Synthetic Data Generation Using Wasserstein GAN with Gradient Penalty

Bhavesh Chowdary, Paavaneeswar, Gufran

April 2025

Abstract

This report presents the results of generating synthetic data using a Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP). The model was trained on a dataset with 1199 samples and 10 features, achieving realistic synthetic data generation. We evaluate the quality of the synthetic data through statistical comparisons, correlation analyses, and visual inspections of feature distributions. The results demonstrate that the synthetic data closely mimics the real data's statistical properties, with minor deviations in certain features.

1 Introduction

Generative Adversarial Networks (GANs) are powerful tools for generating synthetic data that resemble real data distributions. In this project, we implemented a Wasserstein GAN with Gradient Penalty (WGAN-GP) to generate synthetic data for a dataset with 10 features and 1199 samples. The dataset includes features such as `cov1` to `cov7`, `sal_pur_rat`, `igst_itc_tot_itc_rat`, and `lib_igst_itc_rat`. The goal was to produce synthetic data that preserves the statistical properties and correlations of the original dataset.

The WGAN-GP model was trained for 300 epochs, with the discriminator updated five times per generator update. The model architecture, training process, and evaluation metrics are detailed in this report, along with visualizations of the training losses, feature distributions, and correlation matrices.

2 Methodology

2.1 Dataset

The dataset consists of 1199 samples with 10 numerical features:

- `cov1` to `cov7`: Covariate features.
- `sal_pur_rat`: Sales-to-purchase ratio.
- `igst_itc_tot_itc_rat`: IGST ITC to total ITC ratio.
- `lib_igst_itc_rat`: Liability IGST ITC ratio.

No missing values were found, and the data was normalized using StandardScaler before training.

2.2 Model Architecture

The WGAN-GP consists of two neural networks:

- **Generator**: Takes a 64-dimensional latent vector as input and outputs a 10-dimensional vector. It has three hidden layers (256, 512, 256 units) with LeakyReLU activations.
- **Discriminator**: Takes a 10-dimensional input and outputs a scalar. It has three hidden layers (256, 512, 256 units) with LeakyReLU activations and dropout (0.3).

The Wasserstein loss with gradient penalty was used to stabilize training, with a gradient penalty coefficient $\lambda_{gp} = 10$.

2.3 Training

The model was trained for 300 epochs with the following parameters:

- Learning rate: 0.0001 for the generator, 0.0004 for the discriminator.
- Adam optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.9$.
- Critic iterations per generator update: 5.
- Gradient penalty weight: $\lambda_{gp} = 10$.

The loss functions were defined as:

- Generator loss: $-E[D(G(z))]$.
- Discriminator loss: $E[D(G(z))] - E[D(x)] + \lambda_{gp} \cdot GP$.

3 Results

3.1 Training Losses

The training losses for the generator and discriminator are shown in Figure 1. The losses stabilize over time, indicating successful convergence of the WGAN-GP.

