Synthetic Data Generation of EHR Using CTGAN, Transformers and Diffusion Models

Team Members

Arjan Sapkota	(THA077BCT012)
Girban Adhikari	(THA077BCT017)
Jivan Acharya	(THA077BCT019)
Subarna Ghimire	(THA077BCT043)

Supervised By:

Er. Umesh Kanta Ghimire

HOD

Department of Electronics and Computer Engineering
Institute of Engineering, Thapathali Campus

July 22, 2024

Presentation Outlines

- Motivation
- Objectives
- Scope of Project
- Project Applications
- Methodology
- Results
- Discussion of Results
- List of Remaining Tasks
- References

Motivation

- Increasing challenges in leveraging data for AI applications
  - Growing AI model complexity demands larger, high-quality datasets
- Traditional data collection is costly and time-intensive
  - Gathering and processing real-world data requires significant resources
- Ethical and privacy concerns with real data
  - Real data use risks privacy violations and ethical issues

Objectives

- To evaluate the effectiveness of CTGAN, Transformers, and Diffusion Models in generating synthetic EHR data

- To compare the quality and performance of synthetic data from each model for various ML and DL tasks

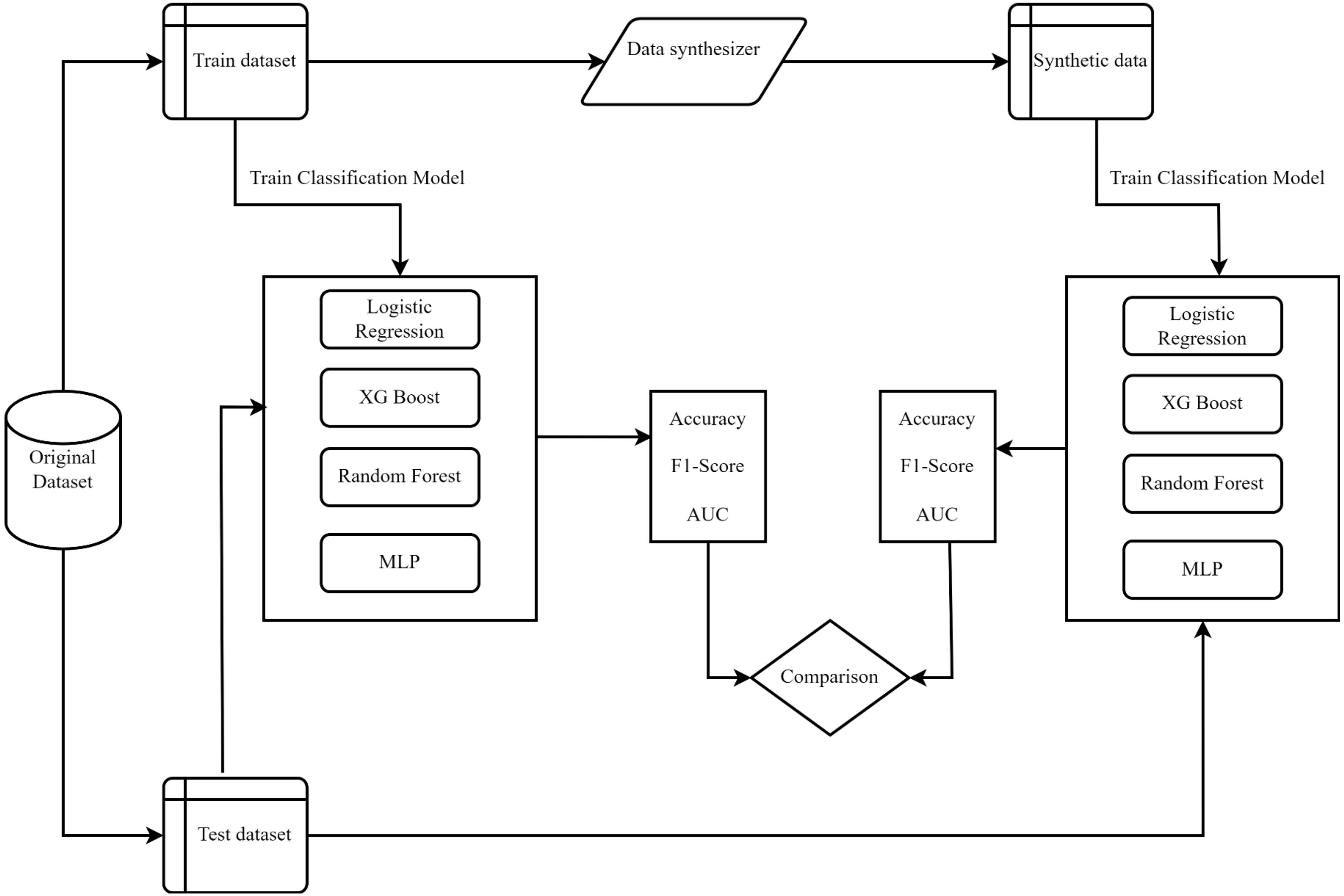
Scope of Project

- Project Capabilities:
  - Generate diverse synthetic data for health related datasets
  - Replace sensitive data to ensure privacy compliance
  - Improve AI model accuracy with augmented synthetic data
- Project Limitations:
  - Synthetic data may lack perfect realism, affecting model performance
  - High-quality generation is computationally intensive and resource-demanding
  - Regulatory bodies may not accept synthetic data for all applications

Project Applications

- Privacy-Preserving Applications
  - Substituting sensitive data with synthetic equivalents to mitigate privacy risks
  - Enhancing AI model training without compromising sensitive health/financial data
- AI Model Training and Performance
  - Augmenting existing datasets with synthetic data to boost model accuracy
  - Facilitating faster iteration and deployment of AI solutions in various fields
- Educational and Training Purposes
  - Providing realistic synthetic datasets for training researchers, students, and professionals
  - Enabling practical experimentation with accessible and diverse datasets

Methodology – [1]
(System Implementation Diagram)



Methodology – [2] (Working Principle)

- Start with the original dataset.
- Split the dataset into training and test datasets.
- Train machine learning models (Logistic Regression, XGBoost, Random Forest, MLP) on the original training dataset.
- Generate synthetic data using a data synthesizer trained on the original training dataset.

