*2025 International Conference on Computing Technologies & Data Communication (ICCTDC)*
*Hassan, India. July 04 - 05, 2025*

# Synthetic Data Generation for Enhanced Model Efficiency: A GAN- Based Approach to Tabular Data

1st Sunil Sangve
*Vishwakarma Institute of Technology*
*Pune, India*
sunil.sangve@vit.edu

2nd Kriya Oswal
*Vishwakarma Institute of Technology*
*Pune, India*
kriya.oswal22@vit.edu

3rd Tejas Sarade
*Vishwakarma Institute of Technology*
*Pune, India*
shriram.tejas22@vit.edu

4th Omkar Tongare
*Vishwakarma Institute of Technology*
*Pune, India*
ashok.omkar22@vit.edu

5th Nilay Sangode
*Vishwakarma Institute of Technology*
*Pune, India*
nilay.sangode22@vit.edu

6th Qusai Shergardwala
*Vishwakarma Institute of Technology*
*Pune, India*
qusai.shergardwala22@vit.edu

***Abstract*—The generation of synthetic data has emerged as a promising approach to enhance the performance of machine learning models, particularly in scenarios with limited original data. This study focuses on using Conditional Tabular Generative Adversarial Networks (CTGANs) for synthetic tabular data generation. The synthetic tabular data produced is supposed to have the same statistical properties as the original data. The aim of this research is to examine the effect of augmented datasets that combine original and synthetic data on the performance of several machine learning models, namely Logistic Regression, Random Forest, and Gradient Boosting Machines. Experimental results reveal how model accuracy, precision, recall, and F1-score vary with the inclusion of synthetic data, thus giving insights into its effectiveness in improving model generalization. Statistical tests, like paired t-tests, validate the significance of performance differences, and experiments varying the proportion of synthetic data provide guidelines for its optimal usage. This work highlights the potential and limitations of synthetic data augmentation in practical machine learning applications. Besides, this paper is based on previous work that applied CTGANs to mitigate data imbalance in machine learning.**

***Index Terms*—Synthetic Data, CTGAN, Tabular Data Augmentation, Machine Learning, Model Accuracy, Data Generalization, Synthetic Data Proportion, Statistical Validation, Data Augmentation, Generative Adversarial Networks (GANs).**

## I. Introduction

Machine learning (ML) has become integral to many domains, pushing forward the frontiers of applications in healthcare, finance, and fraud detection. However, the effectiveness of ML models is often impeded by the quality and distribution of real-world datasets, which often suffer from issues like class imbalance, data sparsity, and privacy constraints. Class imbalance is particularly critical in domains such as medical diagnosis and credit scoring, where minority classes—such as rare diseases or defaulted

loans—are underrepresented. This imbalance biases models toward majority classes, reducing their ability to accurately predict minority class outcomes, often resulting in significant real-world consequences. Synthetic data generation has been found to be a viable solution to overcome these challenges by either augmenting existing datasets or creating new ones, which enables the training of more robust ML models. Traditional oversampling techniques, such as SMOTE, have been widely used but often perform poorly on high-dimensional or mixed-type data commonly found in structured datasets. Recently, Generative Adversarial Networks have gotten much attention owing to their abilities to produce synthetically good-quality data. Majority of GANs' work is with the unstructured tabular data-like images. Therefore, there isn't much studied work about it in the tabular structured case with mixed type features: categorical and numerical. This paper uses the state-of-the-art GAN model, Conditional Tabular Generative Adversarial Networks (CTGAN), to generate synthetic tabular data and assess its impact on the performance of ML models. CTGAN is specifically designed to address the challenges associated with tabular data, including its complex feature dependencies and mixed data types. By generating synthetic samples, we aim to enhance the identification of minority classes and assess how synthetic data influences model accuracy across various scenarios. Our study includes generating synthetic datasets using CTGAN, training multiple ML models on both original and augmented data, and analyzing changes in performance metrics such as accuracy. The results are validated across diverse datasets, including those from credit scoring and fraud detection, and benchmarked against traditional oversampling techniques. This work shows the potential of GAN-based approaches to class imbalance and improved performance of ML models in critical domains, with valuable insights for researchers and practitioners..

