Synthetic Data Generation of Electronic Health Records 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

Head of Department

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

August 9, 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

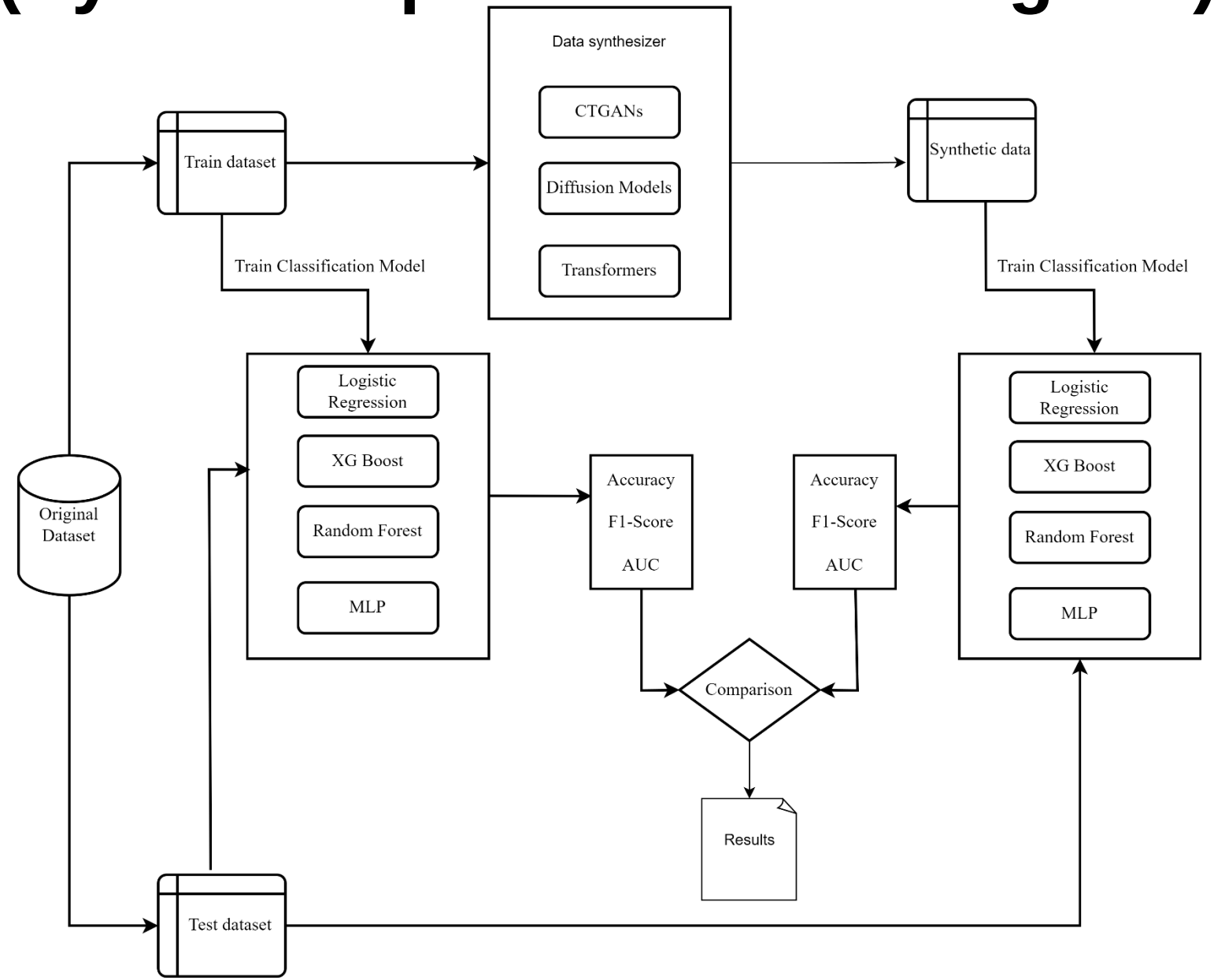
Scope of Project

- 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

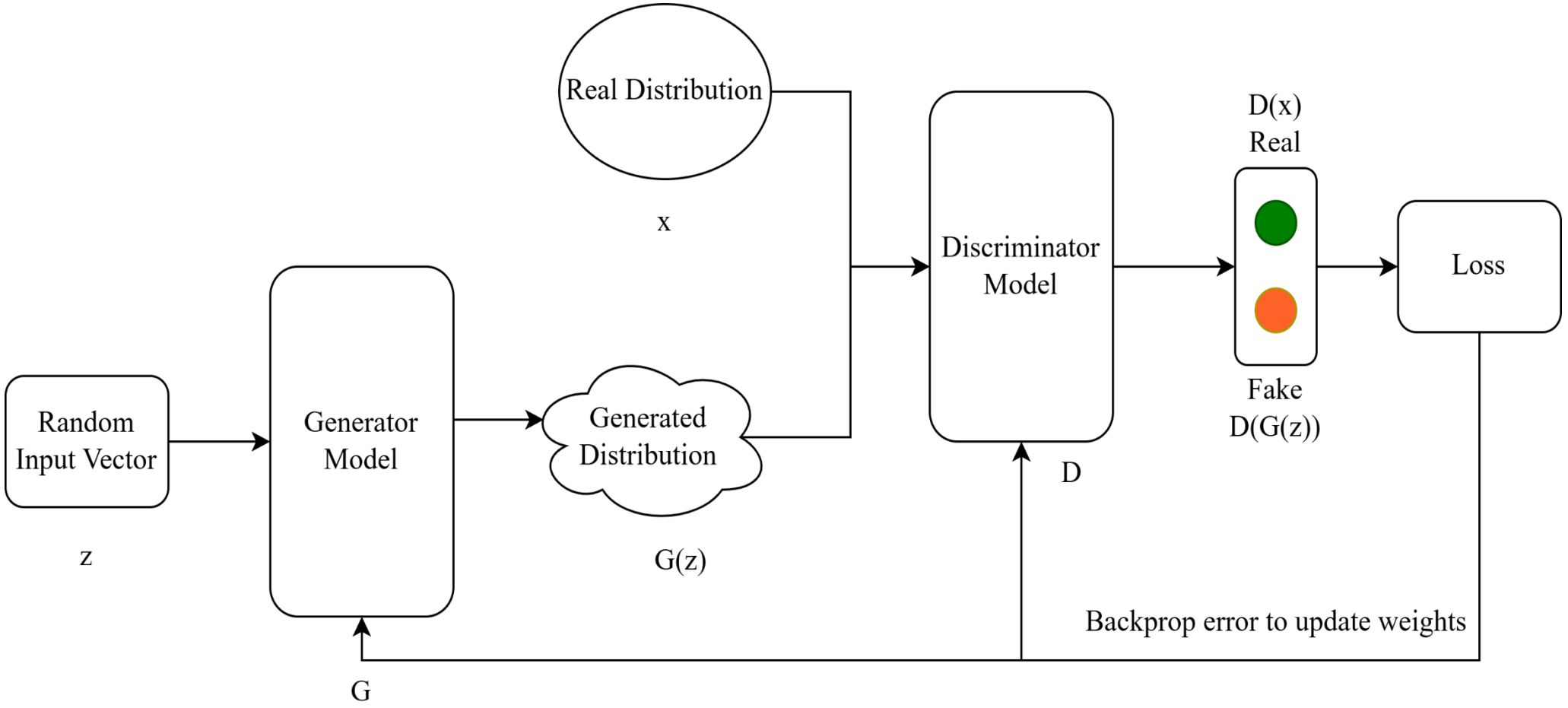
Methodology – [3]
(Working Principle)