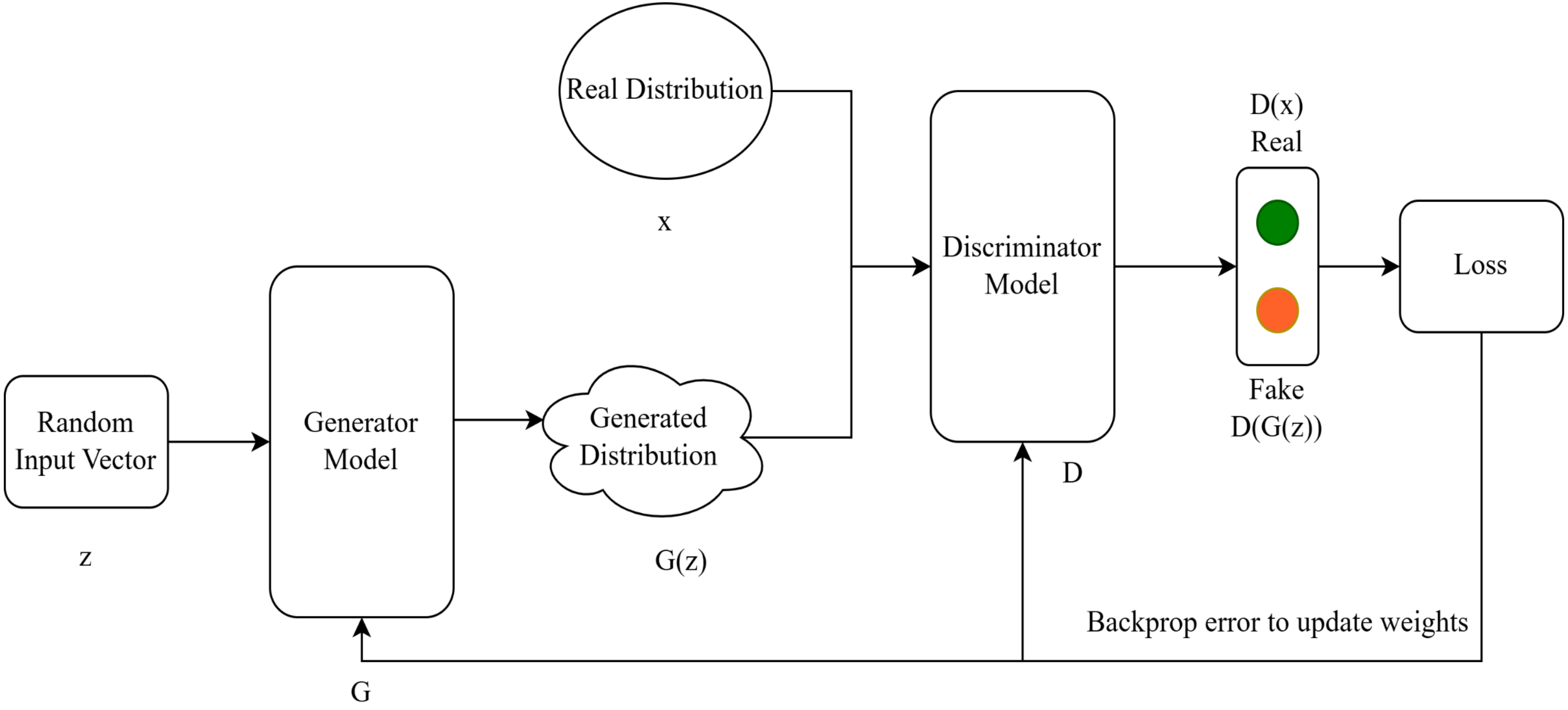
Methodology – [3]
(Working Principle)

- Train machine learning models (Logistic Regression, XGBoost, Random Forest, MLP) on the synthetic dataset.
- Evaluate models trained on both the original and synthetic datasets using Accuracy, F1-Score, and AUC metrics.
- Compare the performance of models trained on original data and synthetic data.

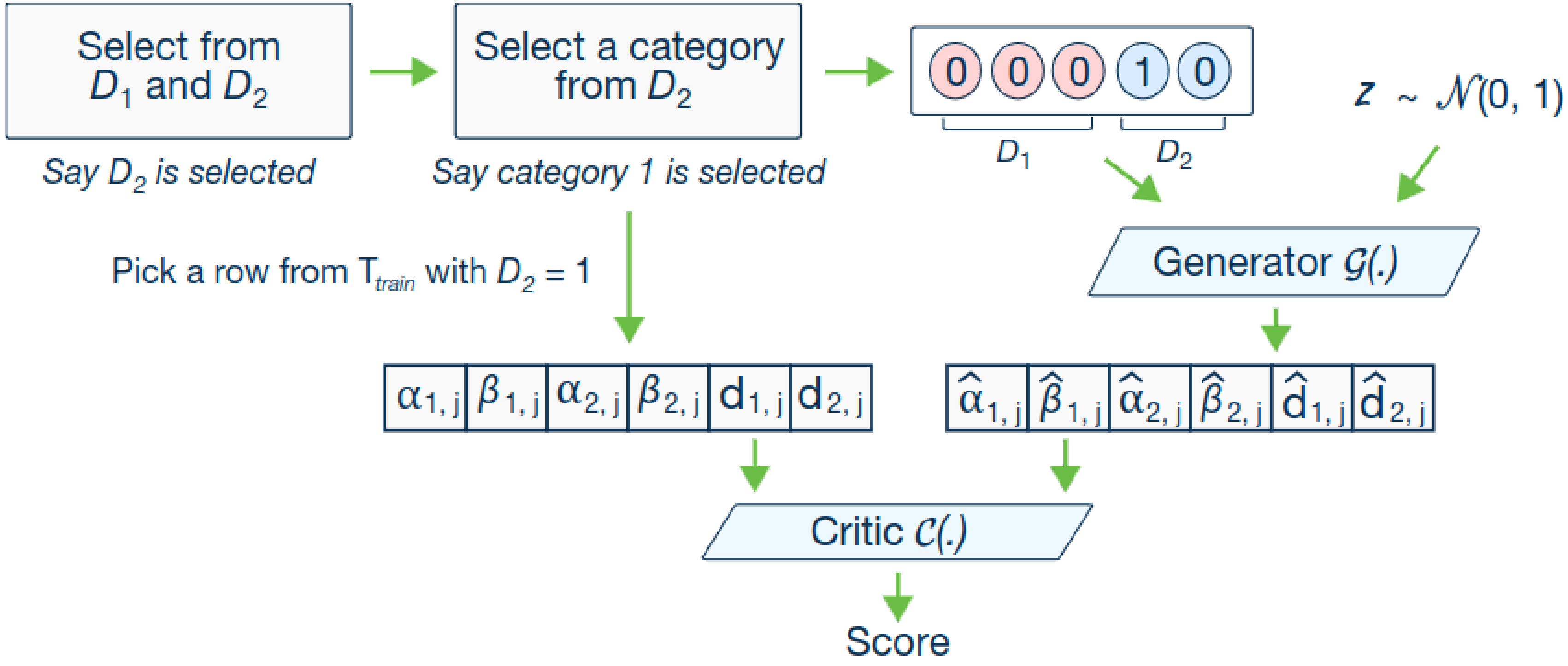
Methodology – [4]
(Data Synthesizers)

