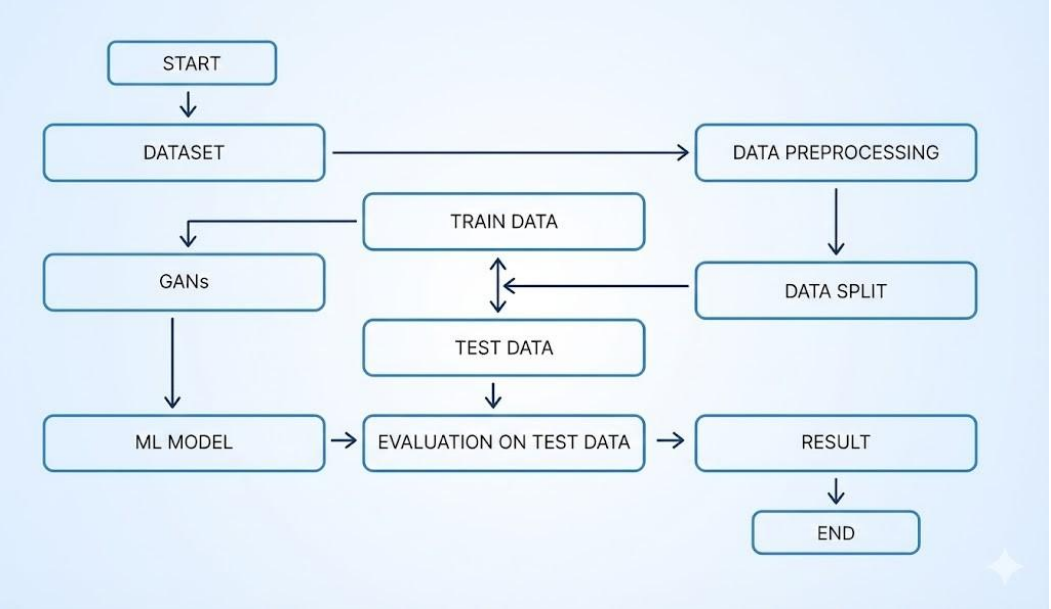
The selected features capture key aspects of:

- **Demographic and personal information** (e.g., gender, number of children, family size, age),
- **Employment and stability indicators** (e.g., days employed, occupation type, organization type),
- **Financial and loan-related attributes** (e.g., income, credit amount, annuity, goods price).

The classes in this dataset are highly unbalanced, as most of the rows, around 91-92%, operate as non-default classes, as indicated by TARGET = 0, while rows corresponding to default classes, as indicated by TARGET = 1, make up only 8-9% of the rows. This realistic dataset, due to its complexity, privacy, and uncaled heterogeneous features, is suitable for testing and validating the performance of GAN-based synthetic data generation with realistic financial data.

5. System Architecture:

The system architecture of this project aims to solve the class imbalance problem on tabular data in a systemic way through traditional machine learning techniques and GAN-based techniques to create synthetic data. The architecture of this project has been designed in a modular and sequential manner so that the data can be processed, models can be trained, synthetic data can be created, and tests can be carried out in a systemic way. Figure X presents the overall architecture of the system in this study.



5.1 Dataset Input Module

Approaches like these, which use real-world data, start by ingesting these datasets with imbalanced class distribution into the architecture. There are three scaled datasets which are to be considered as the significant inputs to this system. They are customer churn data, Adult Census Income data, and customer transaction data. These datasets vary depending on how large they are and to what extent their features are imbalanced.

Each data set will have a combination of numerical and categorical attributes and a binary attribute for the target variable, which represents the class label. The data should be able to load appropriately and be structured afterward, awaiting processing through the data set input module. This module is essentially the backbone for the entire pipeline since the data dictates the quality of the next process.

5.2 Data Preprocessing Module

Then the data is passed on to the data preprocessing component. The data preprocessing component is used for the preprocessing of the data for the learning models as well as the generative models. Data preprocessing components have procedures that include the management of the missing data in the system, the representation of the data, and the management of the contradictions in the data.