- 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

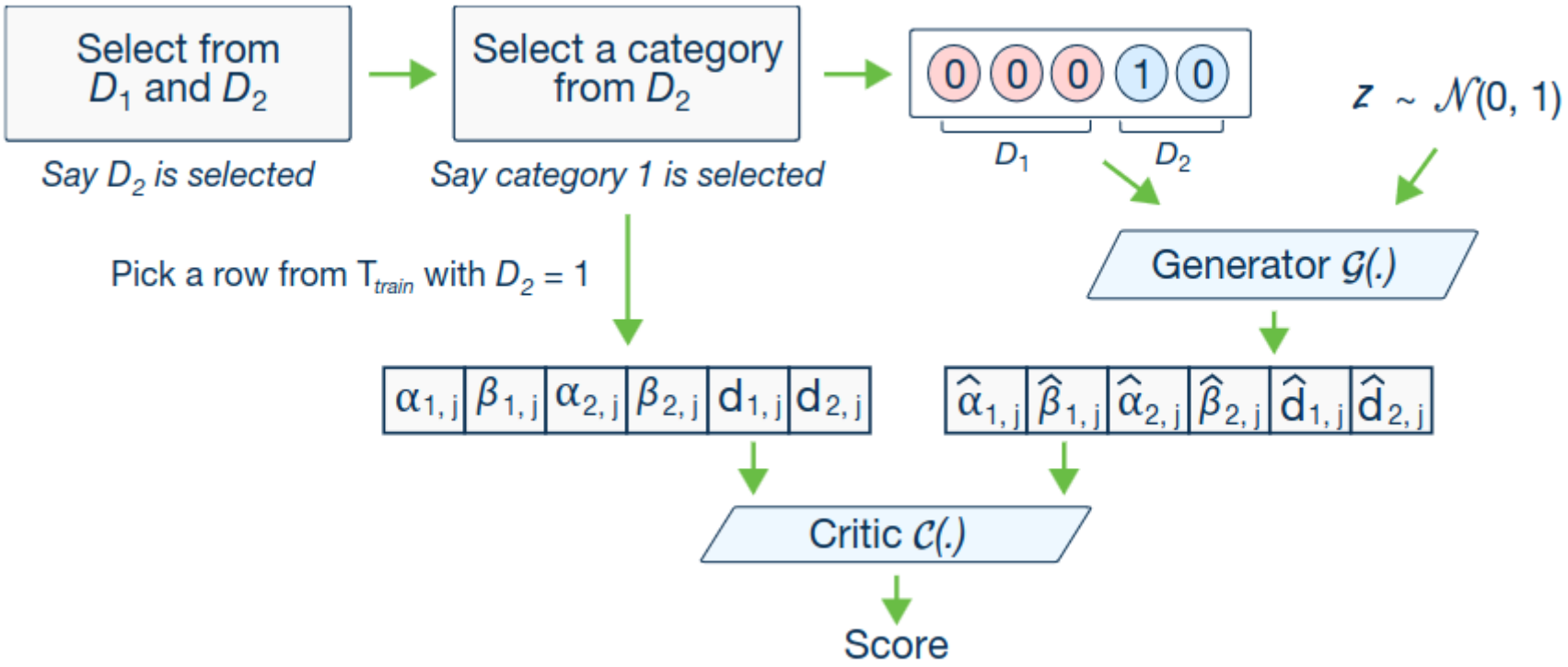
Methodology – [4]
(Data Synthesizers)