- CTGAN
- Transformers based model
- Diffusion based model

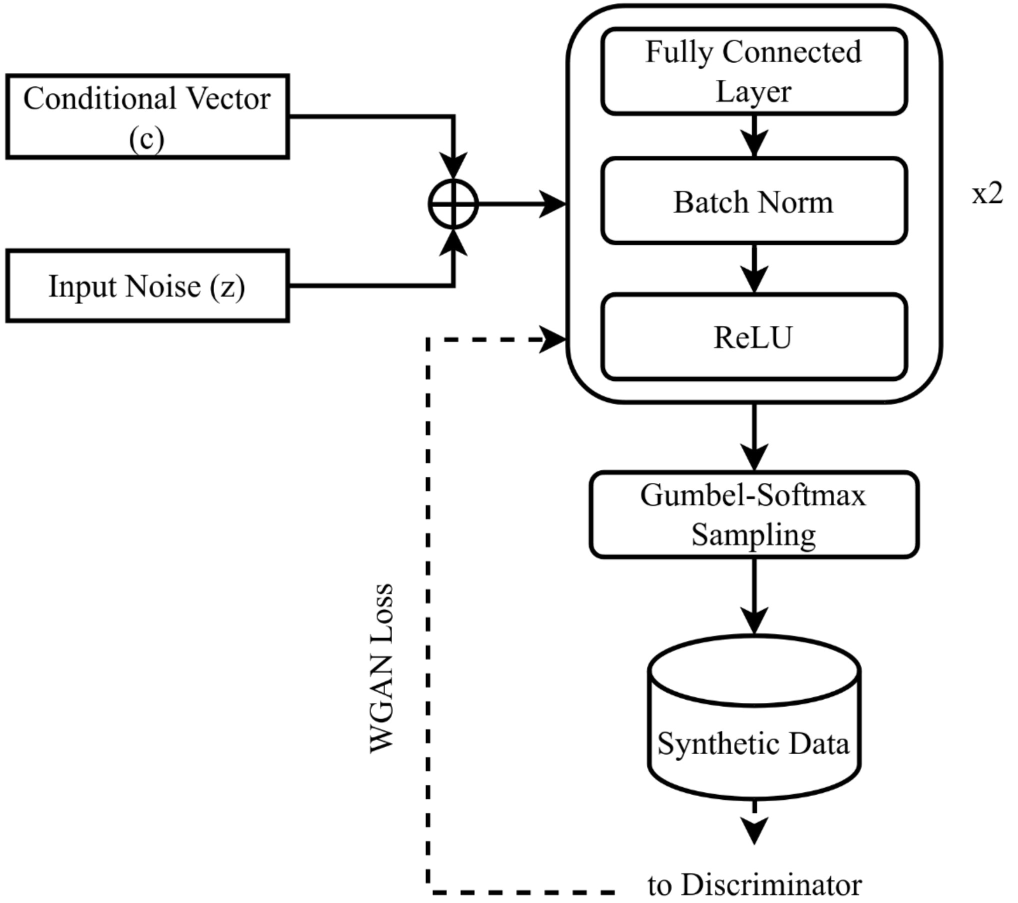
Methodology – [5] (Architecture of GAN)



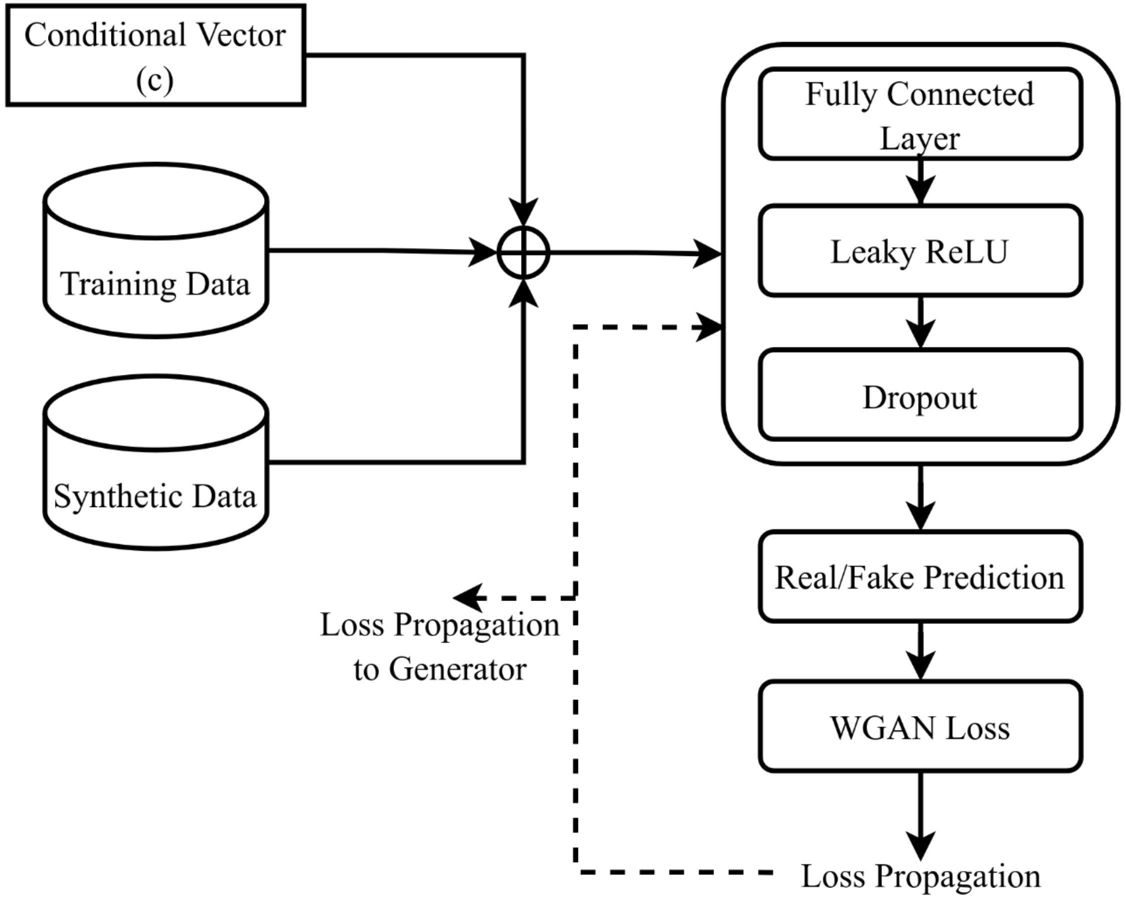
Methodology – [6]
(Architecture of CTGAN)



Methodology – [7]
(Generator of CTGAN)



Methodology – [8] (Discriminator of CTGAN)



Methodology – [9]
(Hardware Requirements)

- Processor:
  - NVIDIA Tesla K80, P100, or T4 (Google Colab)
  - NVIDIA Tesla P100 (Kaggle)
- RAM:
  - Up to 25 GB (Google Colab)
  - 13 GB (Kaggle)
- Persistent Storage:
  - 5 GB per notebook (Kaggle)
- GPU Access:
  - Free access to powerful GPUs (Google Colab)

Methodology – [10]
(Software Requirements)

- Programming Languages: Python
- Development Environments and IDEs: Jupyter Notebook, Google Colab, Kaggle Kernels
- Data Processing and Analysis: Pandas, NumPy, Scikit-learn
- Deep Learning Frameworks: TensorFlow, Keras, PyTorch
- Synthetic Data Generation: GANs - TensorFlow and PyTorch
- Model Training and Evaluation: TensorBoard, Weights & Biases
- Data Storage and Management: Google Drive, Kaggle Datasets
- Version Control: GitHub

Dataset Exploration – [1] (Pima Indian Diabetes Dataset)

Attribute	Details
Dataset Name	Pima Indian Diabetes Dataset
Dataset Type	Tabular
Source	National Institute of Diabetes and Digestive and Kidney Diseases
Size	768 rows
Information Covered	Medical predictor variables and one target variable, Outcome
Context	The dataset includes diagnostic measurements to predict diabetes in female Pima Indians at least 21 years old.
Predictor Variables	Number of pregnancies, BMI, insulin level, age, and other medical measurements

Dataset Exploration – [2] (Indian Liver Patient Dataset)

Attribute	Details
Dataset Name	Indian Liver Patient Dataset
Dataset Type	Tabular
Source	Medical Records
Size	583 rows
Information Covered	Age, Gender, Total Bilirubin, Direct Bilirubin, Total Proteins, Albumin, A/G Ratio, SGPT, SGOT, Alkphos, and Selector
Context	Records of 416 patients diagnosed with liver disease and 167 patients without liver disease
Response	The class label 'Selector' indicating the presence or absence of liver disease