## II. LITERATURE REVIEW

Xu et al. presented CTGAN, a Conditional Generative Adversarial Network aimed at overcoming some of the major

979-8-3315-2798-3/25/$31.00 ©2025 IEEE

challenges in synthesizing tabular data, like mixed data types, multimodal distributions, and imbalanced categorical columns. Through mode-specific normalization and a conditional generator, it surpassed Bayesian networks and other deep learning methods on benchmarks that contained 7 simulated and 8 real-world datasets. The results shown that the system of CTGAN is a powerful remedy for producing realistic tabular data synthesis, capturing complex data distributions but providing high-quality synthetic samples [1]. Notably, to overcome class imbalance in supervised learning, namely credit scoring, Engelmann and Lessmann devised a conditional Wasserstein GAN: cWGAN. Unlike the conventional techniques like SMOTE, cWGAN, while maintaining high-quality synthetic samples for the minority class, easily deals with both numerical and categorical data. Their new method, using an auxiliary classifier loss, which outperforms the existing oversampling strategy in comparison with SMOTE and Random Oversampling, succeeded on seven real-world datasets, demonstrating its effectiveness in unbalanced learning tasks [2]. The present study discusses various techniques for developing synthetic data focusing on how synthetic data can benefit machine learning performances, especially under circumstances where there is a limit to the amount of data in use due to security or privacy concerns. This paper shows synthetic data techniques for the healthcare industry and how these techniques work out to alleviate the problem of scarcity while preserving privacy-sensitive information. Further, this research explores the role of data augmentation in moving forward machine learning applications [3] To deal with privacy and utility issues, this research provides an improved version of conditional GAN (CGAN) that can generate synthetic data (SD) from actual datasets containing both numerical and categorical features. Different from previous research, it proposes an experiment-based evaluation of SD, measuring its utility, privacy, and quality against actual dataThe study shows that SD is feasible for data sharing that protects privacy by providing insights into features like distributions and correlations [4].Synthetic data creation approaches are dealt with by Figueira and Vaz, who take specific interest in GAN-based methods for tabular data. In addition to outlining the challenges in synthetic tabular data generation, including mixed-type data management, non-Gaussian distribution, and imbalance categories, they further discuss significant GAN architectures: TGAN and CTGAN. Besides summarizing the important findings and gaps in research of this area, the paper also discusses the assessment methodology for synthetic data quality. This work is a primary source for synthetic data and GAN researchers [5]. The work is an attempt to find out how hard it is to identify money laundering and fraud in online games when quality data is rare, and class imbalance gives large hurdles. The authors present a novel GAN framework named SDG-GAN that will generate synthetic data for better training of the classifier. It boosts classification performance on benchmark and real-world gambling fraud datasets better than density-based oversampling techniques. This study shows how GANs can overcome class imbalance and data scarcity in fraud detection [6]. This paper discusses how to synthesize synthetic training data for machine learning problems with Generative Adversarial Networks (GANs). GAN-generated data are helpful in reducing dependency on original datasets and resolving imbalanced datasets, while safeguarding sensitive information. for example, in medical data. Experiments on benchmark datasets demonstrate the effectiveness of GANs for data augmentation and privacy preservation; Decision Tree classifiers trained on GAN-generated data attain equivalent, and sometimes superior, accuracy and recall compared to those trained on the original data [7]. This paper introduces the architecture for private data sharing and synthetic data generation designed for fraud detection, Duo-GAN. Duo- GAN models fraudulent and legal financial transactions using two generators of GAN, which balance data. The synthetic data remains useful as the original datasets while maintaining efficacy and privacy. The experiments demonstrate that the proposed framework indeed promises privacy-preserving and utility-retaining data sharing, as the classifiers trained on the synthetic data reach F1 scores whose difference to those trained on the real data is no more than 5Emphasis was placed on deep generative models and topologies of neural networks. Challenges and opportunities for future research, alongside privacy and fairness issues associated with the production of synthetic data, are brought up in this discussion. In order to drive more research into synthetic data creation, this effort strives to push forward the study of synthetic data creation [13]. This paper presents a comprehensive review of methods to generate synthetic data to address the problems of scarcity of data, privacy issues, and algorithmic biases in machine learning. Among the techniques reviewed are large language models, generative adversarial networks, and variational autoencoders. The report presents the pros and cons of the following technologies such as computing requirements, training stability, and keeping the privacy level. The authors claim that removing the abovementioned constraints will push further adoption of the synthetic data creation techniques that promote advancement in machine learning and data-driven solutions [14].The data set in this article includes information on 238 SARS-CoV-2 positive patients who were admitted to the University Hospital of Brussels, Belgium in 2020 to demonstrate the procedure of synthesizing data by Conditional Tabular Generative Adversarial Network. Data quality and consistency were ensured using a multidisciplinary approach called TIMA, and after 100,000 epochs of training, 10,000 synthetic records were generated. The synthetic dataset was highly representative with a median correlation similarity score of 0.97 and a contingency score of 0.94 for continuous variables. Principal Component Analysis was done to verify diversity, and novelty got a high score of 1. Amazingly, the tima committee accepted almost all synthetic data as authentic. These results reveal how efficient and reliable CTGAN is in producing diversified, realistic, and innovative synthetic datasets [15]. This article shows how to create synthetic data using a CTGAN based on data from 238 SARS- CoV-2 positive patients admitted to the University Hospital of Brussels, Belgium, in 2020. The TIMA approach was used