- CTGAN
- Transformers based model
- Diffusion based model

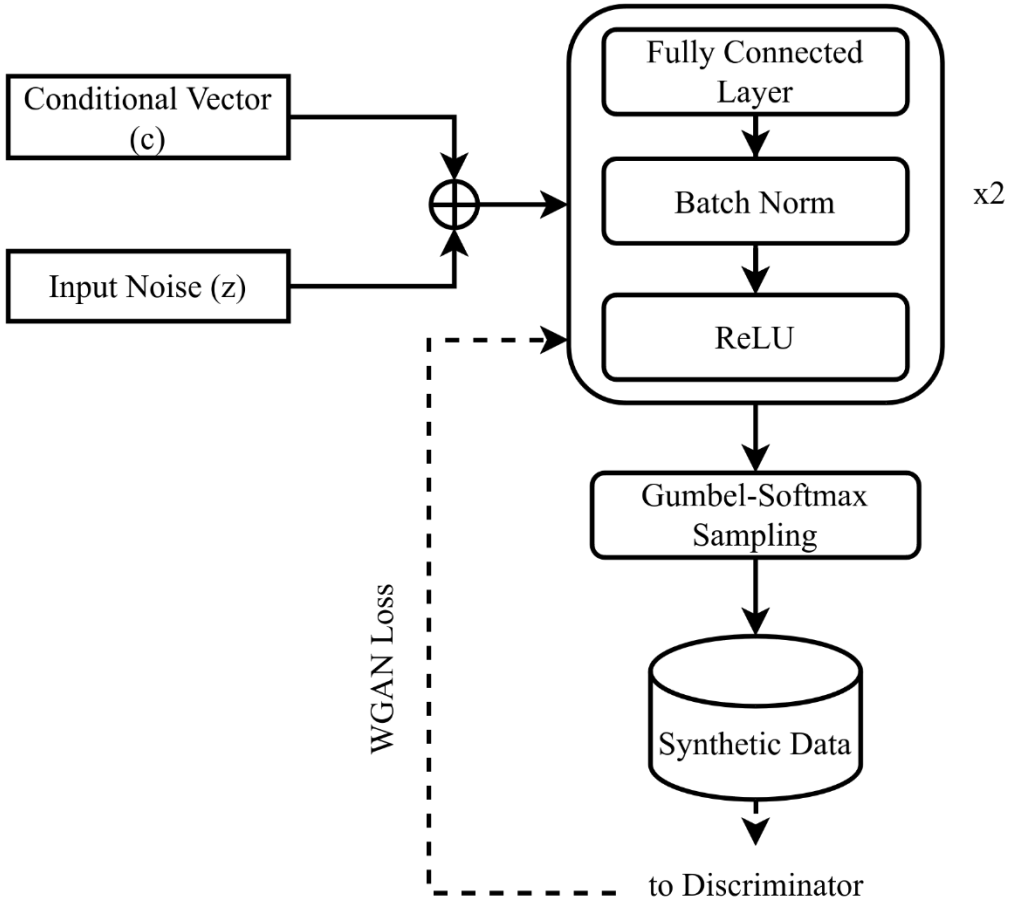
Methodology – [5] (Architecture of GAN)



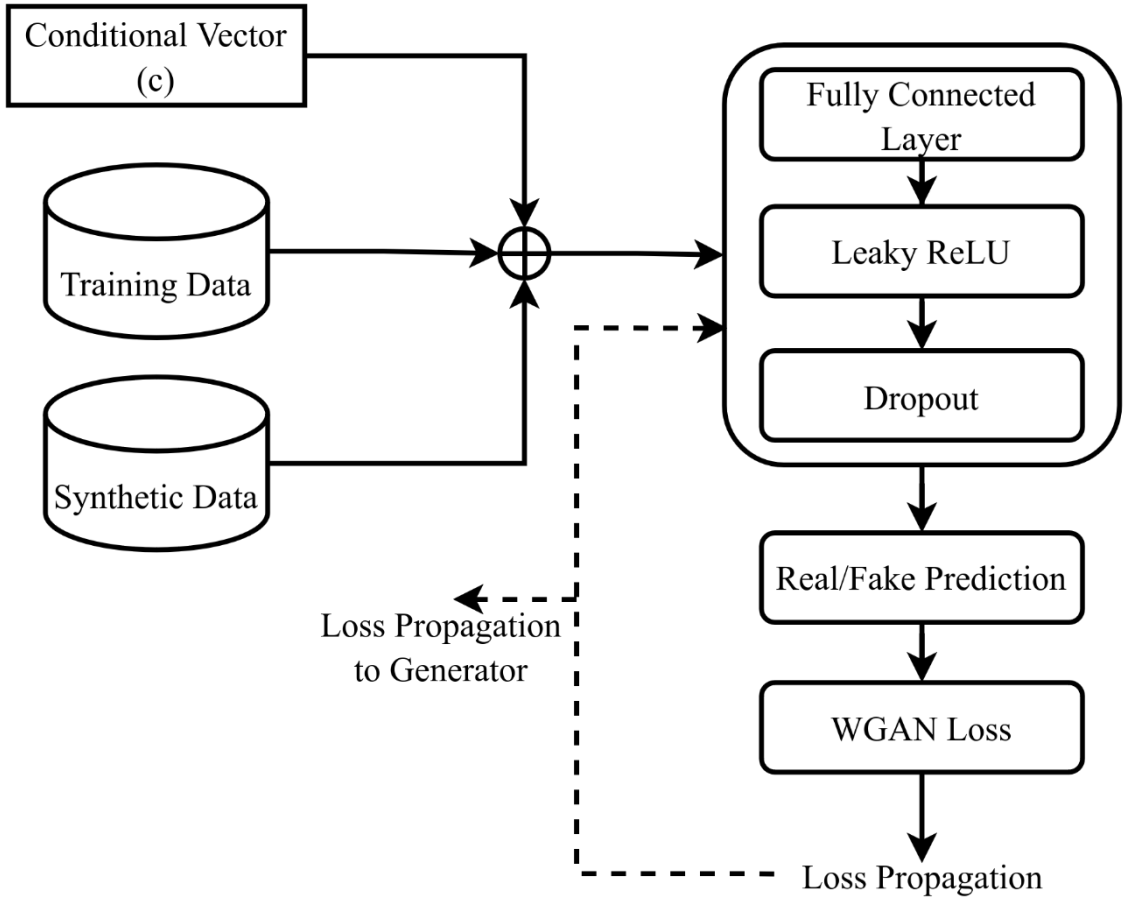
Methodology – [6] (Architecture of CTGAN)



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 Creation – [1] (MIMIC III)

- MIMIC-III (Medical Information Mart for Intensive Care III)
  - A large database comprising health-related data from ICU patients
- Collected from hospital databases
  - Includes demographics, vital signs, medications, and more
- Access Requirements
  - Training: **"Human Research"** training through the Collaborative Institutional Training Initiative (CITI) program
  - Certification: **"Data or Specimens Only Research"** through CITI
  - Access Process: **Data Use Agreement (DUA)**

Dataset Creation – [2] (MIMIC III)

- Structure
  - Tables: Over 50 tables
  - Schema: Organized into various tables such as ADMISSIONS, PATIENTS, ICUSTAYS, CHARTEVENTS, etc
  - Linkages: Tables are linked via unique identifiers (e.g., subject_id, hadm_id, icustay_id)