Dataset Exploration – [3]
(Stroke Prediction Dataset)

Attribute	Details
Dataset Name	Stroke Prediction Dataset
Dataset Type	Tabular
Source	Confidential Source (Use only for educational purposes)
Size	5110 rows
Information Covered	Unique patient identifiers, demographic information, health conditions, lifestyle factors, and stroke occurrence
Context	Each row provides relevant information about a patient, used to predict the likelihood of a stroke based on various input parameters like gender, age, diseases, and smoking status.
Attributes	id, gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status, stroke

Results (Pima Dataset) – [1]

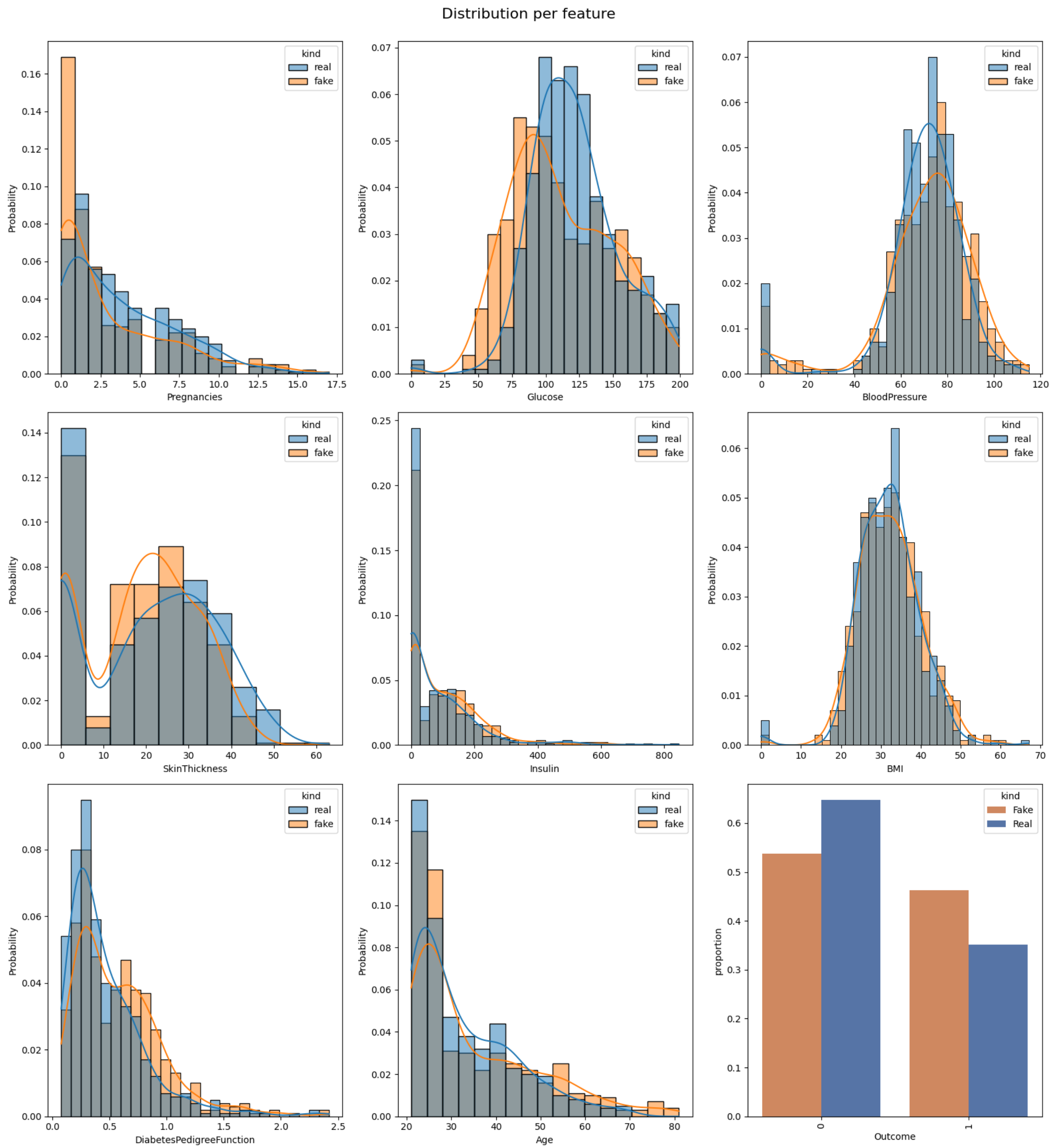
Real

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

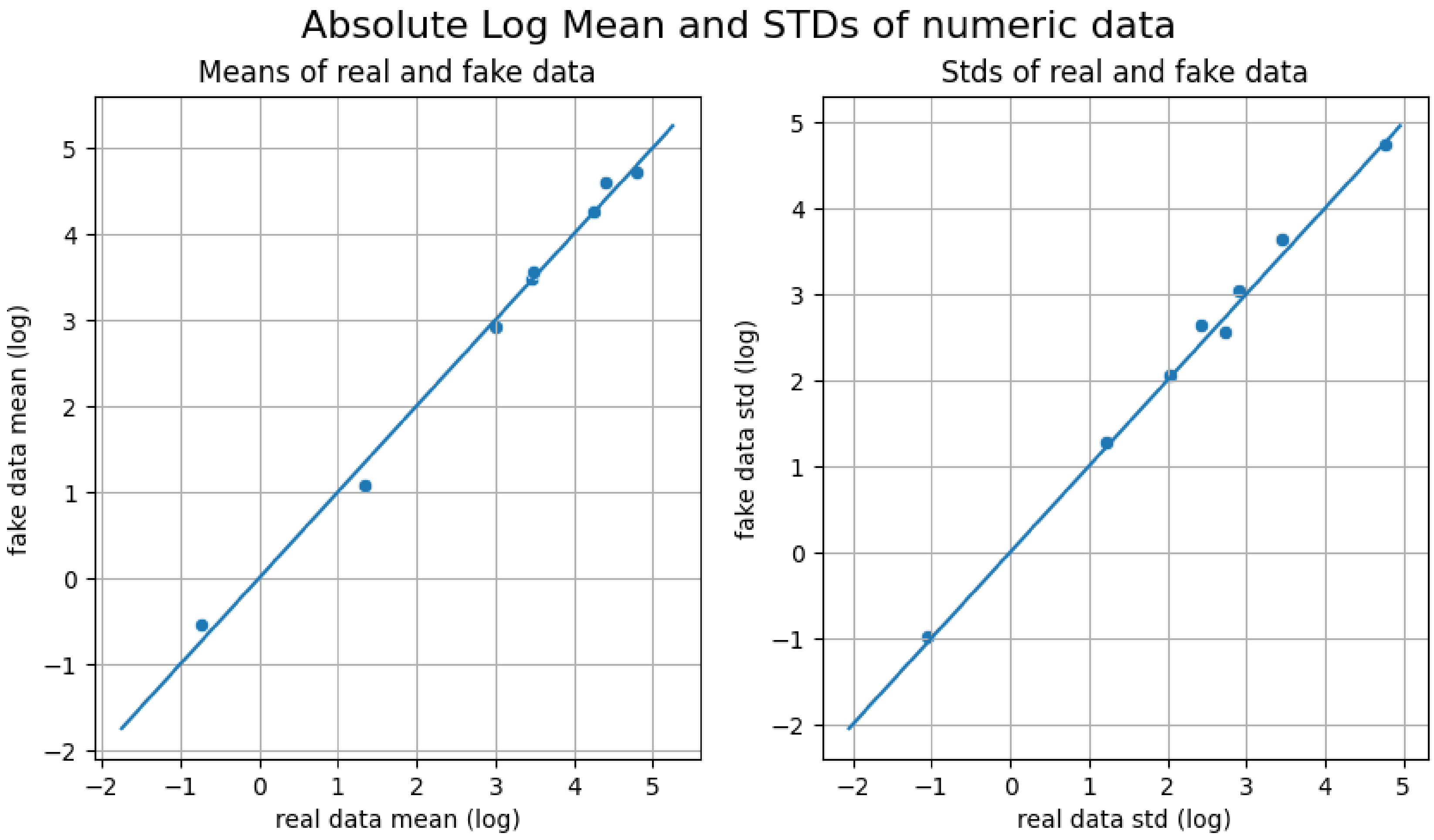
Synthetic

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	500.000000	500.000000	500.00000	500.000000	500.000000	500.000000	500.000000	500.000000	500.000000
mean	2.928000	112.278000	71.06800	18.712000	99.646000	32.351200	0.581848	35.018000	0.462000
