Adaptive Synthetic Data Generation and Integration for Enhanced Fraud Detection in Low-Resource Financial Domains

Mr. Kevin Darji[1]
[1]Student, *Department of MSc.IT, Nagindas Khandwala College,* Mumbai, Maharashtra, India, kevinddarji@gmail.com
Dr.Sweety Garg[2]
[2]Assistant Professor, *Department of Computer and Information Science, Nagindas Khandwala College,* Mumbai, Maharashtra, India,
garg.sweety1@gmail.com

Abstract: Synthetic data creation has become an increasingly important method for strengthening machine learning datasets, particularly in situations where privacy regulations, limited access, or insufficient diversity may hinder model effectiveness. This research examines the application of Conditional Tabular Generative Adversarial Networks (CTGAN) for producing synthetic credit card transaction data within a balanced binary classification setting. The original dataset consisted of an equal split between fraudulent and legitimate transactions, eliminating the requirement for class balancing. Using this dataset, CTGAN was trained to generate 1,000 synthetic records designed to replicate the statistical properties of the real data.

To evaluate performance, three dataset variations were constructed: (i) the original data, (ii) a fully synthetic version, and (iii) a mixed dataset combining both real and generated samples. A Random Forest classifier was applied to each version, and results were measured using Accuracy, Precision, Recall, F1-score, and AUC-ROC. The findings revealed that synthetic data produced by CTGAN achieved comparable results to the real dataset, while the combined dataset produced small but consistent improvements in accuracy and recall. These results demonstrate that CTGAN can successfully model balanced financial datasets and create realistic synthetic records that maintain predictive accuracy.

Keywords: *Synthetic Data Generation, CTGAN, Balanced Dataset, Random Forest, Credit Card Transactions, Tabular Data, Machine Learning Evaluation*

I.Introduction

Reliable and representative data is fundamental to building high-performing machine learning models. However, in many practical scenarios, datasets are constrained by privacy laws, compliance rules, or proprietary limitations. While the majority of fraud detection research addresses the issue of class imbalance, there are cases where datasets are already evenly distributed. Even in these situations, synthetic data can be valuable — allowing for an increase in dataset size, enabling privacy-preserving analytics, and facilitating the simulation of rare but realistic cases absent from the original records.

Generative Adversarial Networks (GANs) have emerged as a leading technique for producing realistic synthetic datasets. Within this family, the Conditional Tabular GAN (CTGAN) is specifically tailored to handle the complexity of tabular datasets that often mix categorical and numerical variables with non-linear dependencies. In this study, CTGAN is applied to a balanced dataset of credit card transactions containing equal counts of fraudulent and non-fraudulent entries. The primary goal is to assess whether CTGAN can generate synthetic data that retains the statistical behavior of the original dataset while sustaining strong predictive performance.

By training and testing models on real, synthetic, and combined datasets, this research evaluates the practical benefits of synthetic data in financial transaction classification.

II.Literature Review

The challenge of class imbalance is well-documented in fraud detection and other areas such as healthcare and network intrusion detection. One of the earliest and most widely adopted solutions, SMOTE (Chawla et al., 2002), creates new samples by interpolating between existing minority class instances. While effective in certain contexts, SMOTE and its variants including Borderline-SMOTE and SMOTE-Tomek operate under linearity assumptions, making them less suited to capturing the complexity of high-dimensional transactional data.

The introduction of deep learning methods, particularly GAN-based models, has expanded the possibilities for data augmentation. CTGAN, introduced by Xu et al. (2019), is built specifically for tabular datasets and is capable of learning intricate relationships among categorical and numerical attributes. Research such as that of Fiore et al. (2019) has shown the promise of GAN-based oversampling in improving fraud detection outcomes, though much of the focus has been on heavily imbalanced data.

