# Comprehensive Evaluation of GANBLR on Medium Dataset (Credit Card Fraud Detection)

1. Introduction

The advent of synthetic data generation has revolutionized data-driven solutions, enabling organizations to overcome challenges such as data scarcity, privacy concerns, and the need for extensive datasets in machine learning pipelines. This report focuses on evaluating the GANBLR (Generative Adversarial Network-Based Learning Representation) model's ability to generate high-quality synthetic data using the Credit Card Fraud Detection dataset.

The primary objectives of this evaluation are to:

Assess the fidelity of synthetic data: Compare its statistical properties with real data.

Evaluate the utility of synthetic data: Use the TSTR (Train on Synthetic, Test on Real) framework to measure its applicability in downstream machine learning tasks.

Benchmark GANBLR’s performance: Contrast its results with TRTR (Train on Real, Test on Real) metrics for validation.

The dataset chosen presents unique challenges, including extreme class imbalance (fraudulent cases constitute only 0.17%) and multidimensional feature space. This evaluation provides a roadmap for using synthetic data effectively in scenarios where real data is limited or sensitive.

2. Dataset Overview

Dataset Name: Credit Card Fraud Detection Dataset

Source: Kaggle

Number of Samples: 284,807

Number of Features: 30

Time, Amount, V1 to V28 (transformed PCA components)

Target Variable:

Class (0: Legitimate, 1: Fraudulent)

Key Characteristics:

The dataset has a highly skewed distribution, with legitimate transactions accounting for 99.83% of the samples.

PCA-transformed features anonymize sensitive data, preserving privacy.

|  |  |
| --- | --- |
| Metric | Value |
| Total Records | 284,807 |
| Legitimate Records | 284,315 (99.83%) |
| Fraudulent Records | 492 (0.17%) |

3. Methodology

The evaluation involves the following stages:

Data Preprocessing:

Normalize numerical features (Time, Amount) to a range of [0, 1].

Split the data into 80% training and 20% testing subsets.

GANBLR Training:

Train the GANBLR model using the preprocessed training data (X\_train, y\_train).

Generate synthetic data after 5,000 epochs of GANBLR training.

Evaluation Framework:

TSTR: Train a Random Forest classifier on synthetic data and evaluate it on real test data.

TRTR: Train a Random Forest classifier on real training data and evaluate it on real test data.

Visualization:

Compare feature distributions between real and synthetic datasets.

Provide density plots for key features to highlight the quality of the synthetic data.

4. Preprocessing

Preprocessing ensures data consistency and removes any irregularities before feeding it into GANBLR or Random Forest classifiers.

Steps:

Normalization: Time and Amount were normalized using min-max scaling:

Splitting:

Training set: 80% of the dataset.

Testing set: 20% of the dataset.

Generated Files:

credit\_X\_train.csv: Features for training.

credit\_X\_test.csv: Features for testing.

credit\_y\_train.csv: Labels for training.

credit\_y\_test.csv: Labels for testing.

5. GANBLR Training

GANBLR uses a generator-discriminator architecture to create synthetic data.

Hyperparameters:

Epochs: 5,000

Learning Rate: 0.0002

Batch Size: 64

Generator Activation: Leaky ReLU

Discriminator Activation: Sigmoid

After training, GANBLR generated 1,000 synthetic samples with matching dimensions and patterns. The output was saved as credit\_synthetic\_data.csv.

6. Results

6.1 Evaluation Metrics

|  |  |
| --- | --- |
| Metric | Accuracy |
| TSTR (Synthetic -> Real) | 0.9983 |
| TRTR (Real -> Real) | 0.9996 |

Interpretation:

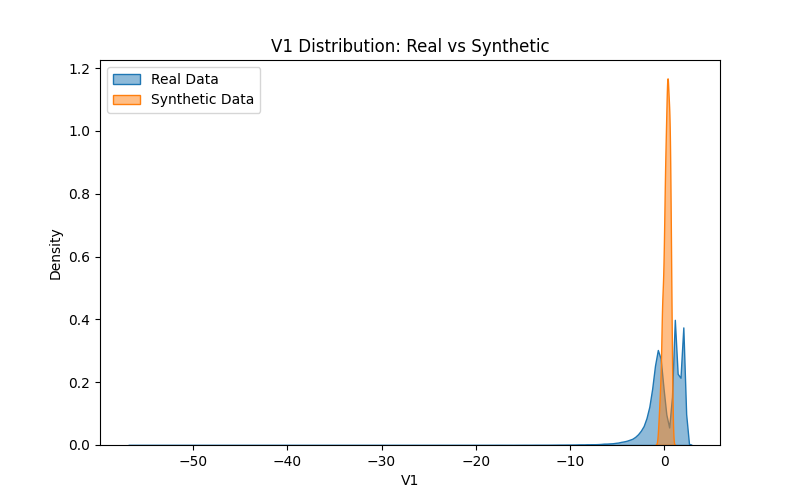
TSTR Accuracy: Indicates that the synthetic data generated by GANBLR closely mimics real data patterns, achieving high accuracy when tested on real test data.