to ensure data quality and consistency, and after 100,000 training epochs, 10,000 synthetic records were produced. The synthetic dataset indicated strong representativeness with a median correlation similarity score of 0.97 and a contingency score of 0.94 for continuous variables. Principal Component Analysis was adopted to confirm variety, and novelty scored a surprisingly high score of 1. Interestingly, nearly all the data in the synthetic data were adjudged authentic by the TIMA committee. These results showcase the reliability and effectiveness with which CTGAN can produce diverse, realistic, and innovative synthetic datasets [16]. The research also proposes a paradigm for assessing synthetic data generators based on four criteria: attribute fidelity, bivariate fidelity, population fidelity, and application fidelity. To assess the utility of four state-of-the-art data synthesizers over 19 datasets of varied sizes and complexity, a sample statistic is selected from each class. To speed up the assessment of synthetic data utility, the paper identifies discrepancies in previous evaluation metrics and explores relationships between the selected metrics. These results provide a systematic approach for comparison of artificial data generators [17]. This study investigates synthetic data as a potential remedy for issues with bias in AI/ML training, privacy protection, and data accessibility. The usefulness and privacy-protecting potential of synthetic data have drawn attention in light of strict privacy regulations and concerns about bias and data quality. The study identifies research needs by reviewing synthetic data creation techniques and related validation criteria. Although synthetic data holds promise, the study raises issues that require further research to prove it is a universal answer for AI and ML applications [18]. This study explores methods of creating synthetic data with an emphasis on maintaining the features of the original datasets while safeguarding user privacy. Unlike models trained on natural data, synthetic data tries to enable machine learning activities without affecting performance noticeably. The paper empirically evaluates the usefulness of synthetic data over a number of supervised learning tasks using publicly available datasets, focusing on its effectiveness in protected privacy data sharing[19]. A common burden in health research, logistic regression prediction models are evaluated in this paper using utility criteria for contrasting approaches to synthetic data generation. Here, three different approaches to synthetic data generation—the Bayesian network, the GAN, and sequential tree synthesis—were evaluated across six utility criteria on 30 health datasets. Based on a Gaussian copula representation, the multivariate Hellinger distance was found to be the best accurate metric for evaluating SDG approaches according to prediction performance. When evaluating alternative SDG approaches on the same dataset, this validated metric provides a reliable technique [20].