Categorical variables undergo binary mapping or one-hot encoding based on the nature of the variables, while the numerical variables remain the same unless transformations are required for a stable model. The output variable is converted into a binary format. Here, it is ensured that the dataset is clean and ready for usage by traditional ML models and GAN-based data synthesizers.

5.3 Data Splitting Module

Following the preprocessing step, the data is then divided into the training dataset and test dataset by using the data splitting method. The data is required to be stratified before it can be separated into the training dataset and test dataset, as the original balance of the data should be represented in the test dataset. This is an important step, especially in imbalanced learning.

The training data will be used only for the purpose of training the models and generating the data. The test data will be unused for any purpose and will be retained.

5.4 GAN-Based Synthetic Data Generation Module

At the core of the overall system architecture lies the GAN-based synthetic data generation

module, which gets activated upon the preparation of the provided training data. Minority class samples are segregated from the overall training dataset for training generative models, including but not limited to, CTGAN, TVAE, and CopulaGAN.

These models are used to learn the underlying distribution characteristics as well as the dependencies of features for a minority class. Furthermore, the generated synthetic data is such that it resembles the actual data. Also, the generated synthetic data gets combined with the actual training data to form a balanced set of data. It thus plays a significant role in tackling the problem of class imbalance while maintaining realistic data characteristics.

5.5 Machine Learning Model Module

The various balance datasets resulting from the GAN-based method and oversampling method are prepared for processing into the machine learning model component. Note that this latter component should support such classifiers as Logistic Regression and Random Forest.

The baseline models were trained with original imbalanced data to provide a standard bar for comparison. The settings of the models were kept exactly alike for all experiments; by doing so, it is ensured that the model performance variance is due to the way data is being resampled or generated and not because of bias introduced by the models themselves.

5.6 Model Evaluation Module

The models learned are tested against the holds out data, referred to as “test” data. This is achieved through the “Evaluation Module”. Various metrics can be used to measure the effectiveness of classification for the data. However, since the main focus of the project is to enhance the detection of the “minority class”, special attention is focused on the performance metrics of the “minority class”.

The evaluation mechanism employed here is similar for both the baseline, SMOTE, as well as GAN-based algorithms. Such a component can accommodate the quantitative evaluation performed to validate the effectiveness of each category of imbalance handling mechanisms.

5.7 Result Analysis and Output Module

The last module within the system architecture is the aggregation and analysis of results, which can be described by the term performance measures evaluated from different experiments. These performance measures are the results of different experiments performed and are used in the evaluation of techniques and trends.

These results are used to reach a conclusion to find out if the SMOTE method and the GAN-based techniques are applicable to different sizes and types of data. This module is helpful in reaching a conclusion based on the results obtained from experiments.

6. Methodology

This study adheres to a scientific approach in its experiments to compare the efficacy of traditional methods of oversampling and data generation techniques of GANs in handling class imbalance in tabular data. The experiment design includes data preprocessing, training of baseline models, handling of class imbalance using SMOTE and GAN-based methods, and evaluation of performance using standard metrics for classification tasks.

The aim is to compare all methods of handling class imbalance in a fair and comparable manner while splitting data into training and testing sets of varying sizes.

6.1 Dataset Selection

Three tabular datasets from real-world scenarios with varying sizes and imbalance ratios were chosen to perform experiments. These datasets are representative of various classification problems in real-world scenarios and are as follows:

- A **small-scale customer churn dataset**
- A **medium-scale Adult Census Income dataset**
- A **large-scale customer transaction dataset**

The datasets vary in terms of the number of samples, number of features, and severity of class imbalance. Testing various datasets enables the study to investigate the scalability of various oversampling methods with respect to various data distributions and sizes of datasets.

6.2 Data Preprocessing

Data preprocessing was done to ensure the quality of the data. The missing values in the data were handled using suitable imputation strategies for the features.

The features that are categorical are encoded using one-hot encoding or binary encoding schemes. The features that are numerical are kept in their original scale.

