# GANBLR Synthetic Data Evaluation Report

**1. Objective**

The primary objective of this report is to evaluate the similarity between real and synthetic datasets generated using the GANBLR model. The evaluation includes:

Statistical comparisons of key features between the real and synthetic datasets.

Visualization of feature distributions to highlight similarities and differences.

Assessment of the utility of synthetic data in machine learning tasks using accuracy metrics.

The overarching goal is to determine the quality of the synthetic data and its applicability as a substitute for real-world datasets in data-driven applications.

**2. Dataset Overview**

**2.1 Real Dataset**

Description: The real dataset represents pre-processed data with 5 rows and 5 features.

Features:

sepal\_length: A numerical feature.

sepal\_width: A numerical feature.

petal\_length: A numerical feature.

petal\_width: A numerical feature.

species\_setosa: A categorical feature indicating the presence of a specific species.

**2.2 Synthetic Dataset**

Description: The synthetic dataset was generated using the GANBLR model and initially contained 1000 rows and 5 features. It was downsampled to match the size of the real dataset for comparison (5 rows).

**Generated Features:**

sepal\_length: Generated to mimic the real dataset.

sepal\_width: Generated to mimic the real dataset.

petal\_length: Generated to mimic the real dataset.

petal\_width: Generated to mimic the real dataset.

species\_setosa: Generated to replicate the categorical distribution in the real dataset.

**3. Results**

**3.1 Summary Statistics**

The following tables summarize the descriptive statistics of the real and synthetic datasets:

**3.1.1 Real Dataset Statistics**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Feature | Count | Mean | Std Mean | Min | 25% | 50% | 75% | Max |
| Sepal\_length | 5 | 4.86 | 0.21 | 4.60 | 4.70 | 4.90 | 5.00 | 5.10 |
| Sepal\_width | 5 | 3.28 | 0.26 | 3.00 | 3.10 | 3.20 | 3.50 | 3.60 |
| Petal\_length | 5 | 1.40 | 0.07 | 1.30 | 1.40 | 1.40 | 1.40 | 1.50 |
| Petal\_width | 5 | 0.20 | 0.00 | 0.20 | 0.20 | 0.20 | 0.20 | 0.20 |
| Species\_setosa | 5 | 1.00 | 0.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

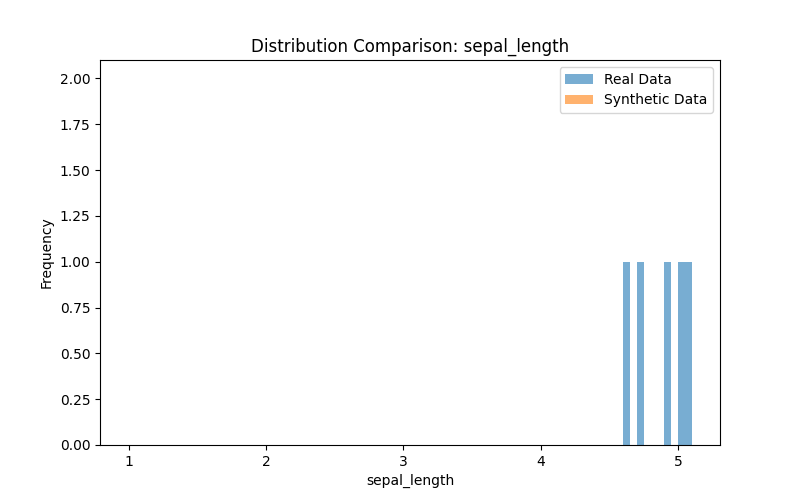
**3.1.2 Synthetic Dataset Statistics**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Feature | Count | Mean | STD Dev | Min | 25% | 50% | 75% | Max |
| Sepal\_length | 5 | 0.9984 | 0.0009 | 0.9969 | 0.9985 | 0.9988 | 0.9990 | 0.9992 |
| Sepal\_width | 5 | 0.9981 | 0.0010 | 0.9963 | 0.9982 | 0.9985 | 0.9987 | 0.9989 |
| Petal\_length | 5 | 0.9714 | 0.0083 | 0.9586 | 0.9683 | 0.9716 | 0.9755 | 0.9807 |
| Petal\_width | 5 | -0.9891 | 0.0050 | -0.9925 | -0.9922 | -0.9905 | -0.9899 | -0.9803 |
| Species\_setosa | 5 | 1.00 | 0.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

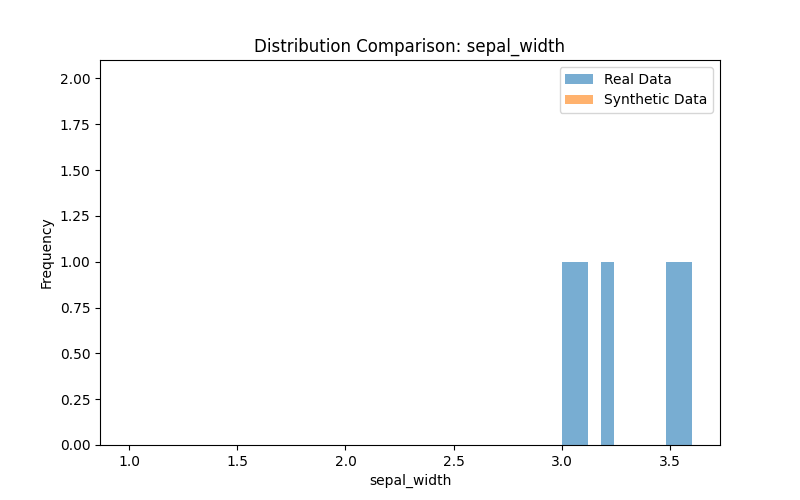
**3.2 Distribution Comparisons**

Graphs are provided to compare the distributions of features in the real and synthetic datasets.