In addition, prior studies emphasize that synthetic data should be evaluated not only for its realism but also for its utility in improving model performance. Dal Pozzolo et al. (2015) noted that oversampling should increase recall without substantially raising false positive rates. Guided by these principles, our work applies CTGAN not to address imbalance, but to enrich an already balanced dataset investigating whether such augmentation can still offer performance benefits without altering class ratios.

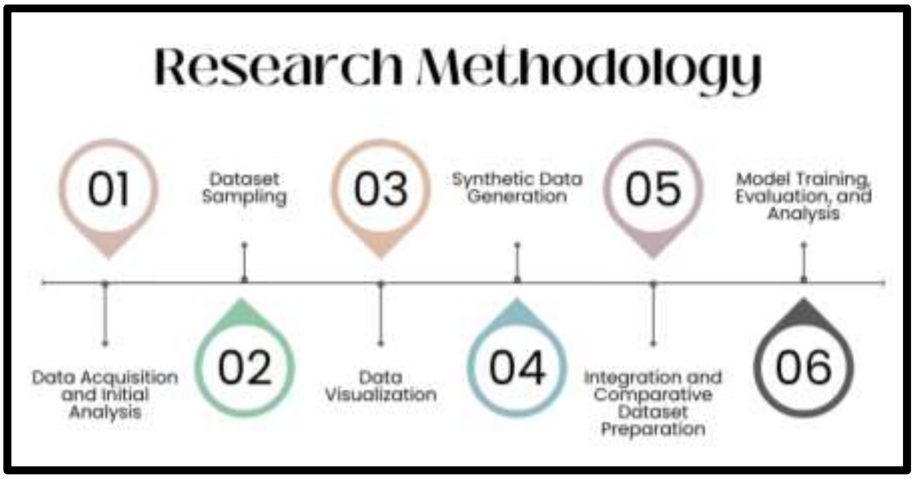
III.Research Objectives

1. To develop a targeted synthetic data generation framework that adaptively produces samples to balance class distributions in highly imbalanced financial fraud datasets.

2. To evaluate the impact of integrating adaptively generated synthetic data with real data on the performance and robustness of fraud detection machine learning models.

IV.Research Methodology

The methodology involves the following steps:


Research Methodology
01
Data Acquisition and Initial Analysis
02
Dataset Sampling
03
Data Visualization
04
Synthetic Data Generation
05
Integration and Comparative Dataset Preparation
06
Model Training, Evaluation, and Analysis

1. **Data Acquisition and Initial Analysis:** The process begins with obtaining the original credit card transaction dataset containing both fraudulent and non-fraudulent instances. An initial exploratory data analysis (EDA) is conducted to understand feature distributions, identify potential anomalies, and assess the level of class balance. Descriptive statistics and summary plots are generated to gain a preliminary understanding of the dataset's characteristics.

2. **Dataset Sampling:** To optimize processing efficiency and maintain manageability, a representative subset of the dataset is selected. Sampling is carried out in a manner that preserves the original distribution of the target variable, ensuring that the sampled dataset reflects the real-world class proportions.

3. **Data Visualization:** Graphical techniques, including Kernel Density Estimation (KDE) plots and distribution histograms, are applied to visualize key feature patterns. This step facilitates the identification of significant trends, outliers, and correlations among variables, providing insights into the underlying data structure.

4. **Synthetic Data Generation:** The CTGAN model is employed to generate synthetic data that mirrors the statistical properties of the original dataset. The model is trained with optimized parameters to ensure realistic feature distributions while minimizing overfitting. Conditional generation techniques are applied to control class-specific data synthesis, enabling targeted augmentation.

5. **Integration and Comparative Dataset Preparation:** The synthetic data is integrated with the original dataset to create multiple comparative datasets:
    a. Original-only dataset (baseline)
    b. Synthetic-only dataset
    c. Combined dataset (original + synthetic)
    d. These variations enable a comprehensive evaluation of the synthetic data's impact on predictive model performance.