The target variable was converted into a binary classification variable for majority and minority classes.

To maintain class distribution during model evaluation, **stratified train–test splitting** was applied. Given a dataset DDD, the stratified split divides it into:

$$D = Dtrain \cup Dtest$$

where

$$Dtrain \cap Dtest = \emptyset$$

and the class distribution of the original dataset is preserved.

6.3 Baseline Model Training

Baseline models were trained using the original imbalanced datasets without applying any resampling techniques. Two widely used machine learning classifiers were selected:

- Logistic Regression
- Random Forest

These models are commonly used in tabular data classification due to their interpretability and robustness.

For Logistic Regression, the probability of a sample belonging to class 111 is given by the **sigmoid function**:

$$P(y = 1|x) = 1 + e - (\beta 0 + \beta Tx)1$$

where

- $\boldsymbol{x}$ represents the input feature vector
- β represents the model parameters

Random Forest works by aggregating predictions from multiple decision trees:

$$\hat{} = \{ 1( ), 2( ),..., ( )\}$$

where Ti(x) represents the prediction of the ithi^{th}ith decision tree.

These models are considered a reference point to assess the impact of class imbalance on minority class prediction performance.

6.4 Traditional Oversampling Using SMOTE

In the case of the class imbalance problem, the Synthetic Minority Oversampling Technique (SMOTE) was used for the training data set. This approach involves the generation of synthetic instances for the minority class.

For a given minority class sample xi and its corresponding nearest neighbor xnn, a synthetic sample xnew is created using the following formula:

xnew=xi+λ(xnn-xi)

where 0≤λ≤1

This technique helps to increase the minority class representation while avoiding duplication of data. Once the oversampling process has been completed, the classifiers are again trained and tested.

6.5 GAN-Based Synthetic Data Generation

In addition to SMOTE, three GAN-based tabular data generation models were used to generate synthetic minority class samples:

- **CTGAN**
- **TVAE**
- **CopulaGAN**

These models learn the distribution of the minority class data and generate realistic synthetic samples that mimic the statistical properties of the original data.

The standard **Generative Adversarial Network (GAN)** objective function is defined as:

$$GminDmaxV(D,G) = Ex{\sim}pdata(x)[logD(x)]$$

where

- D(x) represents the discriminator probability that a sample is real
- G(z) represents the generator producing synthetic samples from noise z

For **Variational Autoencoder based generation (TVAE)**, the loss function is defined as:

$$L = Eq(z|x)[logp(x|z)] - KL(q(z|x)\|p(z))$$

where

- q(z|x) represents the encoder distribution
- p(x|z) represents the decoder reconstruction probability
- KL denotes the Kullback–Leibler divergence

In the case of **CopulaGAN**, the joint distribution of features is modeled using copula functions:

$$(1, 2,\ldots,) = (1(1), 2(2),\ldots, ( ))$$

where C represents the copula function capturing dependencies among features.

The synthetic samples generated by these models were combined with the original training dataset to create balanced datasets used for classifier training.

6.6 Model Evaluation and Comparison

For the assessment of the performance of all the models, the same test data set was used to ensure a fair comparison of the models.

For the assessment of the performance of the models, classification performance metrics were used. These metrics include accuracy, precision, recall, F1 score, and ROC AUC.

Accuracy measures the overall correctness of the model by evaluating the proportion of the correctly classified data to all the data points classified by the model.

Precision measures the correctness of the positive class predictions and the proportion of the data points that were classified as positive to the actual positive data points.

Recall measures the correctness of the model in predicting the data points of the minority class.

F1 score measures the harmonic mean of the precision and the recall. It is a fair measure for the overall performance of the model when the false positive and false negative cases need to be considered.

Apart from these, ROC-AUC was utilized for the assessment of the overall classification performance of the models. ROC-AUC is based on the trade-offs between true and false positives based on the classification thresholds.

A comparative analysis was performed across all datasets in order to assess the impact of the baseline models, SMOTE, and GAN-based synthetic data creation techniques on the classification performance, especially with respect to minority class detection.