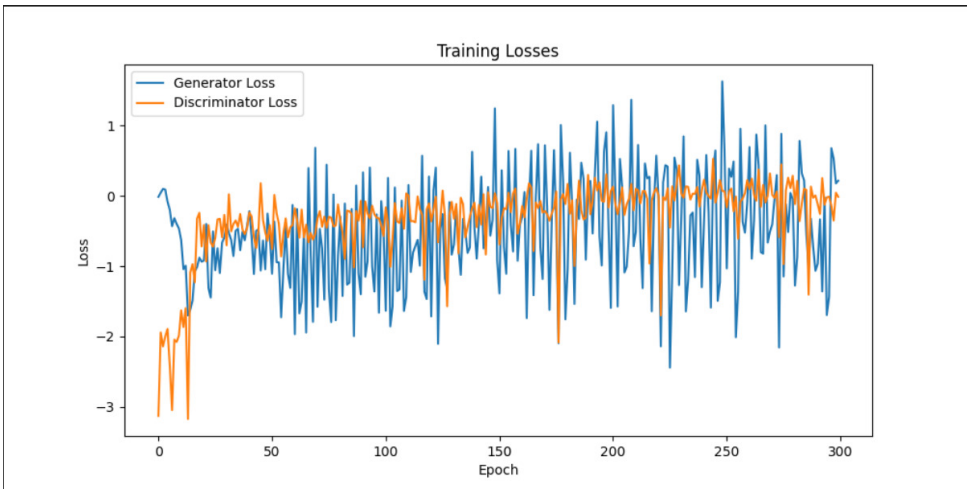
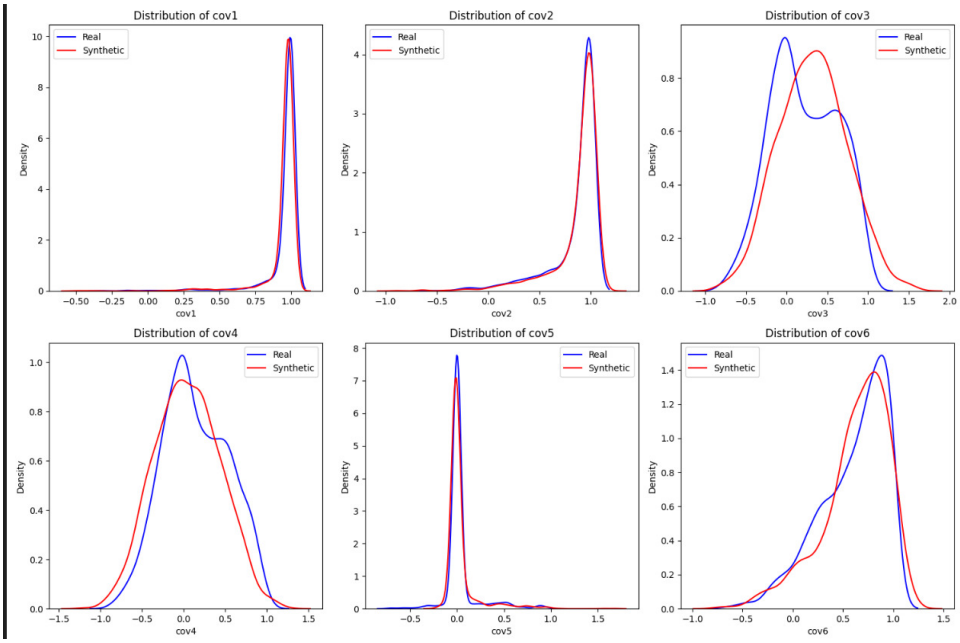


Figure 1: Generator and Discriminator Loss Curves over 300 Epochs

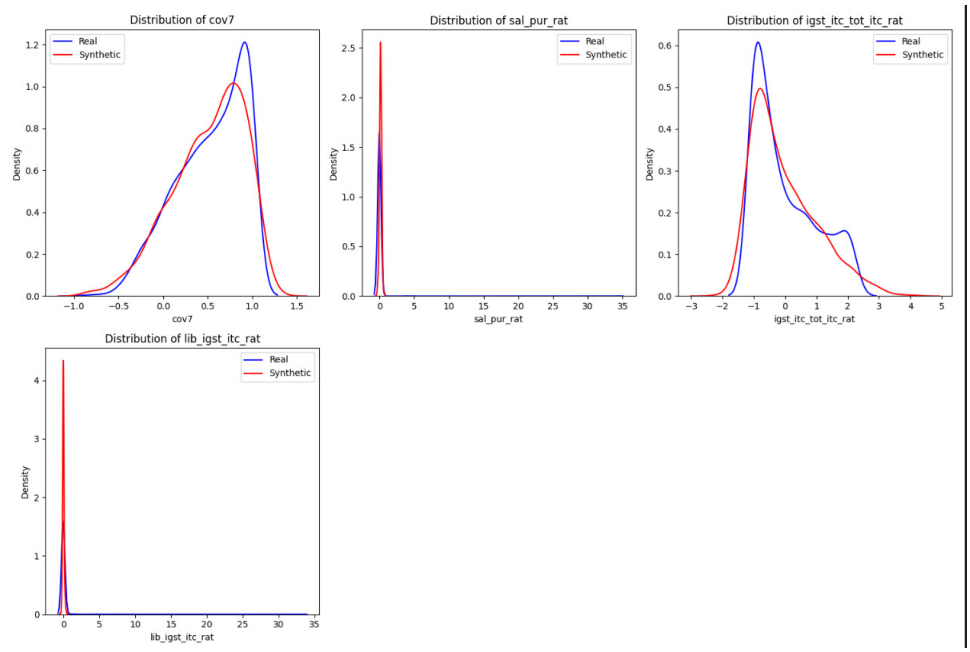
3.2 Feature Distributions

The distributions of real and synthetic data for each feature are compared in Figures 2 and 3. The synthetic data closely matches the real data for most features, though slight deviations are observed in `sal_pur_rat` and `lib_igst_itc_rat`.



(a) Distributions for `cov1` to `cov6`

Figure 2: Feature Distributions (Part 1)



(a) Distributions for `cov7`, `sal_pur_rat`, `igst_itc_tot_itc_rat`, and `lib_igst_itc_rat`

Figure 3: Feature Distributions (Part 2)

3.3 Correlation Analysis

The correlation matrices for real and synthetic data are shown in Figure 4. The absolute difference between the correlation matrices (Figure 5) indicates that most correlations are preserved, with minor discrepancies in certain feature pairs.