std	3.584482	38.044493	20.90755	12.896577	114.092624	7.803099	0.378371	14.112406	0.499053
min	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	0.000000	84.000000	63.00000	3.750000	9.000000	26.875000	0.292750	24.000000	0.000000
50%	1.000000	106.000000	74.00000	20.000000	74.500000	31.900000	0.520500	28.000000	0.000000
75%	5.000000	142.000000	84.00000	28.000000	157.250000	37.200000	0.791750	43.000000	1.000000
max	16.000000	199.000000	115.00000	63.000000	734.000000	60.500000	2.420000	81.000000	1.000000

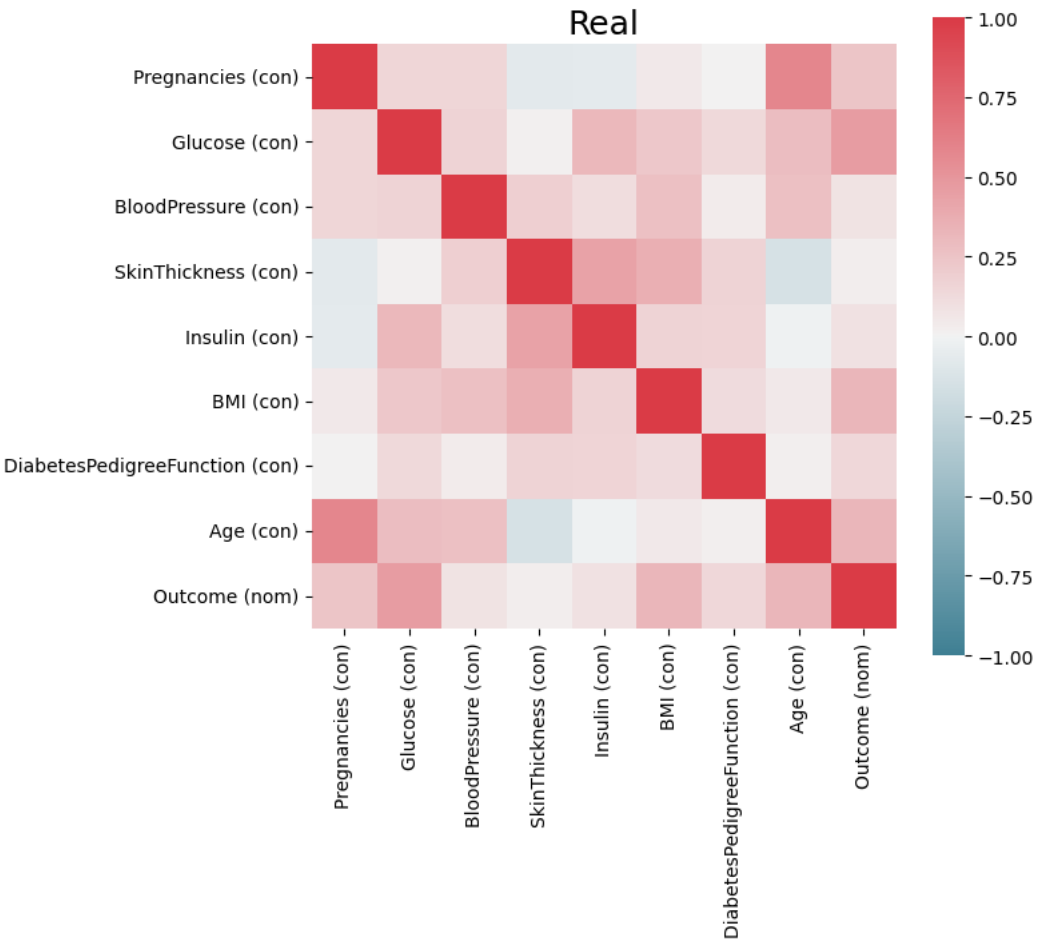
Results – [2]


Distribution per feature
kind
real
fake
Probability
Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age
kind
Fake
Real
proportion
Outcome

Results – [3]


Absolute Log Mean and STDs of numeric data
Means of real and fake data
fake data mean (log)
real data mean (log)
Stds of real and fake data
fake data std (log)
real data std (log)

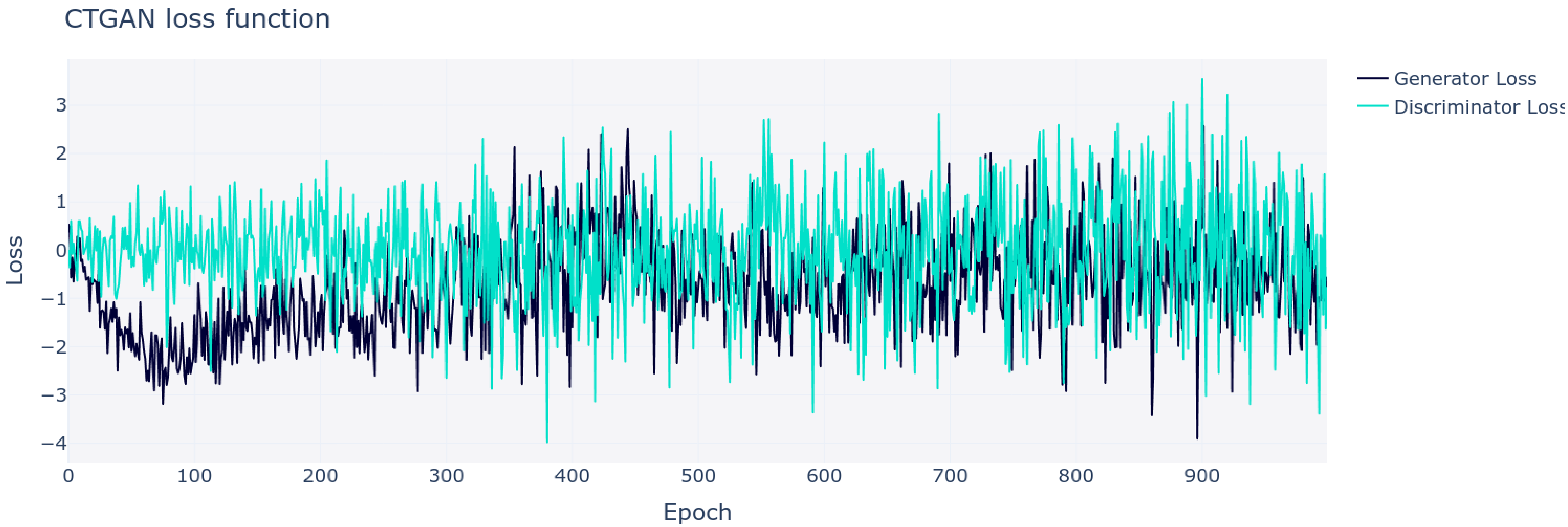
Results – [4]