7. Evaluation Metrics:

To evaluate data utility, classification models were trained on synthetic data and tested on real data, and performance was measured using accuracy, precision, recall, F1-score, and ROC-AUC.

Method	Dataset 1		Dataset 2		Dataset 3		Dataset 4	
	Accuracy	F1 -Score	Accuracy	F1 -Score	Accuracy	F1 -Score	Accuracy	F1 -Score
Baseline	0.79	0.56	0.82	0.55	0/91	0.38	0.91	0.00
SMOTE	0.78	0.59	0.77	0.62	0.86	0.47	0.92	0.07
CTGAN	0.79	0.56	0.77	0.59	0.89	0.38	0.96	0.96
TVAE	0.79	0.56	0.74	0.59	0.57	0.29	0.96	0.95
CopulaGAN	0.79	0.59	0.77	0.60	0.52	0.27	0.96	0.95

In this work, Accuracy and F1-score are used as the primary evaluation metrics for comprehensive assessment of model performance on imbalanced classification tasks.

Accuracy describes the overall correctness of the predictions and informs about the general performance of the model. In the case of highly imbalanced data, accuracy as a single metric can be misleading because models may arrive at high accuracy by always predicting the majority class.

One way to improve on this limitation is by focusing on the F1-score, representing the harmonic mean for precision and recall of the minority class. This metric captures well the ability of a model to identify correctly the instances of the minority class, while its score balances false positives and false negatives.

The present study aims to evaluate the effect of different resampling methods, both conventional and generative, on the accuracy of the model and discrimination of minority classes by testing the models on four data sets of varying sizes and complexities. This would help in the robust comparison of the reliability of the models in handling imbalanced data.

8. Results and Discussion

The present study proposes an extensive assessment of the performance of base-level classifiers, oversampling methods such as SMOTE, and synthetic data approaches using GAN-based methods. Since the principal difficulty in handling imbalanced datasets is to correctly identify minority class instances, metrics such as Precision, Recall, and F1-score for minority class instances were employed to evaluate the performance of the methods.

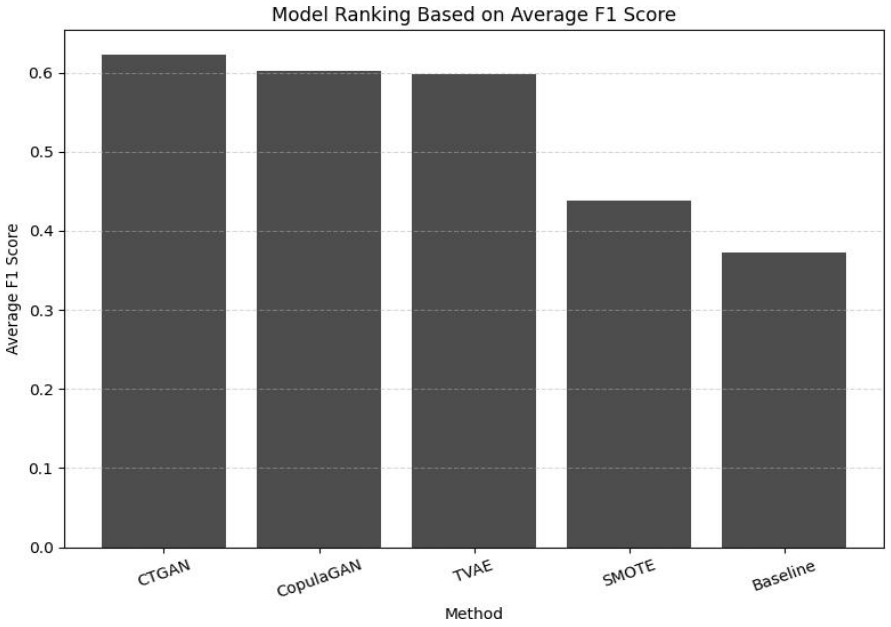


Figure 1.