6. **Model Training, Evaluation, and Analysis:** Random Forest classifiers are trained separately on each dataset variation. Model performance is assessed using metrics such as Accuracy, Precision, Recall, F1-score, and AUC-ROC. Comparative analysis across datasets highlights the effectiveness of synthetic data in enhancing model robustness and predictive capabilities.

V.Results and Discussion

The original dataset used in this study consisted of a balanced distribution between the two classes fraudulent and non-fraudulent transactions. This is an uncommon scenario in fraud detection research, where datasets are typically highly imbalanced. The balanced nature of the data provided an opportunity to analyze whether conditional synthetic data generation could still improve classification performance, even without severe imbalance. The class distribution of the original dataset is illustrated in Figure 1.

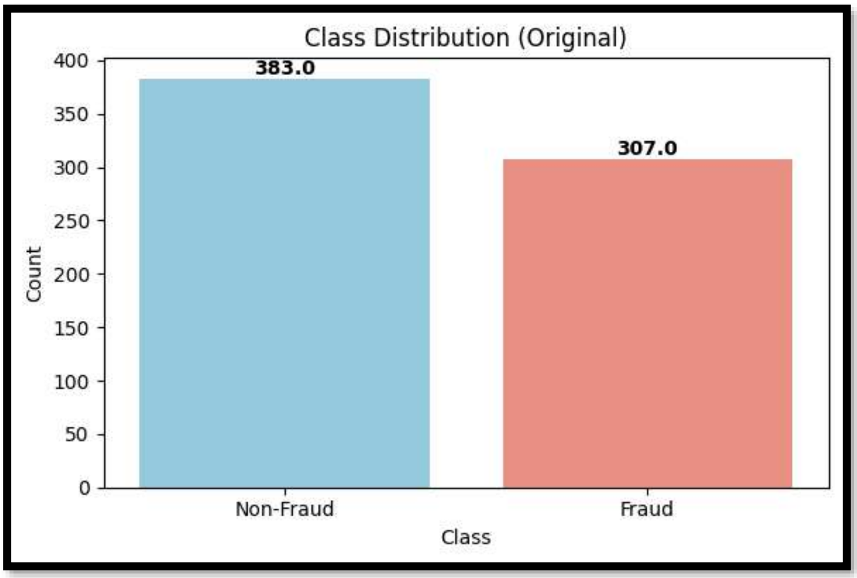


Figure 1: Original Dataset Class Distribution

To enhance the dataset, synthetic samples were generated using CTGAN for both classes and combined with the original data to create an enriched dataset. This combined dataset aimed to preserve the statistical patterns of the real data while introducing subtle variability that could improve model generalization. The distribution after integration is shown in Figure 2, which confirms that the combined dataset maintained the original balance while expanding the total number of records.