## III. METHODOLOGY

### A. Acquisition and Preprocessing of Datasets

A tabular dataset consisting of a mix of numerical and categorical variables is chosen for this study. In order to make the dataset compatible with the CTGAN model and reliable enough to generate synthetic data, several preprocessing steps are applied before it. Among these preprocessing steps, the missing values were first addressed using statistical imputation methods, which used mean values for numerical features and the most frequent category for categorical features. It provided a complete data set without incorporating bias. Second, continuous variables were normalized in a common range to avoid numerically unstable results during model fitting. The transformation of categorical features was done via label encoding such that they get represented as integer values suitable for the CTGAN framework. Some outliers were caught and removed statistically to preserve integrity in the dataset and avoid any skewed distributions from impeding learning by the generative model.

### B. Synthetic Data Generation Using CTGAN

To create synthetic data, a Conditional Tabular Generative Adversarial Network or CTGAN is utilized. Especially when working with mixed data types, such as numerical and categorical features, CTGAN is pretty well-suited for tabular data. Using the preprocessed dataset as training data, it learns to generate synthetic data closely resembling that of the original data.

The CTGAN model is based on two principal components: the generator and discriminator, which are trained adversarially. A generator creates synthetic data samples based on patterns learnt from the real dataset, using random noise as an input. The discriminator, a binary classifier, determines if each sample is real or synthetic. Through this adversarial training, the generator iterates to produce the data closest possible to the original dataset, whereas the discriminator learns to differentiate and distinguish between actual and synthetic data. Hyperparameters need to be tuned to provide optimal performance. Some of these parameters include the batch size, learning rate, and number of epochs. Preliminary results indicate how these values should be optimized.

### C. CTGAN Loss Function

a) The key to training are the loss functions, a set of formulas that tell each network how to improve after each training iteration (or epoch). Each of the discriminator and generator has its own loss values. Epoch after epoch, these networks learn by trying to minimize their loss function.The discriminators loss function is defined as:

$$\mathcal{L}D = -\frac{1}{m}\sum i = 1^m \left[D(x'^{(i)}) - D(x^{(i)})\right]$$

Where $x(i)$ is a real data sample, $x'(i)$ is a synthetic data sample, $D(x)$ is the output of the discriminator for input $x$, and $m$ is the total number of data samples.

b) The generators Loss function is defined as:

$$\mathcal{L}G = -\frac{1}{m}\sum i = 1^m \left[D(x'^{(i)})\right] + H \tag{1}$$

Where H: Entropy term used to stabilize training and D(x(i)): Discriminator's output for synthetic data sample x(i)

### D. Validation of Synthetic Data

Statistical tests, namely the Kolmogorov-Smirnov test, have been used to quantify the similarity in the feature distribution of both the real and the synthetic datasets. The Kolmogorov-Smirnov statistic is given by:

$$\text{KS} = \max |F_{\text{real}}(x) - F_{\text{synthetic}}(x)|$$

Where F real and F synthetic are the cumulative distribution functions of the real and synthetic data, respectively. Besides, correlation matrices for both datasets were calculated and their differences measured as:

$$\text{Correlation Difference} = |C_{\text{real}} - C_{\text{synthetic}}|$$

### E. Experimental Configuration

Three experimental setups were designed to evaluate the effectiveness of synthetic data. In the first setup, machine learning models were trained and tested only on the original dataset. The second setup involved training and testing models on the synthetic dataset generated by CTGAN. In the third setup, the training was performed using an augmented dataset that was obtained by combining the original and synthetic datasets, whereas testing was carried out only on the original test set. This enabled the authors to comprehensively analyze how synthetic data impacts model performance. The datasets were divided into training and testing subsets with an 80:20 split. This ensured a consistent and unbiased evaluation framework across all experiments. They were calculated as:
Training Size=0.8×Total dataset size
Testing Size=0.2×Total Dataset Size

For augmented datasets, the total size was derived by combining the real and synthetic data, defined as: Augmented Dataset Size=Original Dataset Size+Synthetic Data set Size.