Figure 1 presents the comparison of F1 scores obtained by the models on the four datasets, emphasizing the variability of model performance depending on the dataset's scale and the quality of feature representation.

8.1 Performance on Small-Scale Dataset (Customer Churn)

On the small-scale dataset for customer churn, the baseline model's F1 score for the minority class was 0.56, which, while good in terms of overall predictability, indicates the model's limitations in identifying churned customers, as indicated by the poor recall performance.

The use of SMOTE resulted in the recall performance of the minority class being improved to 0.61, with the F1 score being improved to 0.59, which indicates the effectiveness of the SMOTE technique on small-scale data, considering the possibility of interpolating the minority class data.

Of the models considered, CTGAN and TVAE had performance comparable to the baseline model, with minor improvements in the performance on the minority class. CopulaGAN improved the recall rate to 0.57 and had an F1 score of 0.59, comparable to the performance obtained with the use of the SMOTE technique.

However, as illustrated in Figure 1, there is not a significant difference in GAN-based models when it comes to small datasets. This is mainly due to the fact that limited training samples are not enough to allow generative models to effectively learn the underlying feature distribution.

Key Observation:

When it comes to small datasets, oversampling methods based on SMOTE are still considered to be more reliable compared to GAN-based methods.

8.2 Performance on Medium-Scale Dataset (Adult Census Income)

For the medium-scale Adult Census dataset, the baseline model had an F1-score of 0.55 on the minority class, indicating moderate prediction accuracy with low recall.

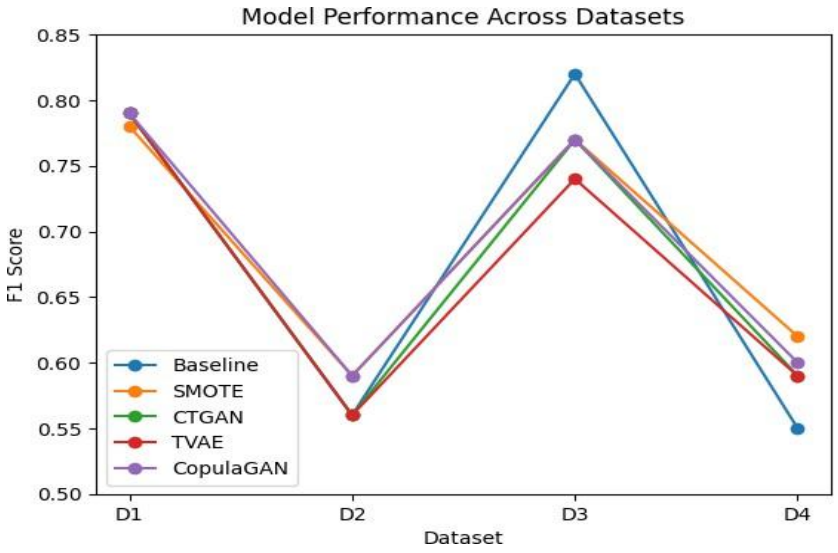


Figure 2.

On the other hand, the use of SMOTE has significantly boosted recall to 0.75. Although it has a slightly lower F1-score of 0.62 compared to the baseline model, it is evident that medium-scale datasets provide enough information to create synthetic data points with good accuracy.

GAN-based models were found to be competitive with other models on this dataset. Although CTGAN and CopulaGAN have almost equal F1-scores to SMOTE, TVAE has a higher recall with slightly lower precision.

As shown in Figure 1 above, it is evident that the performance gap between SMOTE and GAN-based models starts to decrease on medium-scale datasets. It is also evident that GAN-based models start to enjoy the advantage of data availability and learning of feature dependencies.

Key Observation:

For medium-scale datasets, GAN-based models compete with SMOTE. Although it is evident that SMOTE is more stable and efficient compared to GAN-based models, it is clear that GAN-based models have started to enjoy the advantage of data availability and learning of feature dependencies.