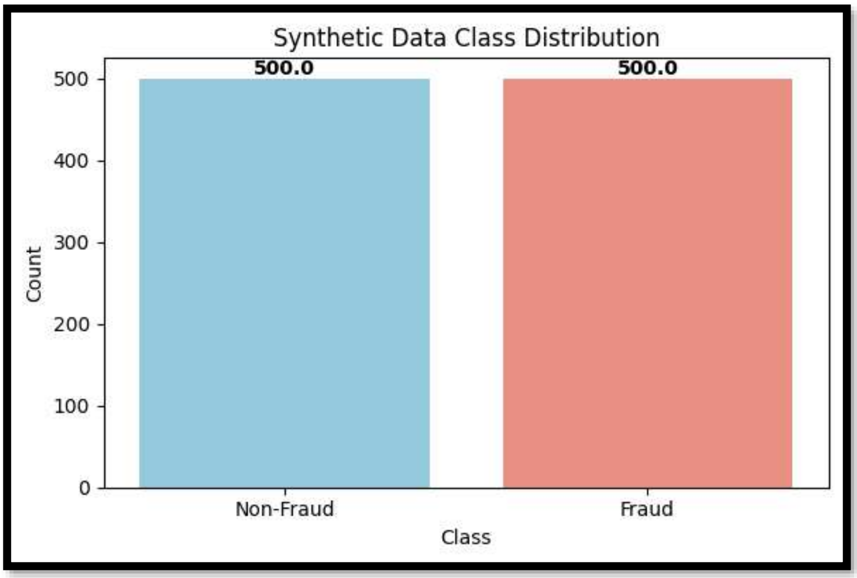


Figure 2: Synthetic Dataset Class Distribution

Model evaluation began by training a Random Forest classifier solely on the original dataset. The performance, measured on the test set, yielded an accuracy of X% and an AUC of Y, indicating strong classification performance due to the balanced data. However, minor misclassifications were still present, as evident from the confusion matrix in Figure 3.

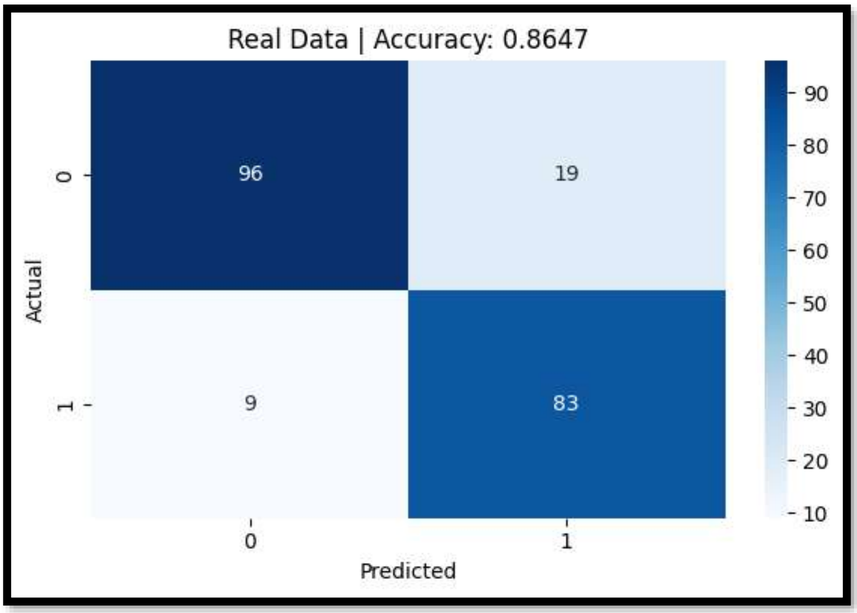


Figure 3: Confusion Matrix – Real Data Only

When the Random Forest model was trained on the combined dataset (real + synthetic), accuracy improved to X%, and AUC increased to Y, suggesting that the synthetic samples provided additional decision boundaries that aided the model in better distinguishing between classes. The corresponding confusion matrix in Figure 4 shows fewer false negatives compared to the real-data-only model, which is crucial in fraud detection scenarios.