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 x 9
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]
(Pima Indian Diabetes Dataset)

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1

Dataset Exploration – [3]
(Indian Liver Patient Dataset)

Attribute	Details
Dataset Name	Indian Liver Patient Dataset
Dataset Type	Tabular
Source	Medical Records
Size	583 x 11
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 – [4]
(Indian Liver Patient Dataset)

	Age	Gender	TB	DB	Alkphos	Sgpt	Sgot	TP	ALB	A/G Ratio	Selector
0	65	Female	0.7	0.1	187	16	18	6.8	3.3	0.90	1
1	62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
2	62	Male	7.3	4.1	490	60	68	7.0	3.3	0.89	1
3	58	Male	1.0	0.4	182	14	20	6.8	3.4	1.00	1
4	72	Male	3.9	2.0	195	27	59	7.3	2.4	0.40	1
5	46	Male	1.8	0.7	208	19	14	7.6	4.4	1.30	1
6	26	Female	0.9	0.2	154	16	12	7.0	3.5	1.00	1
7	29	Female	0.9	0.3	202	14	11	6.7	3.6	1.10	1
8	17	Male	0.9	0.3	202	22	19	7.4	4.1	1.20	2
9	55	Male	0.7	0.2	290	53	58	6.8	3.4	1.00	1

Dataset Exploration – [5] (Stroke Prediction Dataset)

Attribute	Details
Dataset Name	Stroke Prediction Dataset
Dataset Type	Tabular
Source	Confidential Source (Use only for educational purposes)
Size	5110 x 12
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

Dataset Exploration – [6]
(Stroke Prediction Dataset)

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
5	56669	Male	81.0	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1
6	53882	Male	74.0	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
7	10434	Female	69.0	0	0	No	Private	Urban	94.39	22.8	never smoked	1
8	27419	Female	59.0	0	0	Yes	Private	Rural	76.15	NaN	Unknown	1
9	60491	Female	78.0	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1

Dataset Exploration – [7] (MIMIC-III Dataset)

Attribute	Details
Dataset Name	MIMIC-III (Medical Information Mart for Intensive Care III)
Dataset Type	Tabular
Source	Beth Israel Deaconess Medical Center, Boston, MA
Size	58976 x 16
Information Covered	Unique patient identifiers, demographic information, admission and discharge times, diagnoses, procedures, medications, vital signs, lab results, length of stay, and mortality
Attributes	Subject_ID, HADM_ID, ICUSTAY_ID, Age, ADMISSION_TYPE, MARITAL_STATUS, HOSPITAL_EXPIRE_FLAG, GENDER, NUMCALLOUT, NUMCPTEVENTS, NUMDIAGNOSIS, NUMOUTEVENTS, NUMRX, NUMPROCEVENTS, NUMMICROLABEVENTS, NUMPROC, NUMTRANSFERS, NUMINPUTEVENTS, NUMLABEVENTS, NUMNOTEVENTS

Dataset Exploration – [8] (MIMIC-III Dataset)

	ADMISSION_TYPE	MARITAL_STATUS	HOSPITAL_EXPIRE_FLAG	GENDER	NUMCALLOUT	NUMCPTEVENTS	NUMDIAGNOSIS	NUMOUTEVENTS	NUMRX
0	EMERGENCY	MARRIED	0	F	0.0	0.0	7	7.0	0.0
1	ELECTIVE	MARRIED	0	M	0.0	0.0	8	62.0	69.0
2	EMERGENCY	MARRIED	0	M	1.0	6.0	10	29.0	69.0
3	EMERGENCY	SINGLE	0	M	0.0	4.0	4	2.0	26.0
4	EMERGENCY	MARRIED	0	M	0.0	4.0	4	59.0	67.0

NUMRX	NUMPROCEVENTS	NUMMICROLABEVENTS	NUMPROC	NUMTRANSFERS	NUMINPUTEVENTS	NUMLABEVENTS	NUMNOTEV
0.0	0.0	1.0	3.0	2	6.0	91.0	
69.0	0.0	1.0	7.0	4	180.0	208.0	
69.0	4.0	1.0	1.0	5	0.0	221.0	
26.0	0.0	0.0	6.0	3	50.0	99.0	
67.0	0.0	2.0	9.0	4	483.0	315.0	

Results (Pima Dataset Description) – [1]