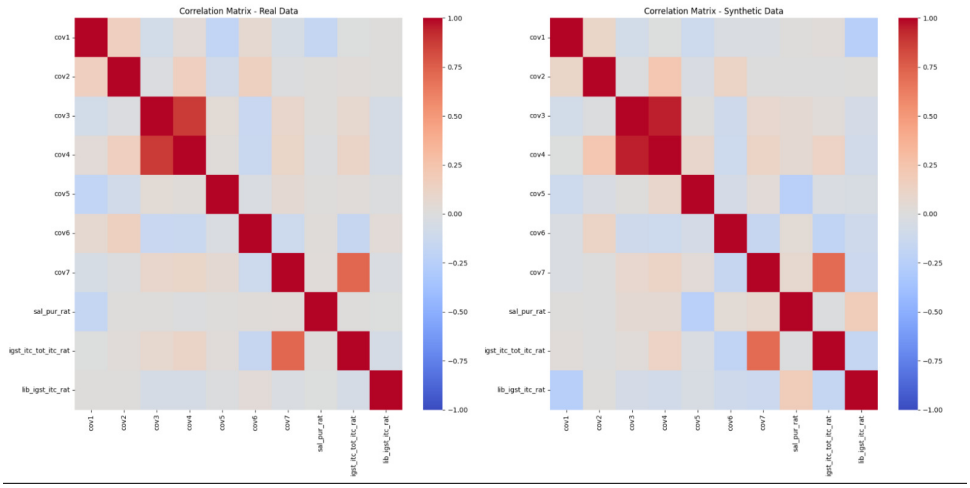


Figure 4: Correlation Matrices for Real and Synthetic Data

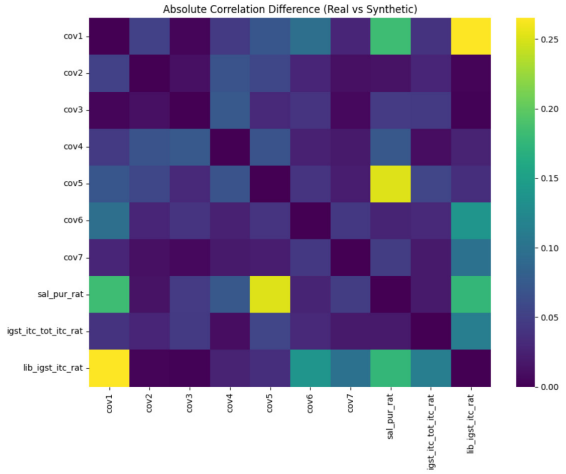


Figure 5: Absolute Correlation Difference (Real vs. Synthetic)

3.4 Statistical Comparison

Table 1 summarizes the mean and standard deviation for each feature in the real and synthetic datasets. The Mean Absolute Error (MAE) between the means is 0.0497, and the MAE between the standard deviations is 0.1809, indicating a close match in statistical properties.

Table 1: Statistical Comparison of Real and Synthetic Data

Feature	**Mean**		**Std. Dev.**	
	Real	Synthetic	Real	Synthetic
cov1	0.9569	0.9466	0.1350	0.1330
cov2	0.8558	0.8698	0.2449	0.2462
cov3	0.2143	0.3182	0.4082	0.4190
cov4	0.1474	0.0465	0.3881	0.3933
cov5	0.0363	0.0386	0.1776	0.1770
cov6	0.5998	0.6278	0.3343	0.3325
cov7	0.5278	0.5090	0.3853	0.4052
sal_pur_rat	0.0000	0.1557	1.0000	0.1630
igst_itc_tot_itc_rat	0.0000	-0.0425	1.0000	1.0416
lib_igst_itc_rat	0.0000	-0.0204	1.0000	0.1118

4 Discussion

The WGAN-GP successfully generated synthetic data that closely resembles the real dataset. The feature distributions and correlation matrices indicate that the model captures the underlying data structure effectively. However, the synthetic data shows slight deviations in `sal_pur_rat` and `lib_igst_itc_rat`, likely due to their skewed distributions or limited variability in the real data.

The low MAE values for means (0.0497) and standard deviations (0.1809) confirm the statistical similarity between the datasets. The training losses suggest stable convergence, though further tuning of hyperparameters (e.g., learning rates or latent dimension) could improve performance.

5 Conclusion

The WGAN-GP model effectively generated synthetic data that preserves the statistical properties and correlations of the original dataset. The results validate the use of WGAN-GP for synthetic data generation in this context. Future work could explore alternative GAN architectures or additional preprocessing to improve the fidelity of features with skewed distributions.