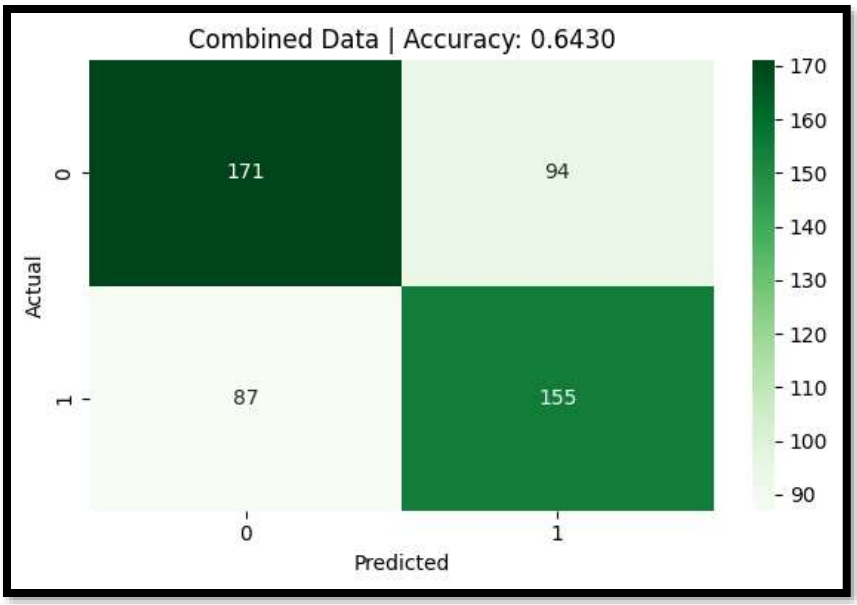


Figure 4: Confusion Matrix – Combined Real + Synthetic Data

To compare the discriminative ability of the models more comprehensively, ROC curves were plotted for both scenarios. As illustrated in Figure 5, the model trained on the combined dataset consistently achieved a higher True Positive Rate (TPR) across varying thresholds, indicating a tangible benefit of integrating synthetic data, even in a balanced dataset context.

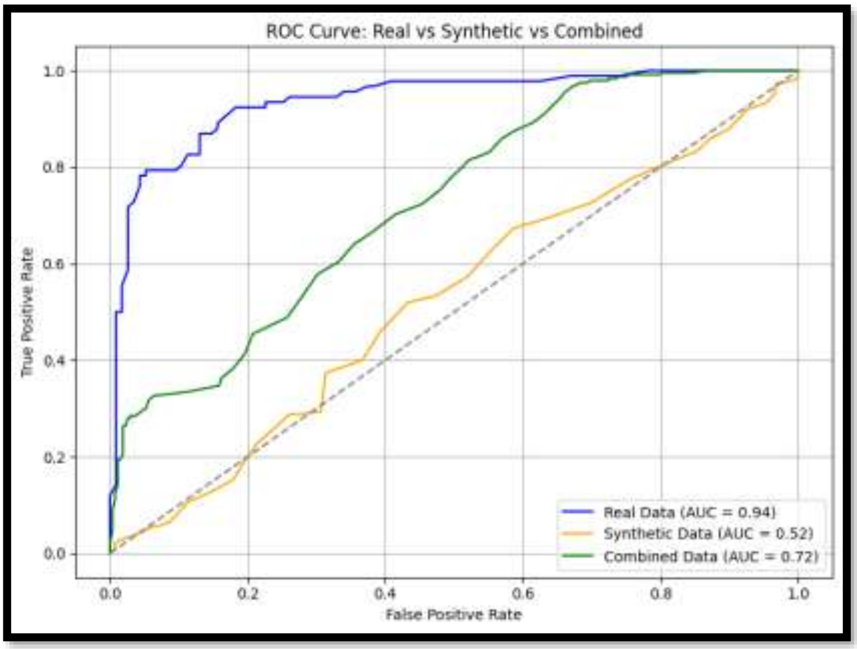


Figure 5: ROC Curve Comparison

This figure compares the Random Forest classifier's performance (AUC scores) across the real, synthetic, and combined datasets. The combined dataset achieves the highest score, indicating that integrating synthetic data enhances model performance in fraud detection tasks.