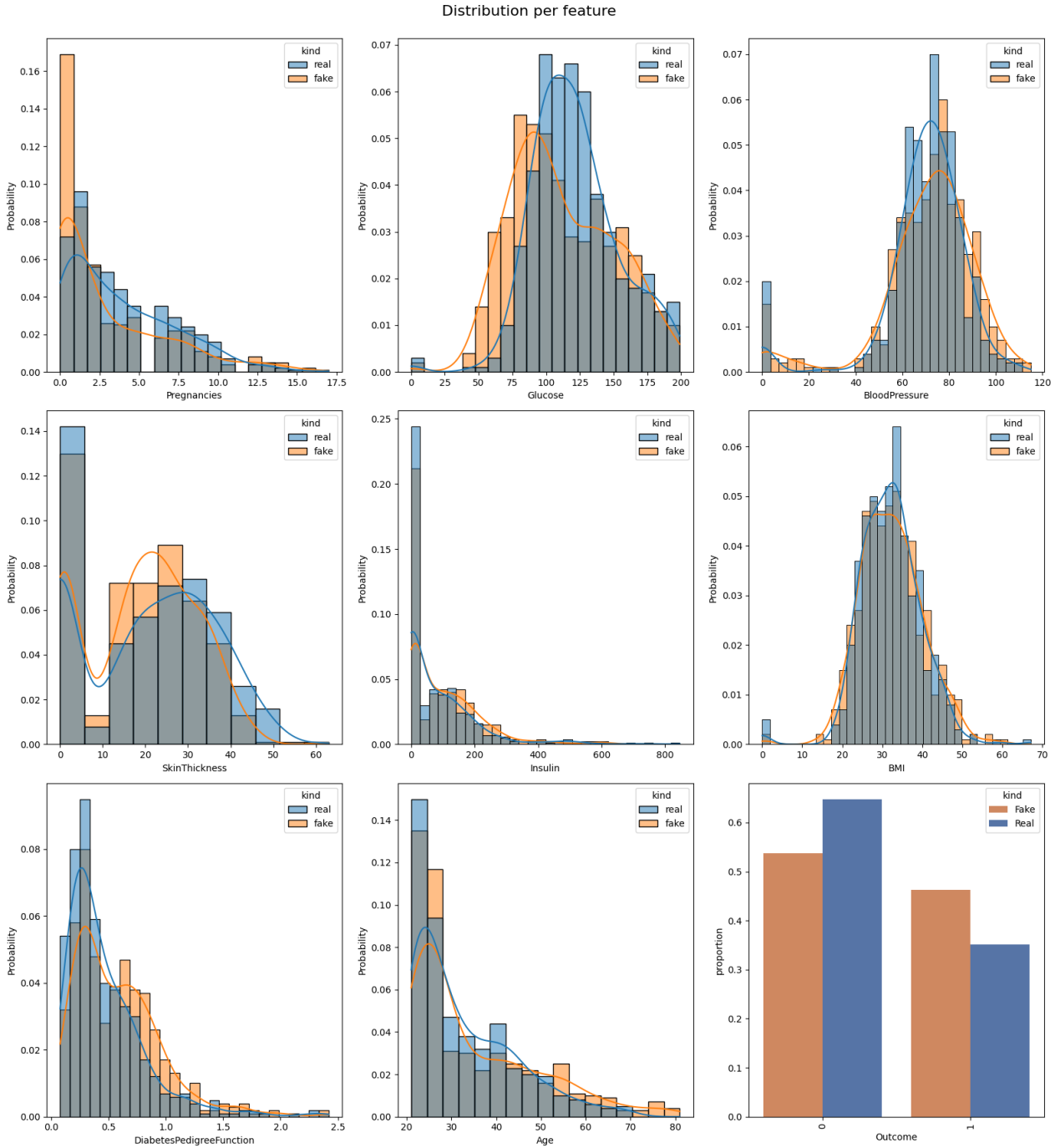
Real

	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

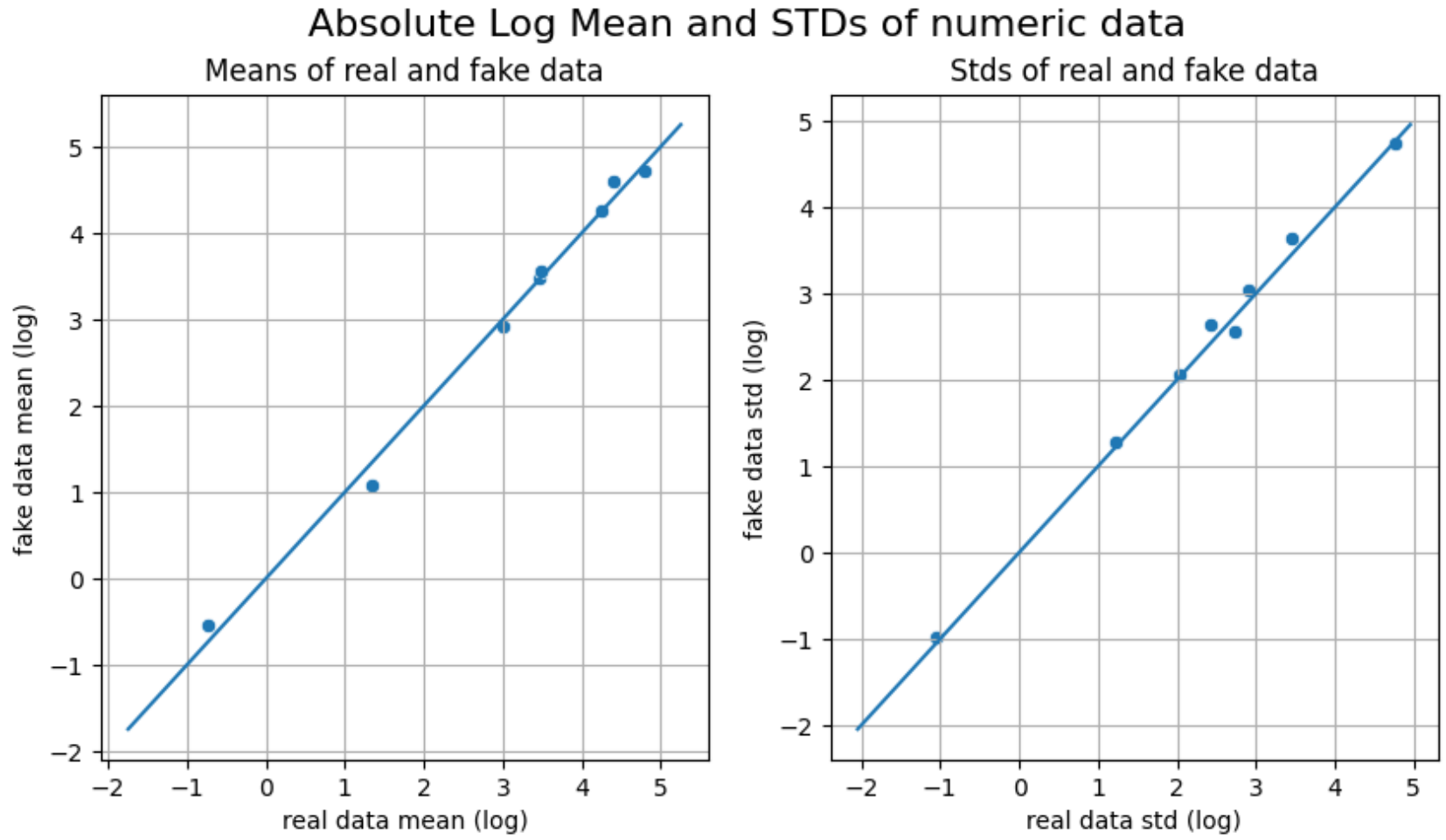
Synthetic