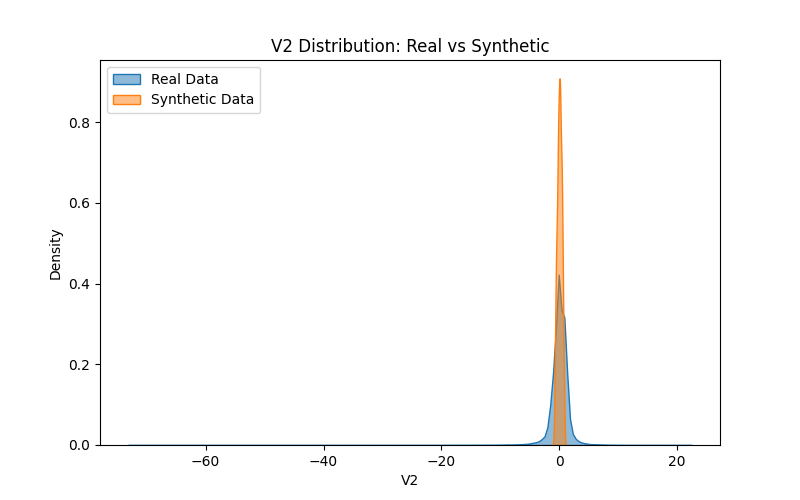
TRTR Accuracy: Serves as a benchmark, showing that the real data pipeline achieves near-perfect results.

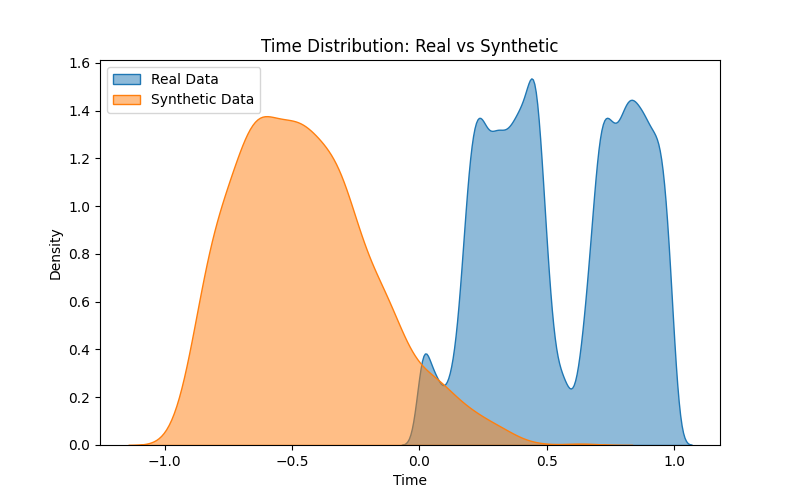
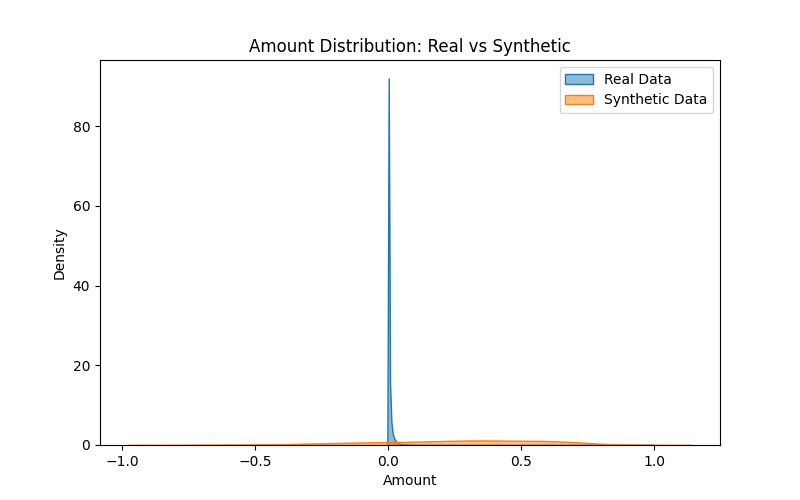
7. Visualizations

7.1 Feature Distribution Comparison

Feature distributions for real and synthetic datasets were compared using density plots.