Real
Pregnancies (con)
Glucose (con)
BloodPressure (con)
SkinThickness (con)
Insulin (con)
BMI (con)
DiabetesPedigreeFunction (con)
Age (con)
Outcome (nom)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

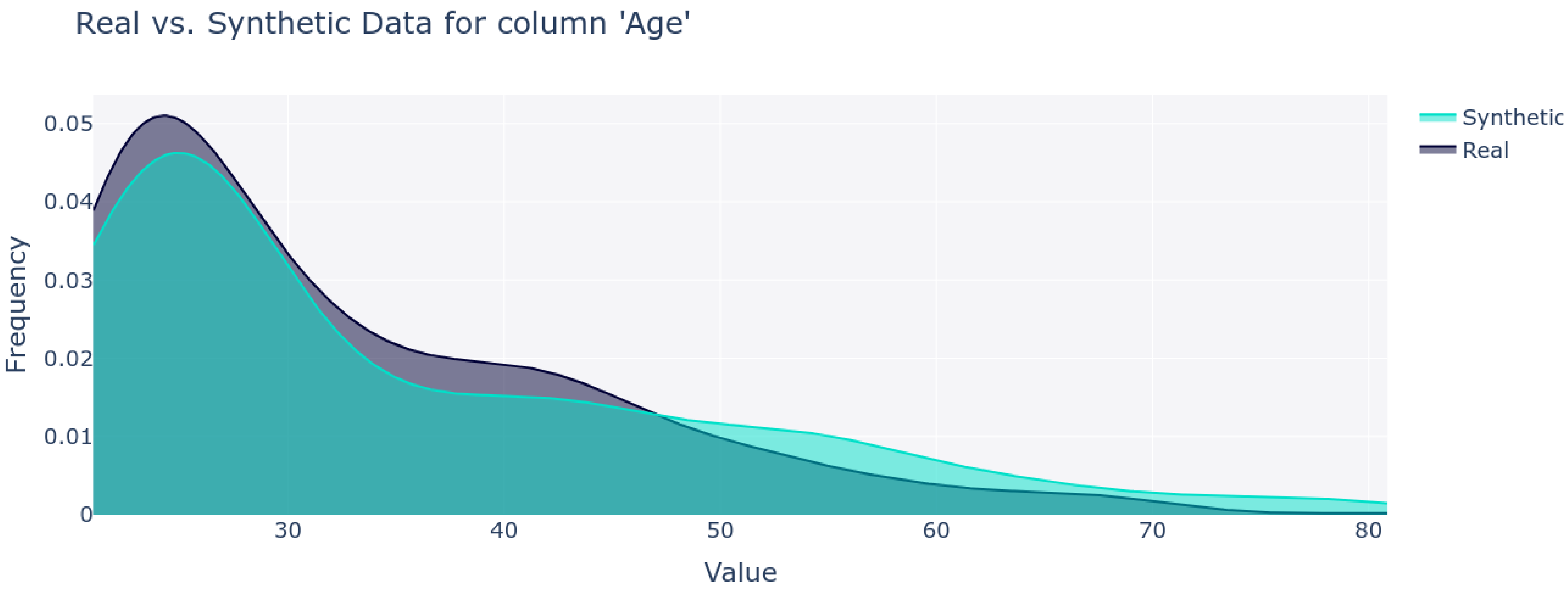
Fake
Pregnancies (con)
Glucose (con)
BloodPressure (con)
SkinThickness (con)
Insulin (con)
BMI (con)
DiabetesPedigreeFunction (con)
Age (con)
Outcome (nom)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Results – [5]