### F. Machine Learning Model Evaluation

The performance of both synthetic and augmented datasets is evaluated using a suite of widely used machine learning models, such as Random Forest and Gradient Boosting, which were chosen for their robustness and applicability to tabular data. For each experimental setup, the models were trained and tested with standard machine learning metrics: accuracy, precision, recall, and F1-score. Comparative analysis was performed to identify improvements in model performance that could be attributed to the addition of synthetic data. Paired t-tests were also used to validate the statistical significance of differences in performance across the experimental setups. This ensured that the reported improvements were not due to random variations in the data or model behavior.To probe into the significance of performance improvement, paired t-tests were used with:

$$t = \frac{\bar{x}_d}{s_d/\sqrt{n}} \tag{2}$$

where x bar d is the mean difference, sd is the standard deviation, and n is the number of paired observations. Relative improvements in metrics were calculatedas:

$$\text{Relative Improvement } (\%) = \frac{\text{Metric (Aug.)} - \text{Metric (Org.)}}{\text{Metric (Org.)}} \times 100 \tag{3}$$

These measures ensured a statistically robust comparison of model performance across datasets.

### G. Tools and Reproducibility

The study was implemented using open-source tools and frameworks. Data preprocessing and machine learning workflows were carried out using Python 3.7.3, with core libraries such as Pandas and Scikit-learn. Synthetic data generation was done using the CTGAN library, and all experiments were run in a Jupyter Notebook environment. The computations were accelerated using an NVIDIA T4 GPU, which enabled efficient model training and data generation. All code and datasets were maintained in a version-controlled repository to ensure reproducibility of results. It detailed the preprocessing steps, model configurations, and evaluation procedures, so that other researchers could replicate and validate the findings.

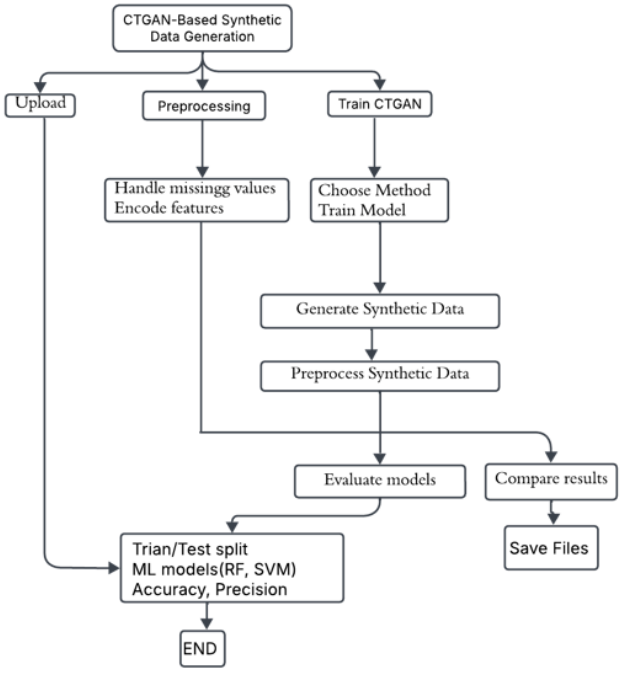


Fig. 1. Flow chart

## IV. IMPLEMENTATION

The implementation of this project involves preprocessing a tabular dataset with both numerical and categorical variables for compatibility with CTGAN. It involved statistical imputation for missing values, normalization of numerical features, and encoding of categorical variables. The optimized architecture of CTGAN was then trained with 256 hidden units, a learning rate of 0.0002, and 70 training epochs to generate

synthetic data that closely mimics real data distributions. Machine learning models, including Logistic Regression, Random Forest, and Gradient Boosting, were trained on the original, synthetic, and augmented datasets. Performance evaluation was conducted using accuracy, precision, recall, and F1-score, with statistical validation performed using paired t-tests. All experiments were run in a Python-based environment utilizing the libraries of Scikit-learn, Pandas, and CTGAN, to ensure reproducibility and scalability.