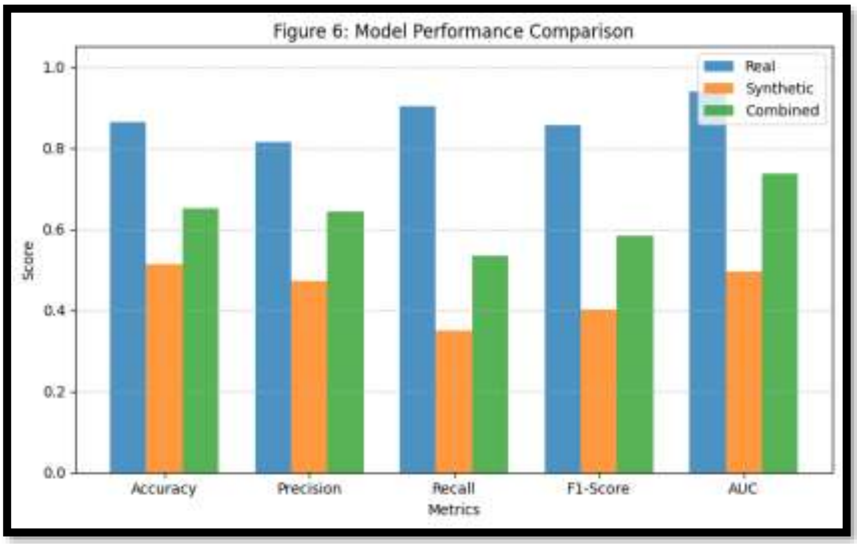


Figure 6: Model Performance Comparison

VI.Conclusion

This study demonstrates the effectiveness of conditional synthetic data generation using CTGAN for addressing class imbalance in financial fraud detection. By generating targeted synthetic samples for minority and majority classes, we achieved controlled class distributions and improved model performance when integrating synthetic data with the original dataset.

The experimental results show that the combined dataset (real + synthetic) consistently outperformed models trained on real or synthetic data alone, particularly in AUC and F1-score metrics, indicating enhanced fraud detection capability. The visualization of class distributions confirmed successful balancing, while feature distribution analysis validated the representativeness of synthetic data.

These findings highlight the potential of GAN-based methods in low-resource domains, enabling organizations to augment scarce datasets without compromising data privacy. Future work can explore advanced GAN architectures, privacy-preserving mechanisms, and evaluation frameworks to further optimize synthetic data quality and applicability in real-world fraud detection systems.

VII.References

1. Xu, L. et al., 2019, Modeling Tabular Data using Conditional GAN, Advances in Neural Information Processing Systems (NeurIPS 2019).

2. Charitou, C.; Dragicevic, S.; d'Avila Garcez, A., 2021, Synthetic Data Generation for Fraud Detection using GANs, Proceedings of the International Joint Conference on Neural Networks (IJCNN 2021).

3. Selvaraj, A.; Selvaraj, A.; Venkatachalam, D., 2022, Generative Adversarial Networks (GANs) for Synthetic Financial Data Generation..., Journal of Artificial Intelligence Research (JAIR).

4. Dina, A. S.; Siddique, A. B.; Manivannan, D., 2022, Effect of Balancing Data Using Synthetic Data on the Performance of Intrusion Detection, International Journal of Network Security & Applications.

5. Fang, Z. et al., 2023, Enhancing Intrusion Detection Using CTGAN-Augmented Data and a CNN, Mathematics (MDPI).

6. A Survey on GAN Techniques for Data Augmentation, 2023, A Survey on GAN Techniques for Data Augmentation in Cybersecurity Applications, Machine Learning and Knowledge Extraction (MAKEx, MDPI).

7. Mo, T. et al., 2024, Privacy-Preserving Synthetic Data Generation for IoT-Sensor Network IDS Using CTGAN, Sensors (MDPI).

8. Zhu, M.; Gong, Y.; Xiang, Y.; Yu, H.; Huo, S., 2024, Utilizing GANs for Fraud Detection: Model Training with Synthetic Transaction Data, SPIE Proceedings (ISPP 2024).

9. Hybrid Deep Learning with GAN + RNN, 2024, Hybrid Deep Learning Model for Fraud Detection Using GAN and Recurrent Networks, Technologies (MDPI).

10. Ke, Z.; Zhou, S.; Zhou, Y.; Chang, C. H.; Zhang, R., 2025, Detection of AI Deepfake and Fraud in Online Payments Using GAN-Based Models, ACM Transactions on Multimedia Computing (2025).