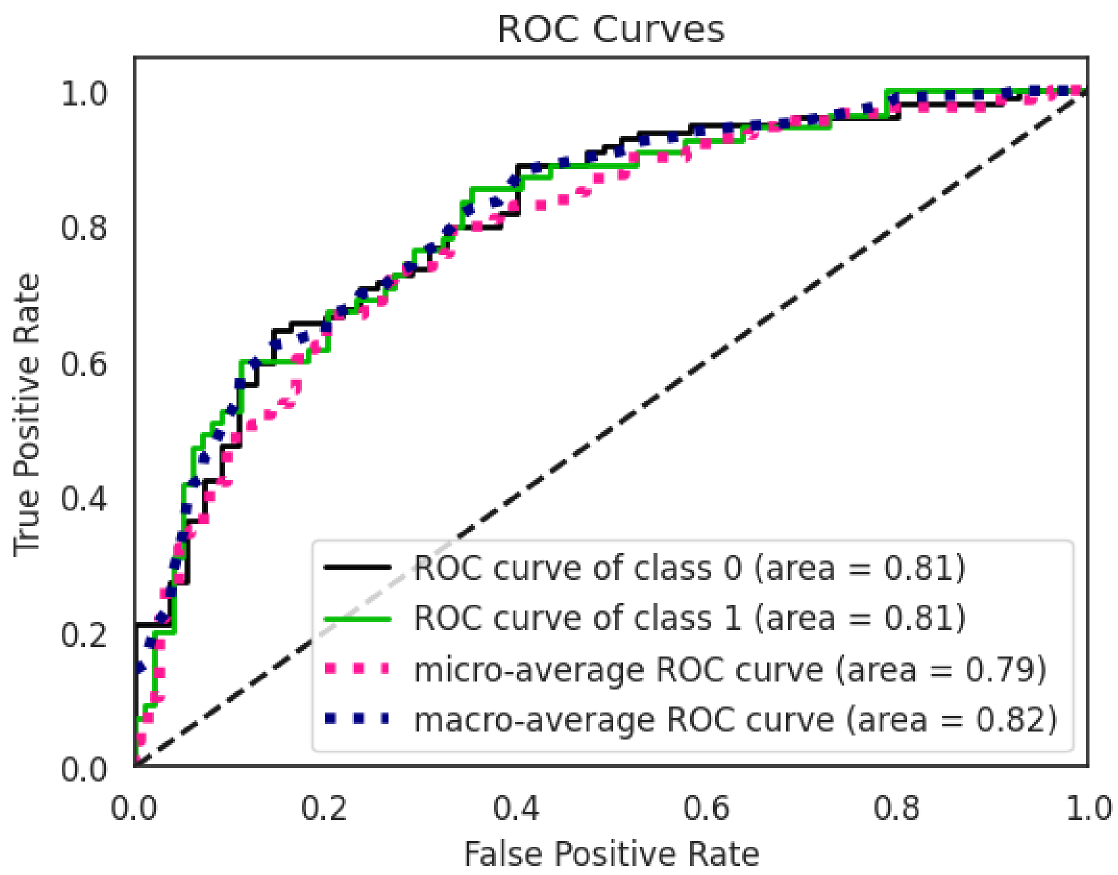


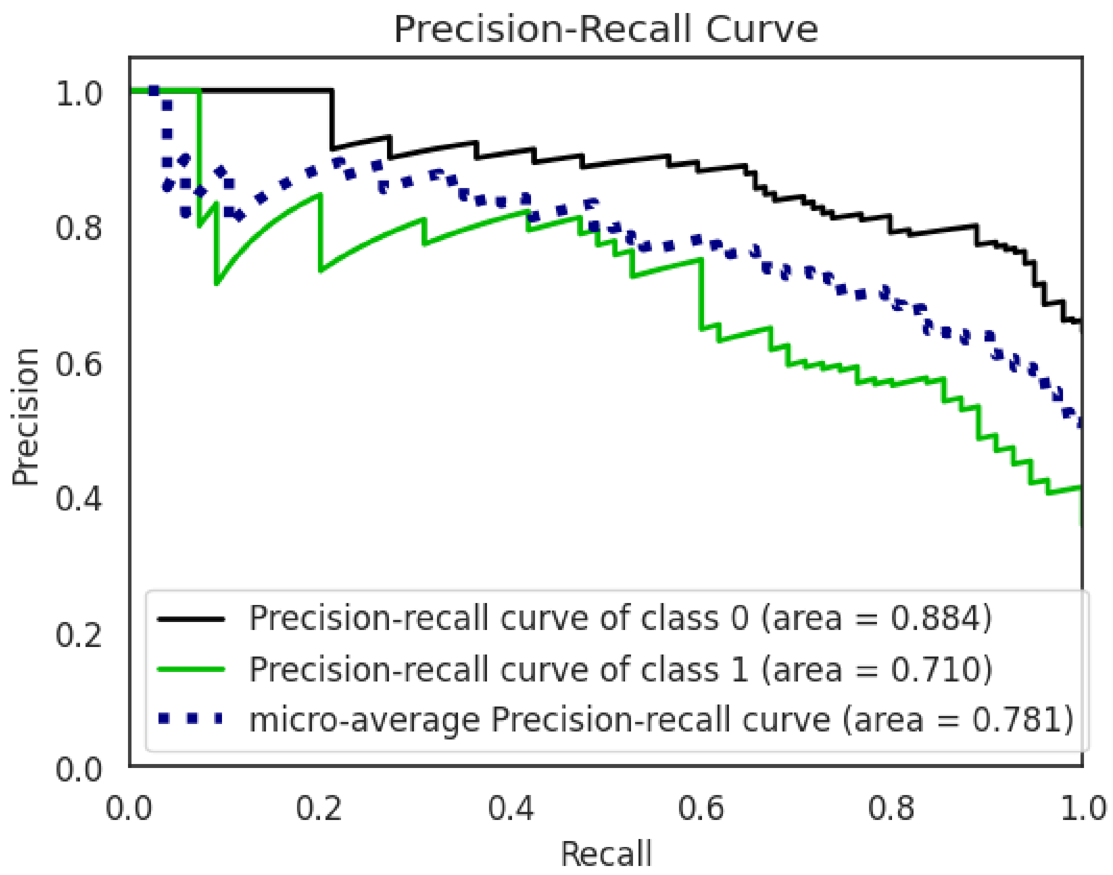
CTGAN loss function
Generator Loss
Discriminator Loss
Loss
Epoch
3
2
1
0
−1
−2
−3
−4
0
100
200
300
400
500
600
700
800
900

Results – [6]

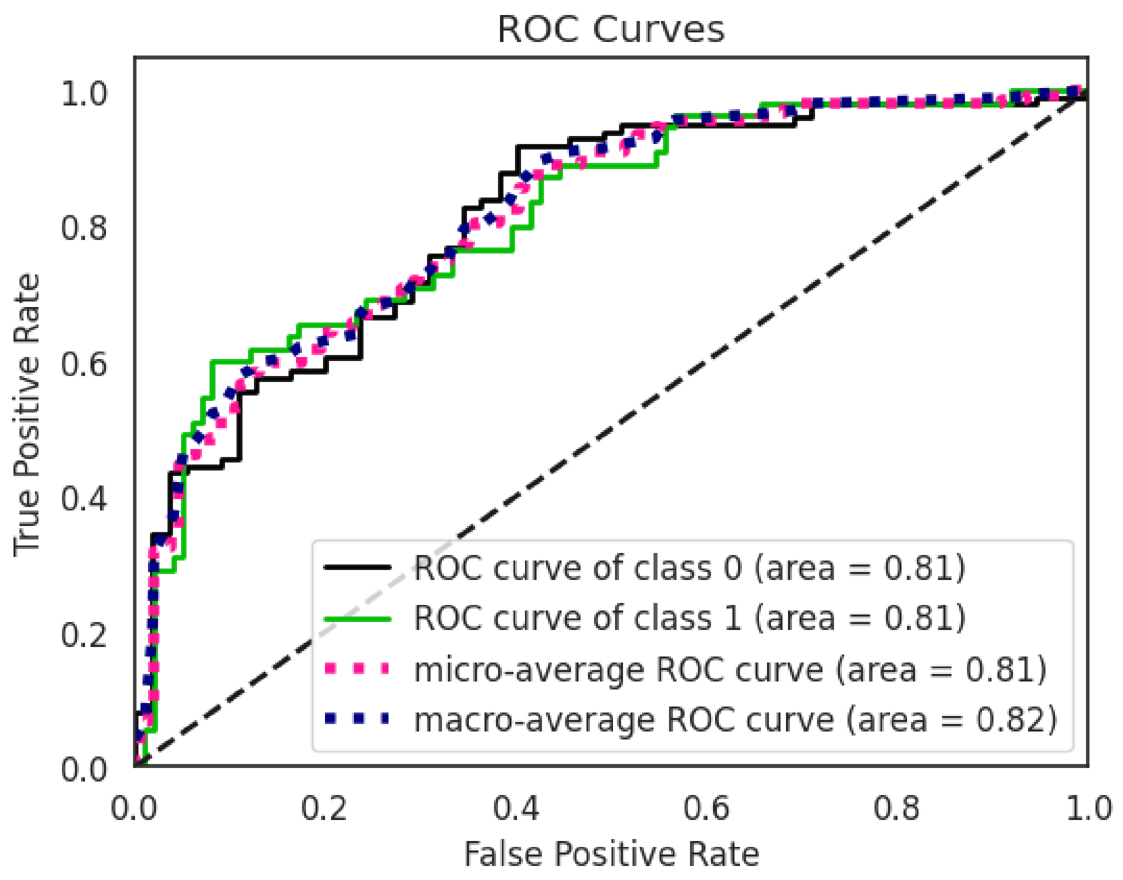

Real vs. Synthetic Data for column 'Age'
Synthetic
Real
Frequency
0.05
0.04
0.03
0.02
0.01
0
30
40
50
60
70
80
Value