	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

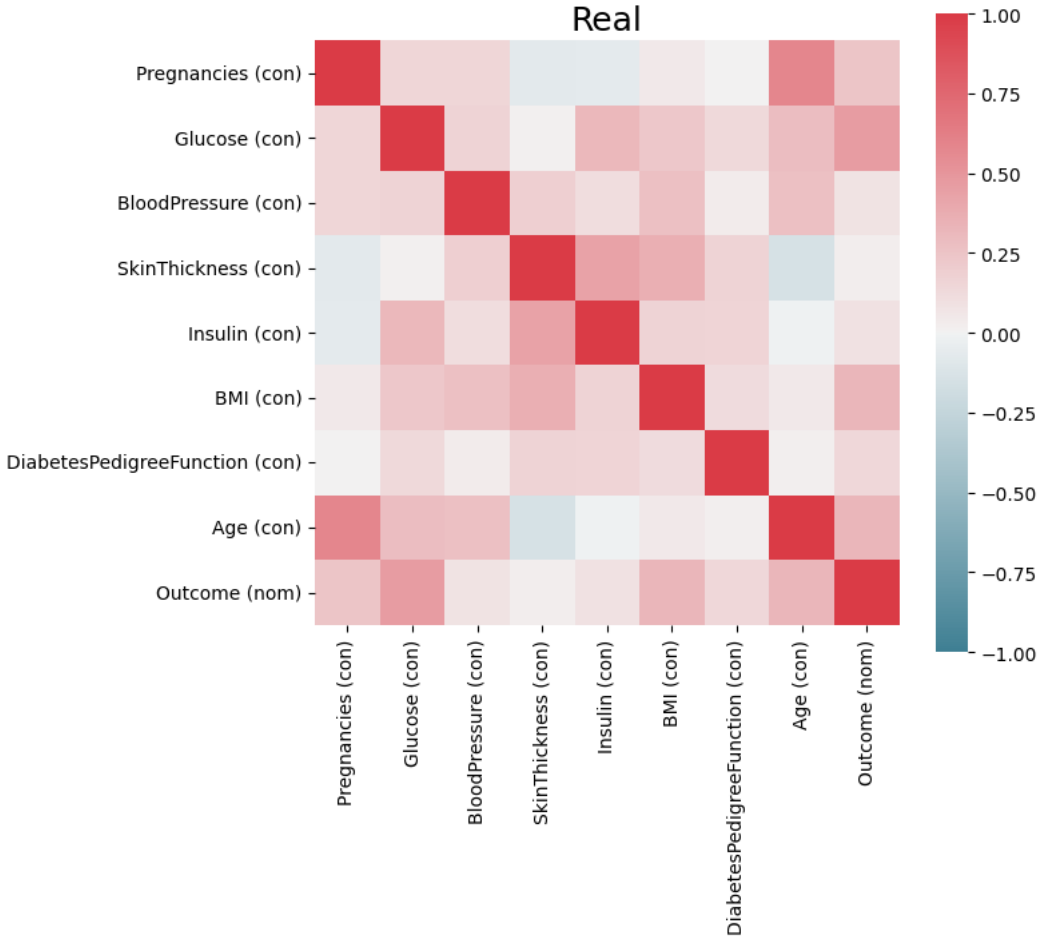
Results (Pima) – [2]

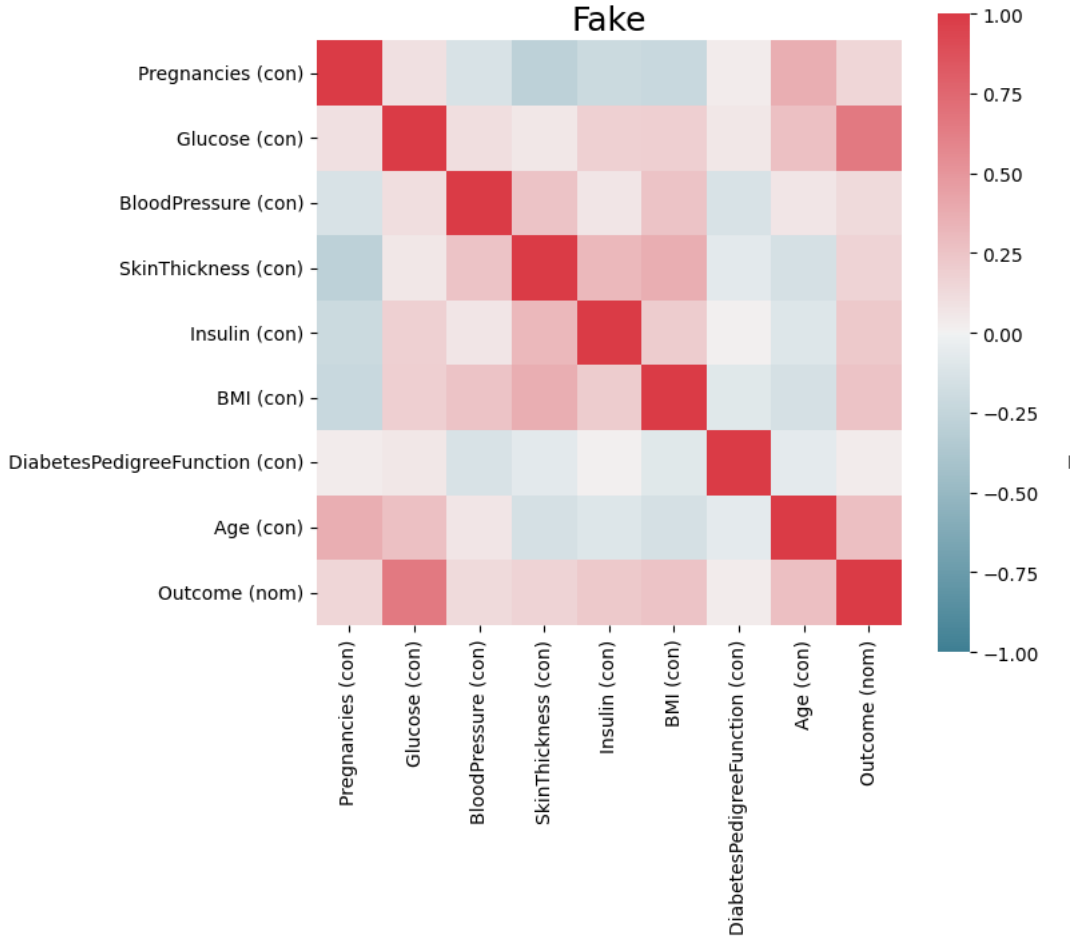


Results (Pima) – [3]