8.3 Performance on Large-Scale Dataset (Transaction Data – Pre-Scaled)

The high-dimensional data set posed another challenge due to its highly pre-scaled anonymized features. Although the classifier achieved high accuracy of 91%, its recall for the minority class was extremely low at 0.26.

The use of SMOTE resulted in better performance with respect to recall at 0.62 and F1-score at 0.47.

The performance of the GAN-based models was inconsistent. Although CTGAN achieved comparable performance to the baseline model with moderate improvements in recall values, TVAE and CopulaGAN showed unstable performance with high values of recall and low precision.

From the performance of the models shown in Figure 1, it is evident that the GAN family of models is unable to achieve better performance compared to traditional oversampling methods for handling data imbalance in this data set. This may be attributed to the loss of semantics in data features due to heavy pre-scaling and anonymization.

8.4 Performance on Large-Unscaled Dataset (Credit Risk Data)

For this large unscaled credit risk data set, severe class imbalance and complex high-dimensional feature distribution characteristics were observed. The baseline classifier showed high accuracy (92%) but was unable to learn instances of minority classes, resulting in zero recall and F1 scores.

For this data set, SMOTE showed marginal improvement over baseline, resulting in increased recall (0.04) and F1 scores (0.07). The accuracy remained similar to baseline. The low precision indicates that SMOTE's simple interpolation approach between instances may cause class boundaries to overlap in high-dimensional feature space.

For this data set, GAN-based model approaches showed significant improvement in classification performance. CTGAN showed high precision (1.00), and high recall (0.91), resulting in an F1 score of 0.96. TVAE and CopulaGAN showed high accuracy (96%) with high minority class detection capability.

This is clearly evident from Figure 1, where GAN-based model approaches significantly outperform baseline and SMOTE on this large data set with preserved feature distribution.

Observation:

GAN-based model approaches perform exceptionally well on large data sets with preserved feature distributions.

8.5 Overall Comparative Analysis

Overall, for all datasets, certain trends have been observed.

For small-scale datasets, SMOTE performs better compared to GAN-based methods due to low data diversity for training GAN models.

For medium-scale datasets, GAN-based methods start performing better compared to SMOTE since data is sufficient for learning relationships between features.

For large pre-scaled datasets, SMOTE performs better compared to GAN-based methods since the latter face difficulties in learning due to distorted feature data.

For large unscaled datasets, GAN-based methods perform better compared to both baseline and SMOTE methods in terms of minority class detection.

The above trends have been represented in Figure 2 in terms of the average F1-score ranking of all models for all datasets. From the graph, it is clear that CTGAN performs better

compared to CopulaGAN and TVAE. SMOTE and baseline methods perform with relatively low performance compared to GAN-based methods.

8.6 Computational Cost and Stability Considerations

Although GAN-based methods show high performance for larger datasets, they consume much more computational resources and time in comparison to SMOTE. Moreover, training GAN models is associated with some challenges concerning instabilities in training. On the other hand, SMOTE is computationally efficient, deterministic, and simple to use. It is scalable to larger datasets and is associated with performance improvements.
In conclusion, despite showing performance improvements over some cases, GAN-based methods should be used after taking into account the computational complexities associated with them.

9. Conclusion and Future Work:

This project provided a thorough investigation of the classical oversampling techniques and GAN-based synthetic data generation techniques for addressing the imbalance problem while performing tabular classification tasks. The investigation was carried out on three real-world datasets with different sizes, including the customer churn dataset (small-scale dataset), Adult Census Income dataset (medium-scale dataset), and customer transaction dataset (large-scale dataset).

From the experiment results, the accuracy of the machine learning models is not sufficient because the accuracy is biased to the majority class in the unbalanced dataset. The inadequacy of the accuracy metric further validates the inadequacy of the accuracy metric in the imbalanced learning scenario. The importance of the recall metric and F1-score in the performance assessment is emphasized.