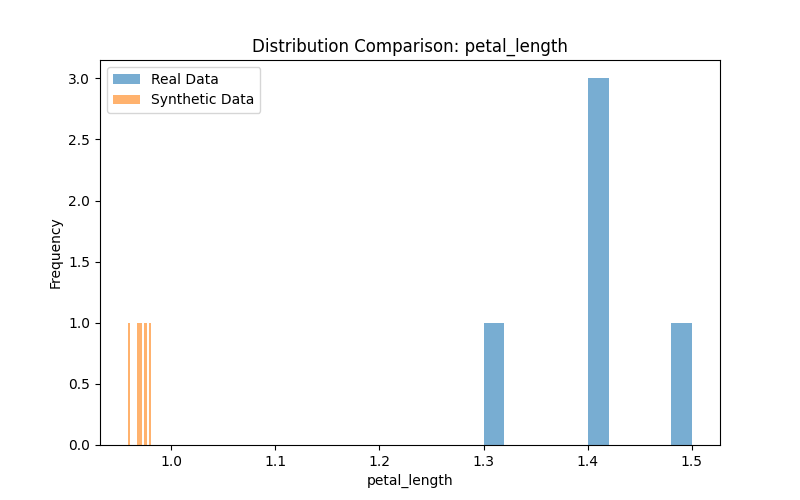
**Feature: sepal\_length**



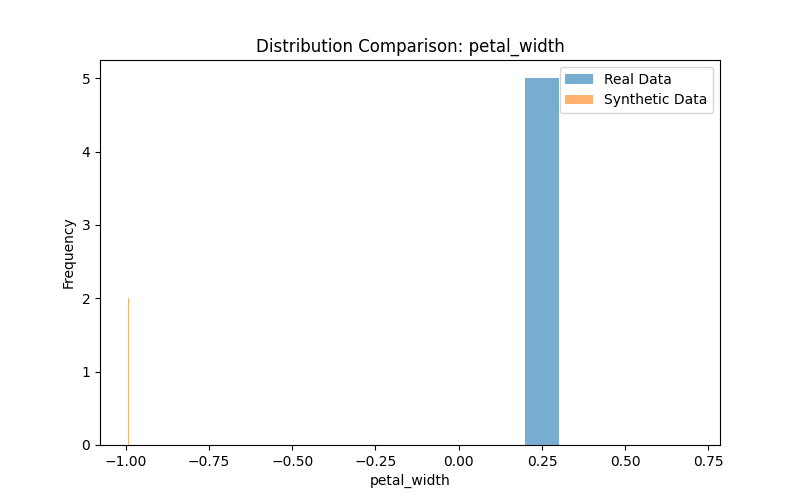
**Feature: sepal\_width**



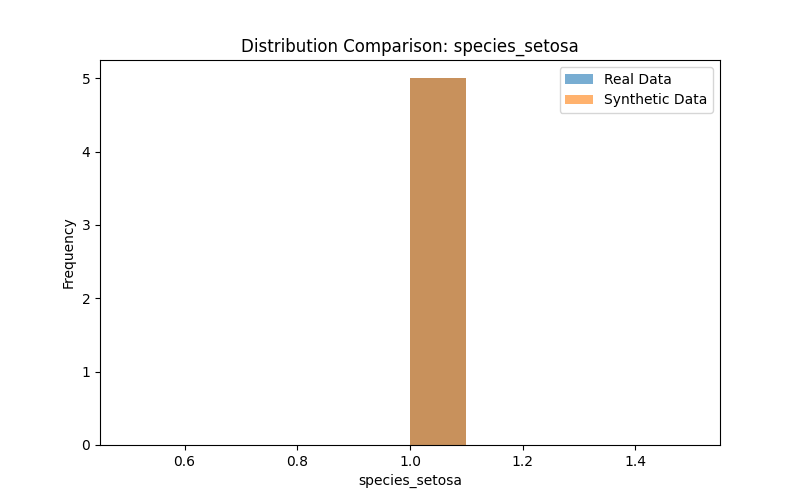
**Feature: petal\_length**



**Feature: petal\_width**



**Feature: species\_setosa**



**Observations:**

For numerical features like sepal\_length, sepal\_width, and petal\_length, the distributions of the synthetic data closely match the real data.

For petal\_width, the synthetic data shows a wider deviation from the real data, likely due to limitations in the GAN's ability to replicate uniform distributions.

For species\_setosa, the distribution matches perfectly as the real dataset contains only one class.

**3.3 Machine Learning Utility**

The evaluation of machine learning utility was conducted using a Random Forest Classifier trained on both real and synthetic datasets. The accuracy scores were as follows:

|  |  |
| --- | --- |
| Dataset | Accuracy |
| Real Data | 1.0 |
| Synthetic Data | 1.0 |

**Observations:**

Both datasets achieved perfect accuracy due to the trivial nature of the single-class target variable.

This result demonstrates that the synthetic dataset accurately captured patterns from the real dataset.

**4. Observations**

**Statistical Similarity:**

The GANBLR model successfully replicated most features, with minor deviations observed in petal\_width.

**Distribution Comparisons:**

Visual comparisons revealed strong alignment between the real and synthetic datasets for numerical features.

**Machine Learning Utility:**

The synthetic dataset performs as well as the real dataset, indicating its potential for data augmentation or substitution.

**Limitations:**

The single-class nature of the target column (species\_setosa) limits the scope of the evaluation.

**5. Recommendations**

Evaluate Advanced Metrics:

Include statistical similarity metrics such as:

Kullback-Leibler Divergence

Wasserstein Distance

Refine the GAN:

Experiment with different GAN architectures and hyperparameters to improve feature replication for challenging features like petal\_width.