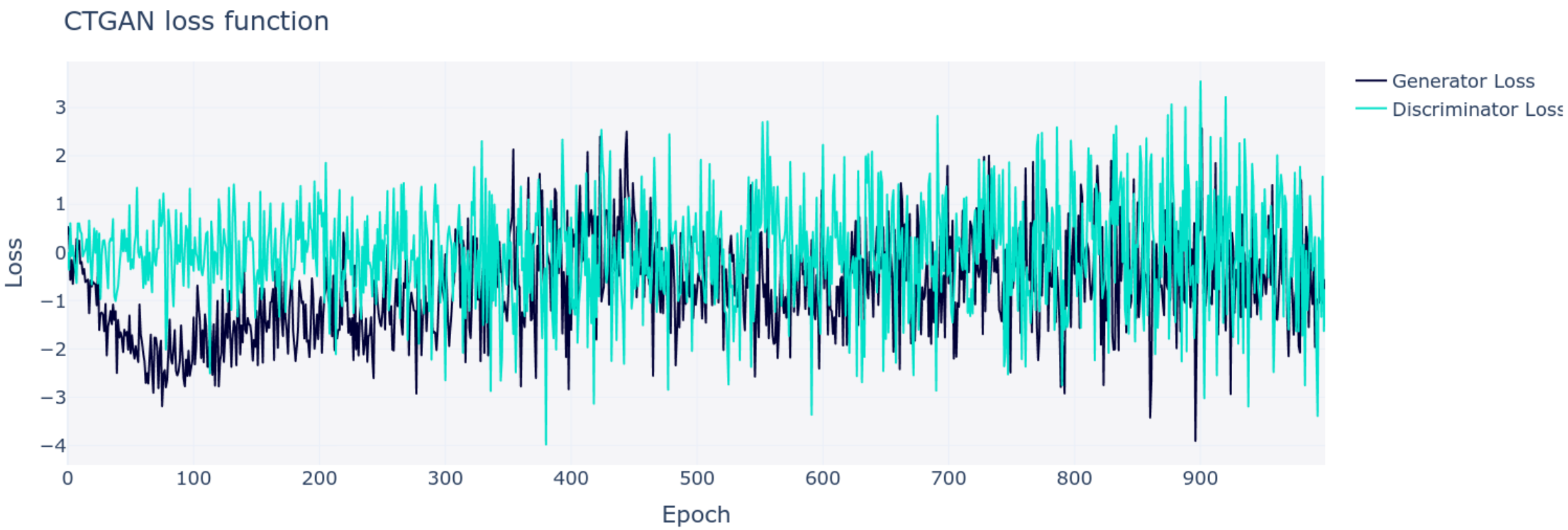


Results (Pima) – [4]