Three of the popular GAN-based synthetic data generation methodologies were employed, i.e., CTGAN, TVAE, and CopulaGAN. Overall, the performance of the three models indicated great potential on medium-sized datasets considering the array of feature types. Nevertheless, the performance of each model was found to be extremely sensitive in relation to the dataset, dimension of the features, and the preprocessing strategy. When applied on the large transaction data set, the availability of pre-scaled features restricted the capacity of the GAN model in learning data distributions, thereby compromising the classification performance.

Overall, the results demonstrate that, unlike popular perceptions, GAN-based oversampling is not categorically better than other techniques, especially considering that it is quite dataset-dependent, with much consideration being taken when accounting for the representation of the dataset and their respective scales. This study has therefore emphasized or made it evident that the selection of imbalance handling techniques, with consideration of datasets, should not be taken lightly, and this observation pertains to the learning scheme referred to as synthetic data generation techniques.

10. References

1. Figueira, A.; Vaz, B. Survey on Synthetic Data Generation, Evaluation Methods and GANs. *Mathematics* **2022**, *10*, 2733. https://doi.org/10.3390/math10152733

2. Abedi, Masoud, Lars Hempel, Sina Sadeghi, and Toralf Kirsten. 2022. "GAN-Based Approaches for Generating Structured Data in the Medical Domain" *Applied Sciences* 12, no. 14: 7075. https://doi.org/10.3390/app12147075

3. Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, Debbie Rankin,
Synthetic data generation for tabular health records: A systematic review,Neurocomputing,Volume 493,2022,
(https://www.sciencedirect.com/science/article/pii/S0925231222004349)

4. Generation of synthetic full-scale burst test data for corroded pipelines using the tabular generative adversarial network,
Engineering Applications of Artificial Intelligence,
Volume 115,
2022,
(https://www.sciencedirect.com/science/article/pii/S0952197622003529)

5. Generation of synthetic full-scale burst test data for corroded pipelines using the tabular generative adversarial network,
Engineering Applications of Artificial Intelligence,
Volume 115,
2022,
(https://www.sciencedirect.com/science/article/pii/S0952197622003529)

6. Achuthan, S., Chatterjee, R., Kotnala, S. *et al.* Leveraging deep learning algorithms for synthetic data generation to design and analyze biological networks. *J Biosci* 47, 43 (2022). https://doi.org/10.1007/s12038-022-00278-3

7. Mohammad Esmaeilpour, Nourhene Chaalia, Adel Abusitta, François-Xavier Devailly, Wissem Maazoun, Patrick Cardinal,
Bi-discriminator GAN for tabular data synthesis, Pattern Recognition Letters,
Volume 159,
2022,
(https://www.sciencedirect.com/science/article/pii/S0167865522001830)

8. TabFairGAN: Fair Tabular Data Generation with Generative Adversarial Networks
Amirarsalan Rajabi, Ozlem Ozmen Garibay
https://doi.org/10.48550/arXiv.2109.00666

9. Dmitry Anshelevich, Gilad Katz,
Synthetic tabular data generation using a VAE-GAN architecture,
Knowledge-Based Systems, Volume 326,
2025,
(https://www.sciencedirect.com/science/article/pii/S0950705125010421)

10. G.Charbel N. Kindji, Lina M. Rojas-Barahona, Elisa Fromont, Tanguy Urvoy,
Tabular data generation models: An in-depth survey and performance benchmarks with extensive tuning,
Neurocomputing,
Volume 658,
2025,
(https://www.sciencedirect.com/science/article/pii/S0925231225023276)

11. Jian'en Yan, Haihui Huang, Kairan Yang, Haiyan Xu, Yanling Li,
Synthetic data for enhanced privacy: A VAE-GAN approach against membership inference attacks,
Knowledge-Based Systems,
Volume 309,
2025,
(https://www.sciencedirect.com/science/article/pii/S0950705124015338)

12. Subhajit Chatterjee, Debapriya Hazra, Yung-Cheol Byun,
GAN-based synthetic time-series data generation for improving prediction of demand for electric vehicles,
Expert Systems with Applications, Volume
264,
2025,
(https://www.sciencedirect.com/science/article/pii/S0957417424027052)