## V. RESULTS AND DISCUSSIONS

This section discusses the performance that machine learning models receive in terms of synthetic data generated by Conditional Tabular Generative Adversarial Networks (CTGAN). First and foremost, the primary goal of this research was to establish whether it was possible for synthetic data to enhance or maintain the predictive accuracy of models in comparison to the original dataset. For this purpose, a comparative analysis was performed with seven machine learning algorithms: Random Forest, Gradient Boosting, Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, and Naïve Bayes. All the algorithms were trained and tested on both the original dataset and the synthetic dataset. For this study, accuracy was chosen as the evaluation metric because it provides a clear measure of model performance. It follows that the synthetic data generated by CTGAN not only preserves the statistical characteristics of the original data but also improves the performance of models in some scenarios. This ensures that CTGAN is a sound method for generating synthetic data when real data availability is limited or restricted.

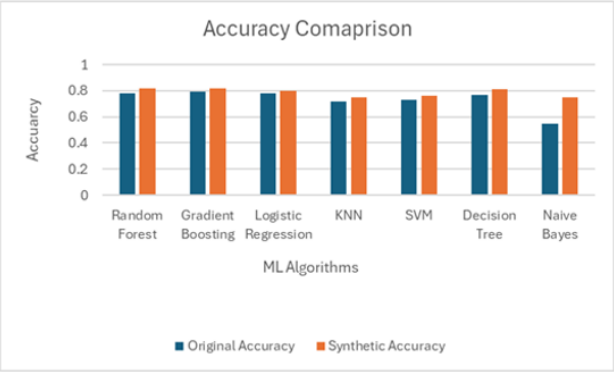


Fig. 2. Accuracy comparison

Figure 2 shows a comparative ability of models on both original and synthetic datasets. As found out in Figure 1,models trained on synthetic dataset either showed more consistent or improved accuracy in comparison with the original dataset. For example:Random Forest and Gradient Boosting yielded better performance over synthetic data. KNN, SVM, and Decision Tree gave similar performances among the two data sets. There was a strong improvement of the Naive Bayes classifier from synthetic training indicating that the distribution of classes CTGAN produces over the original ones might not be optimal.

TABLE I
PERFORMANCE COMPARISON

| Model | Acc(O) | Acc(S) | Pre(O) | Pre(S) | Rec(O) | Rec(S) | F1(O) | F1(S) |
|---|---|---|---|---|---|---|---|---|
| RF | 0.79 | 0.83 | 0.79 | 0.83 | 0.79 | 0.83 | 0.79 | 0.83 |
| GB | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 |
| LR | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 |
| KNN | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 |
| SVM | 0.76 | 0.80 | 0.76 | 0.80 | 0.76 | 0.80 | 0.76 | 0.80 |
| DT | 0.81 | 0.81 | 0.81 | 0.81 | 0.81 | 0.81 | 0.81 | 0.81 |
| NB | 0.66 | 0.78 | 0.66 | 0.78 | 0.66 | 0.78 | 0.66 | 0.78 |