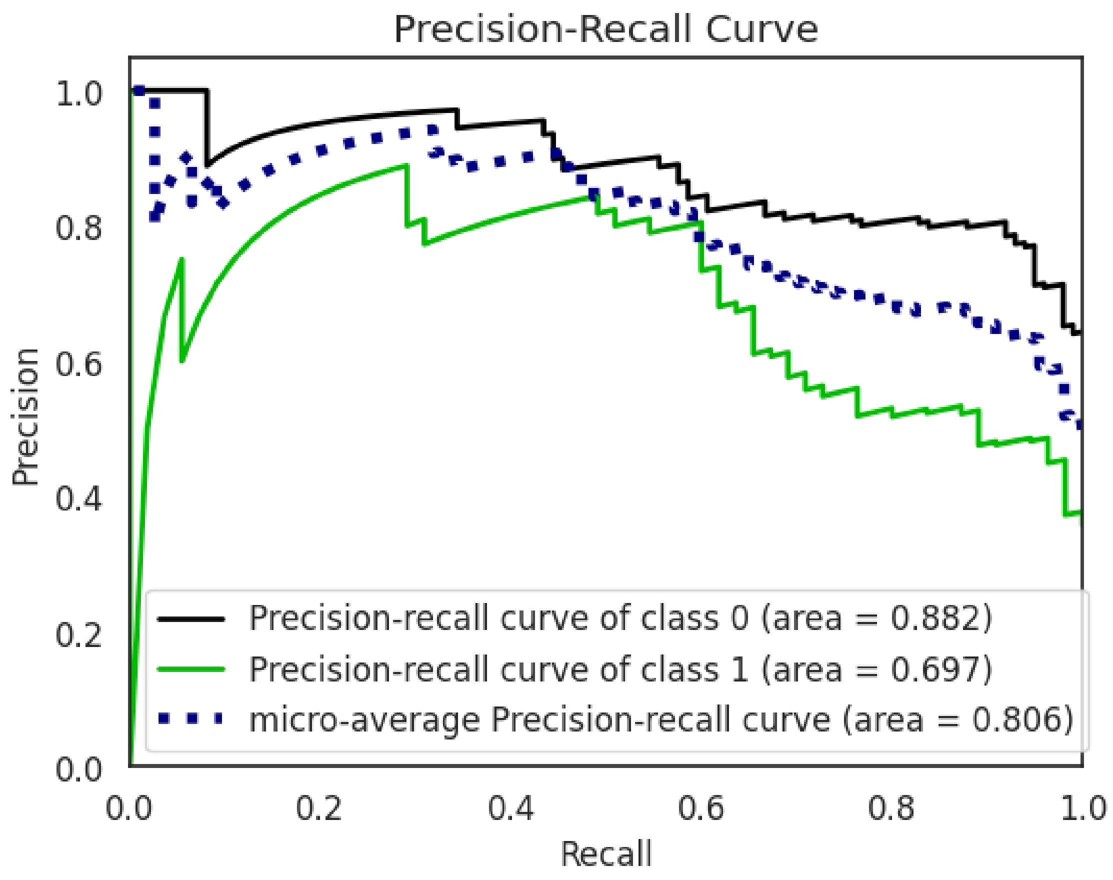
Results – [7]
(ROC & PR Curve on Real Data)





Results – [8]
(ROC & PR Curve on Synthetic Data)

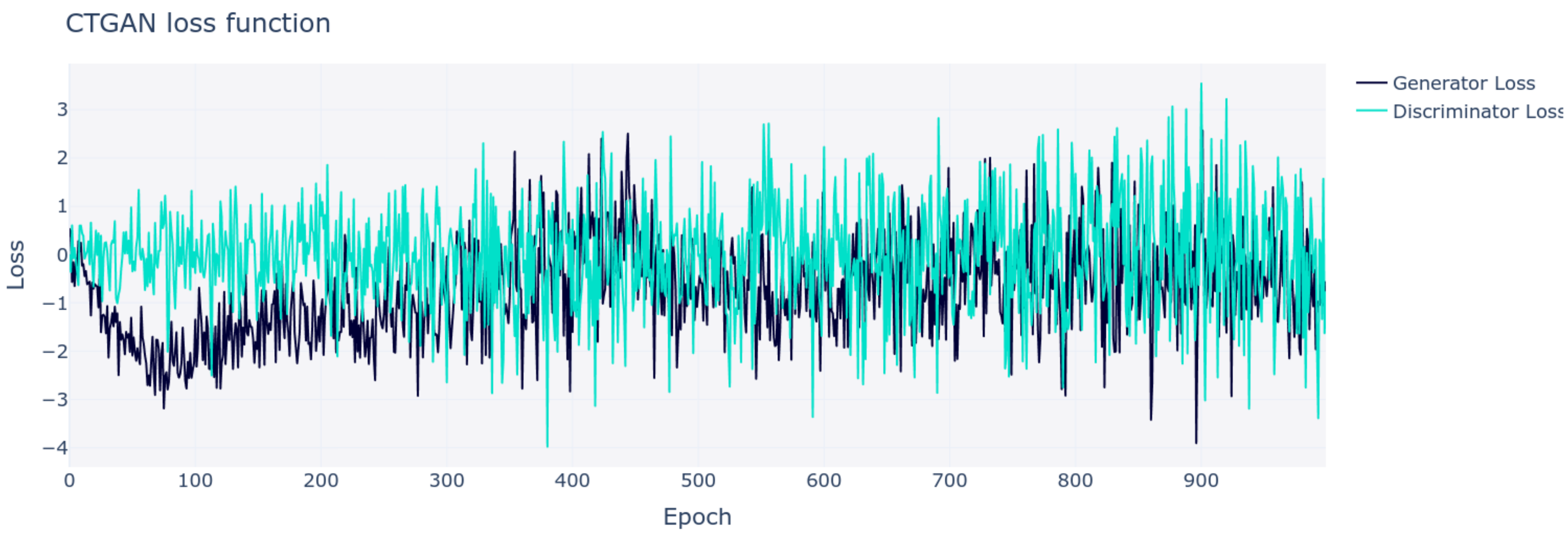




Discussion of Results – [1]

Model	Dataset Type	Accuracy	F1-Score	ROC-AUC
Logistic Regression	Real	0.71	0.71	0.81
	Synthetic	0.73	0.73	0.81
XG Boost	Real	0.75	0.75	0.79
	Synthetic	0.72	0.73	0.82
Neural Network	Real	0.73	0.73	0.77
	Synthetic	0.74	0.72	0.8
Random Forest	Real	0.77	0.77	0.83
	Synthetic	0.73	0.73	0.81

Discussion of Results – [2]



- Epochs – 1000
- Batch Size – 20
- Generator Loss = (-0.38)
- Discriminator Loss = (0.18)
- Training Set – 614 (80%)
- Test Set – 154 (20%)

List of Remaining Tasks

- Implement synthetic data generation with Transformers and Diffusion Models
- Compare performance of CTGAN, Transformer, and Diffusion Model-generated synthetic data
- Explore advanced evaluation metrics for synthetic data quality

References – [1]

[1] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *International Conference on Learning Representations (ICLR)*, 2013.

[2] I. J. Goodfellow, J. Pouget-Abadie and M. Mirza, "Generative Adversarial Networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 2672-2680

[3] L. Xu, . M. Skoularidou, A. Cuesta-Infante and . K. Veeramachaneni, "Modeling tabular data using conditional GAN," in *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

References – [2]

[4] M. Arjovsky, S. Chintala and L. Bottou, "Wasserstein GAN," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, Sydney, Australia, 2017.

[5] "Pima Indians Diabetes Database," Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

[6] Ramana,Bendi and Venkateswarlu,N.. (2012). ILPD (Indian Liver Patient Dataset). UCI Machine Learning Repository. https://doi.org/10.24432/C5D02C.

References – [3]

[7] F. Soriano, "Stroke Prediction Dataset," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.