13. Lee, S., & Min, M. (2025). CG-TGAN: Conditional Generative Adversarial Networks with Graph Neural Networks for Tabular Data Synthesizing. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(17), 18145-18153. https://doi.org/10.1609/aaai.v39i17.33996

14. Saifur Rahman, Shantanu Pal, Shubh Mittal, Tisha Chawla, Chandan Karmakar,
SYN-GAN: A robust intrusion detection system using GAN-based synthetic data for IoT security,
Internet of Things, Volume
26,
2024,
(https://www.sciencedirect.com/science/article/pii/S2542660524001537)

15. Muhammad Ahtazaz Ahsan, Amna Arshad, Adnan Noor Mian,
Leveraging tabular GANs for malicious address classification in ethereum network,
Computer Networks,
Volume 254,
2024,
(https://www.sciencedirect.com/science/article/pii/S1389128624006455)

16. Alex X. Wang, Stefanka S. Chukova, Colin R. Simpson, Binh P. Nguyen,
Challenges and opportunities of generative models on tabular data,
Applied Soft Computing,
Volume 166,
2024,
(https://www.sciencedirect.com/science/article/pii/S1568494624009979)

17. Vasileios C. Pezoulas, Dimitrios I. Zaridis, Eugenia Mylona, Christos Androutsos, Kosmas Apostolidis, Nikolaos S. Tachos, Dimitrios I. Fotiadis,
Synthetic data generation methods in healthcare: A review on open-source tools and methods, Computational and Structural Biotechnology Journal,
Volume 23,

2024,
(https://www.sciencedirect.com/science/article/pii/S2001037024002393)
18. Saifur Rahman, Shantanu Pal, Shubh Mittal, Tisha Chawla, Chandan Karmakar,
SYN-GAN: A robust intrusion detection system using GAN-based synthetic data for IoT security,
Internet of Things, Volume
26,
2024,
(https://www.sciencedirect.com/science/article/pii/S2542660524001537)
19. Nasimov, R.; Nasimova, N.; Mirzakhalilov, S.; Tokdemir, G.; Rizwan, M.; Abdusalomov, A.; Cho, Y.-I. GAN-Based Novel Approach for Generating Synthetic Medical Tabular Data. *Bioengineering* **2024**, *11*, 1288.
https://doi.org/10.3390/bioengineering11121288 20.Ha Ye Jin Kang, Erdenebileg Batbaatar, Dong-Woo Choi, Kui Son Choi, Minsam Ko, Kwang Sun Ryu,
Synthetic Tabular Data Based on Generative Adversarial Networks in Health Care: Generation and Validation Using the Divide-and-Conquer Strategy,
JMIR Medical Informatics, Volume
11,
2023,
(https://www.sciencedirect.com/science/article/pii/S2291969423000571)
21. Hajra Murtaza, Musharif Ahmed, Naurin Farooq Khan, Ghulam Murtaza, Saad Zafar, Ambreen Bano, Synthetic data generation: State of the art in health care domain,
Computer Science Review, Volume
48,
2023,
(https://www.sciencedirect.com/science/article/pii/S1574013723000138)
22. Aryan Pathare, Ramchandra Mangrulkar, Kartik Suvarna, Aryan Parekh, Govind Thakur, Aruna Gawade,
Comparison of tabular synthetic data generation techniques using propensity and cluster log metric,
International Journal of Information Management Data Insights,
Volume 3, Issue 2,
2023,
(https://www.sciencedirect.com/science/article/pii/S2667096823000241)
23. Fonseca, Joao & Bação, Fernando. (2023). Tabular and latent space synthetic data generation: a literature review. Journal of Big Data. 10. 10.1186/s40537-023-00792-7.
24. Y. Zhang, N. A. Zaidi, J. Zhou and G. Li, "GANBLR: A Tabular Data Generation Model," 2021 IEEE International Conference on Data Mining (ICDM), Auckland, New Zealand, 2021, pp. 181-190, doi: 10.1109/ICDM51629.2021.00103.