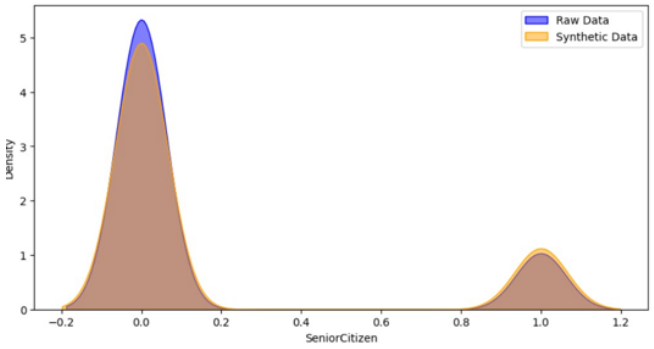


Fig. 3. Distribution comparison: Senior citizen

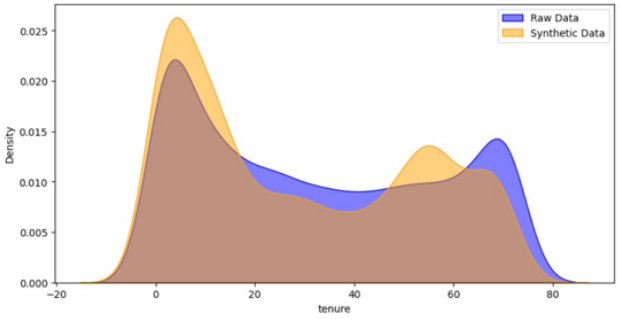


Fig. 4. Distribution comparison: Tenure

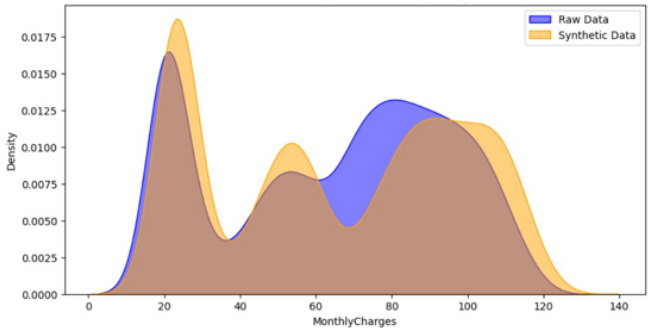


Fig. 5. Distribution comparison: Monthly changes

The following figures compare the distributions of three features—"SeniorCitizen," "tenure," and "MonthlyCharges"—across the original and synthetic datasets. The large overlap between the original data (blue) and synthetic data (yellow) across all three graphs indicates that CTGAN is doing a good job of replicating feature distributions accurately. This means that the synthetic data will maintain the important properties of the original dataset.

"SeniorCitizen"Feature: The distribution is bimodal, with most values near 0 (non-senior citizens) and a smaller peak at 1 (senior citizens). The synthetic data closely mirrors the original, maintaining densities of 5 for non-senior citizens and 0.2 for senior citizens, preserving demographic proportions essential for classification and analysis. "MonthlyCharges"Feature: This feature is highly skewed with peaks at $20 and$ 80. The synthetic data fits well, retaining densities of 0.0175 at $20 and$ $0.010 at 80$, which shows that CTGAN preserves the pricing patterns that are critical for financial analysis and revenue predictions.

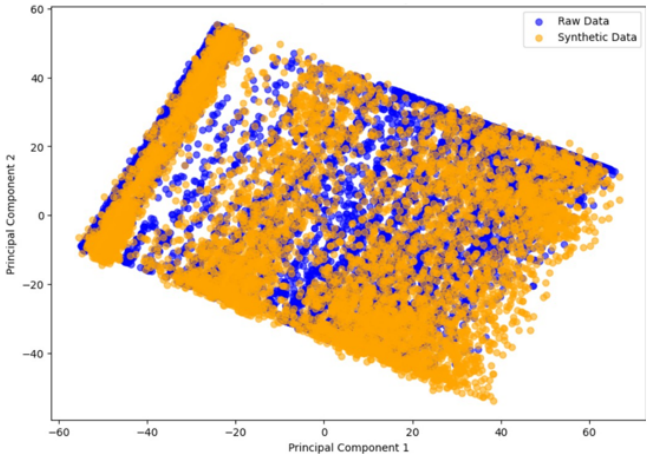


Fig. 6. PCA decomposition comparison

## VI. FUTURE SCOPE

This study demonstrates the applicability of CTGAN for producing good quality synthetic tabular data. Future work may focus on improving the fidelity of synthetic data by integrating advanced generative models like Variational Autoencoders or diffusion models with CTGAN. Further, the performance analysis on synthetic data in real-world applications, including financial modeling and fraud detection, could be profound for understanding its practical impact. Differential privacy and similar privacy-preserving techniques can be integrated to strengthen data security with utility preservation. Automated benchmarking frameworks for measuring synthetic data against various datasets can help establish standard evaluation metrics. Synthetic data generation will become much more scalable and applicable across many industries if such aspects are considered.

## VII. CONCLUSION

This paper proves that the generated synthetic tabular data is useful for training machine learning models in preserving statistical properties of the original datasets. It shows experimentally that performance for models trained on synthetic data can be close to those trained on real data, which would indicate the value of generative models in settings with limited data access. The difficulties are to improve data quality, ensure applicability in domains, and adapt the evaluation approach. Future research directions would include making CTGAN more adaptable and its performance in real-world deployments so that the maximum possible impact of synthetic data is derived from machine learning.