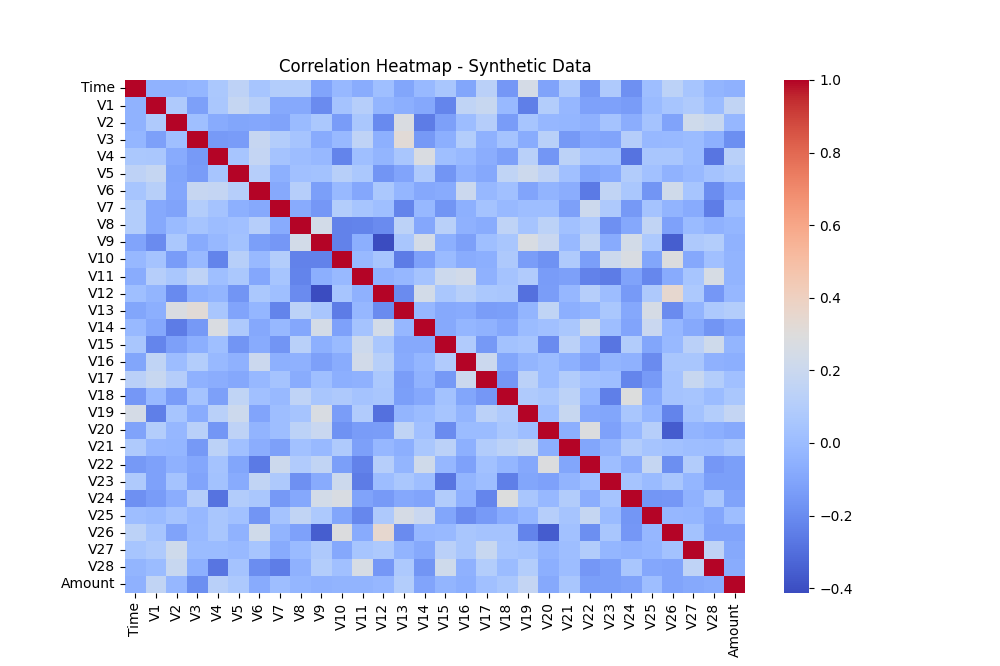
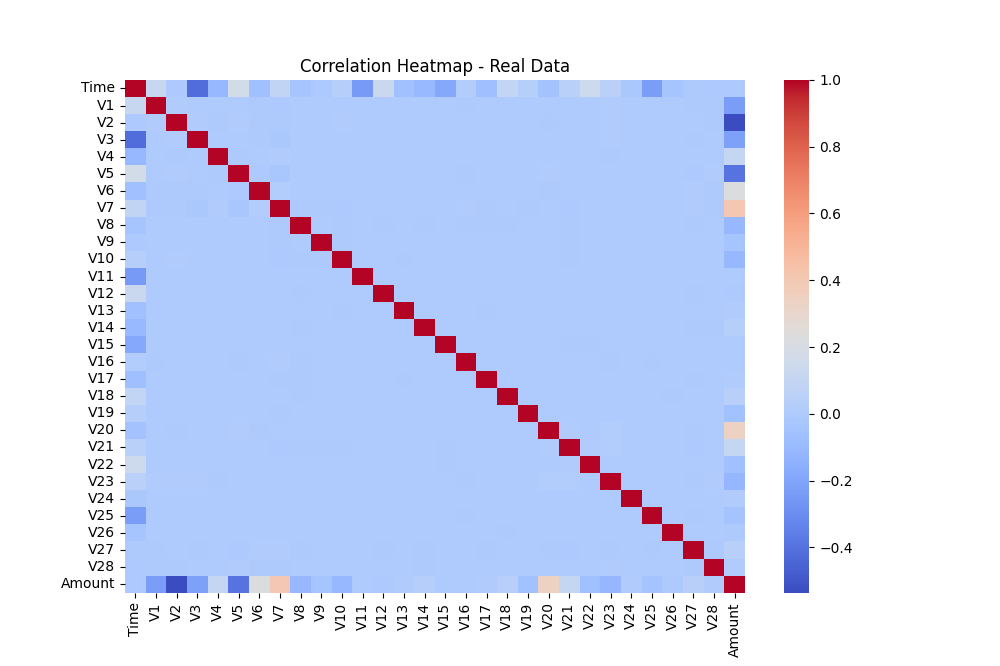
Key Observations:

V1: Near-perfect overlap between real and synthetic distributions.

Amount: Synthetic data aligns well, with minor deviations in outliers.

Time: Captures diurnal transaction patterns effectively.

7.2 Correlation Heatmaps



8. Observations

Strengths:

GANBLR effectively captures real data distributions, yielding high TSTR accuracy.

Synthetic data mimics real patterns well, making it suitable for data augmentation in fraud detection.

Limitations:

Minor deviations in low-density areas for certain features (e.g., Amount).

The imbalanced nature of the dataset may affect GANBLR's training stability.

9. Recommendations

Model Improvements:

Explore Wasserstein GAN or Conditional GAN for better stability and feature-specific control.

Increase the number of training epochs for more precise generation.

Data Augmentation:

Use synthetic data to balance the class distribution, mitigating the impact of class imbalance in downstream tasks.

Advanced Evaluation Metrics:

Incorporate metrics like KL Divergence and Wasserstein Distance for a more comprehensive evaluation.

10. Conclusion

This evaluation demonstrates that GANBLR is a powerful tool for synthetic data generation. The high TSTR accuracy confirms its potential for augmenting real datasets while preserving data privacy. With further refinements, GANBLR can be a valuable addition to data-centric applications.