## REFERENCES


[1] Xu, Lei, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. "Modeling tabular data using conditional gan." Advances in neural information processing systems 32 (2019).

[2] Engelmann, Justin, and Stefan Lessmann. "Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning." Expert Systems with Applications 174 (2021):114582.

[3] Abufadda, Mohammad, and Khalid Mansour. "A survey of synthetic data generation for machine learning." In 2021 22nd international arab conference on information technology (ACIT), pp. 1-7. IEEE, 2021.

[4] Majeed, Abdul, and Seong Oun Hwang. "Moving Conditional GAN Close to Data: Synthetic Tabular Data Generation and its Experimental Evaluation." IEEE Transactions on Big Data (2024).

[5] Figueira, Alvaro, and Bruno Vaz. "Survey on synthetic data generation, evaluation methods and GANs." Mathematics 10, no. 15 (2022): 2733.

[6] Charitou, Charitos, Simo Dragicevic, and Artur d'Avila Garcez. "Synthetic data generation for fraud detection using gans." arXiv preprint arXiv:2109.12546 (2021).

[7] Charitou, Charitos, Simo Dragicevic, and Artur d'Avila Garcez. "Synthetic data generation for fraud detection using gans." arXiv preprint arXiv:2109.12546 (2021).

[8] Figueira, Alvaro, and Bruno Vaz. "Survey on synthetic data generation, evaluation methods and GANs." Mathematics 10, no. 15 (2022): 2733.

[9] Douzas, Georgios, and Fernando Bacao. "Effective data generation for imbalanced learning using conditional generative adversarial networks." Expert Systems with applications 91 (2018): 464-471.

[10] Miletic, Marko, and Murat Sariyar. "Challenges of Using Synthetic Data Generation Methods for Tabular Microdata." Applied Sciences 14, no. 14 (2024): 5975.

[11] Bourou, Stavroula, Andreas El Saer, Terpsichori- Helen Velivassaki, Artemis Voulkidis, and Theodore Zahariadis. "A review of tabular data synthesis using GANs on an IDS dataset." Information 12, no. 09 (2021): 375.

[12] Espinosa, Erica, and Alvaro Figueira. "On the quality of synthetic generated tabular data." Mathematics 11, no. 15 (2023): 3278.

[13] Lu, Yingzhou, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, and Wenqi Wei. "Machine learning for synthetic data generation: a review." arXiv preprint arXiv:2302.04062 (2023).

[14] Goyal, Mandeep, and Qusay H. Mahmoud. "A systematic review of synthetic data generation techniques using generative AI." Electronics 13, no. 17 (2024): 3509.

[15] Parise, Orlando, Rani Kronenberger, Gianmarco Parise, Carlo de Asmundis, Sandro Gelsomino, and Mark La Meir. "CTGAN-Driven Synthetic Data Generation: A Multidisciplinary, Expert-Guided Approach (TIMA)." Computer Methods and Programs in Biomedicine (2024): 108523.

[16] Lee, J. S., and O. Lee. "Ctgan vs tgan? which one is more suitable for generating synthetic eeg data." J. Theor. Appl. Inf. Technol 99, no. 10 (2021): 2359-2372.

[17] Dankar, Fida K., Mahmoud K. Ibrahim, and Leila Ismail. "A multi-dimensional evaluation of synthetic data generators." IEEE Access 10 (2022): 11147-11158.

[18] Dankar, Fida K., Mahmoud K. Ibrahim, and Leila Ismail. "A multi-dimensional evaluation of synthetic data generators." IEEE Access 10 (2022): 11147-11158.

[19] Hittmeir, Markus, Andreas Ekelhart, and Rudolf Mayer. ”On the utility of synthetic data: An empirical evaluation on machine learning tasks.” In Proceedings of the 14th International Conference on Availability, Reliability and Security, pp. 1-6. 2019.

[20] El Emam, Khaled, Lucy Mosquera, Xi Fang, and Alaa El-Hussuna. ”Utility metrics for evaluating synthetic health data generation methods: validation study.” JMIR medical informatics 10, no. 4 (2022): e35734.