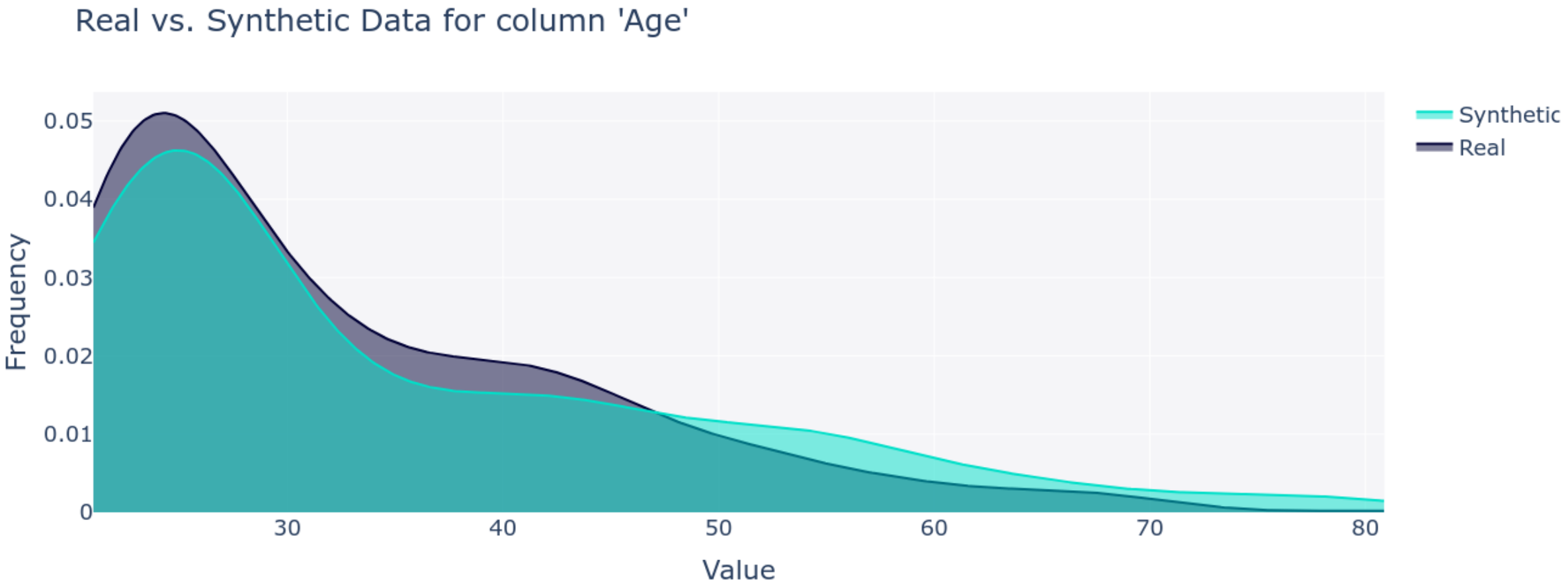




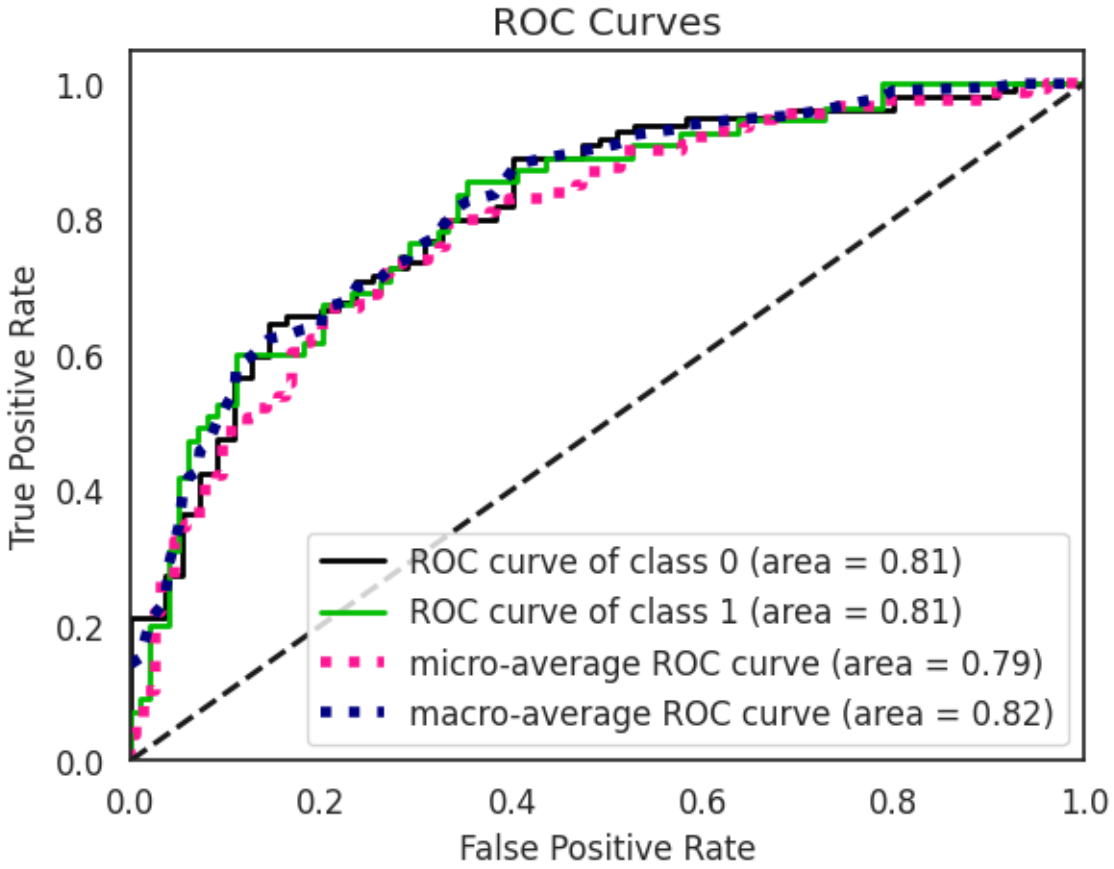
Results (Pima) – [5]

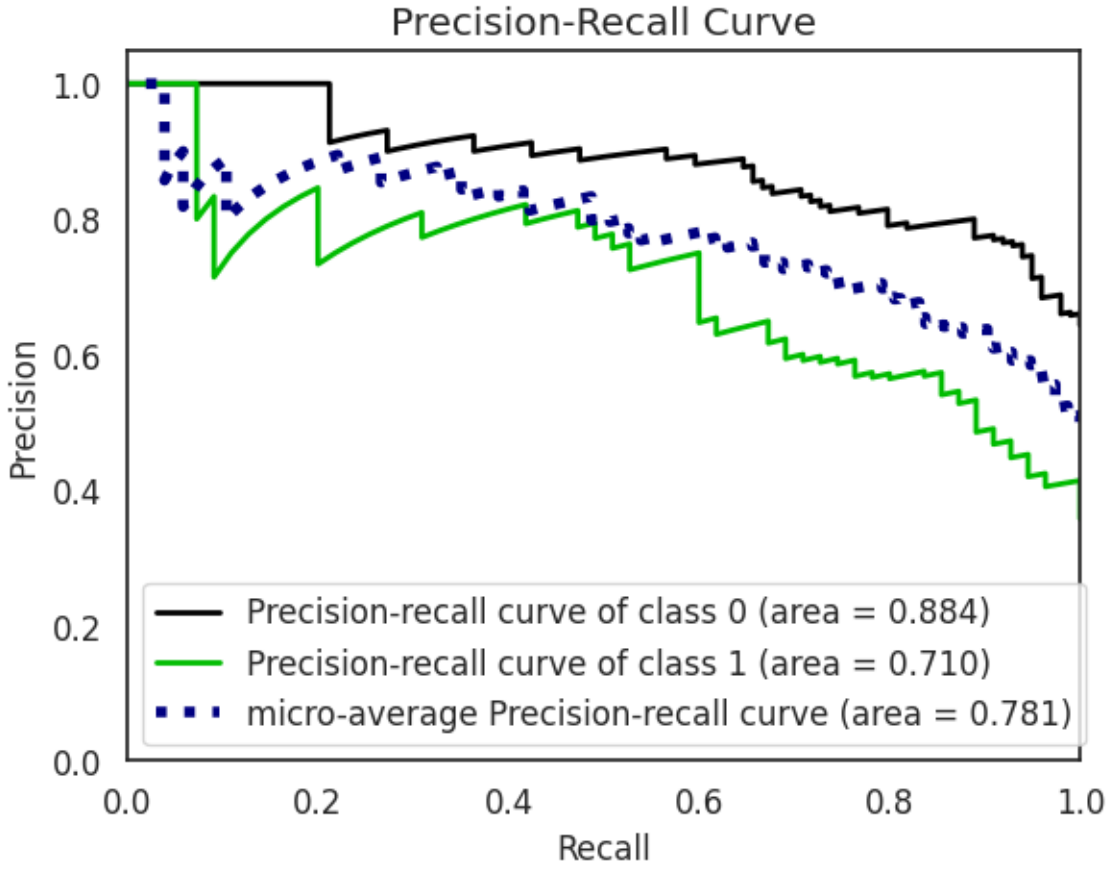


Results (Pima) – [6]

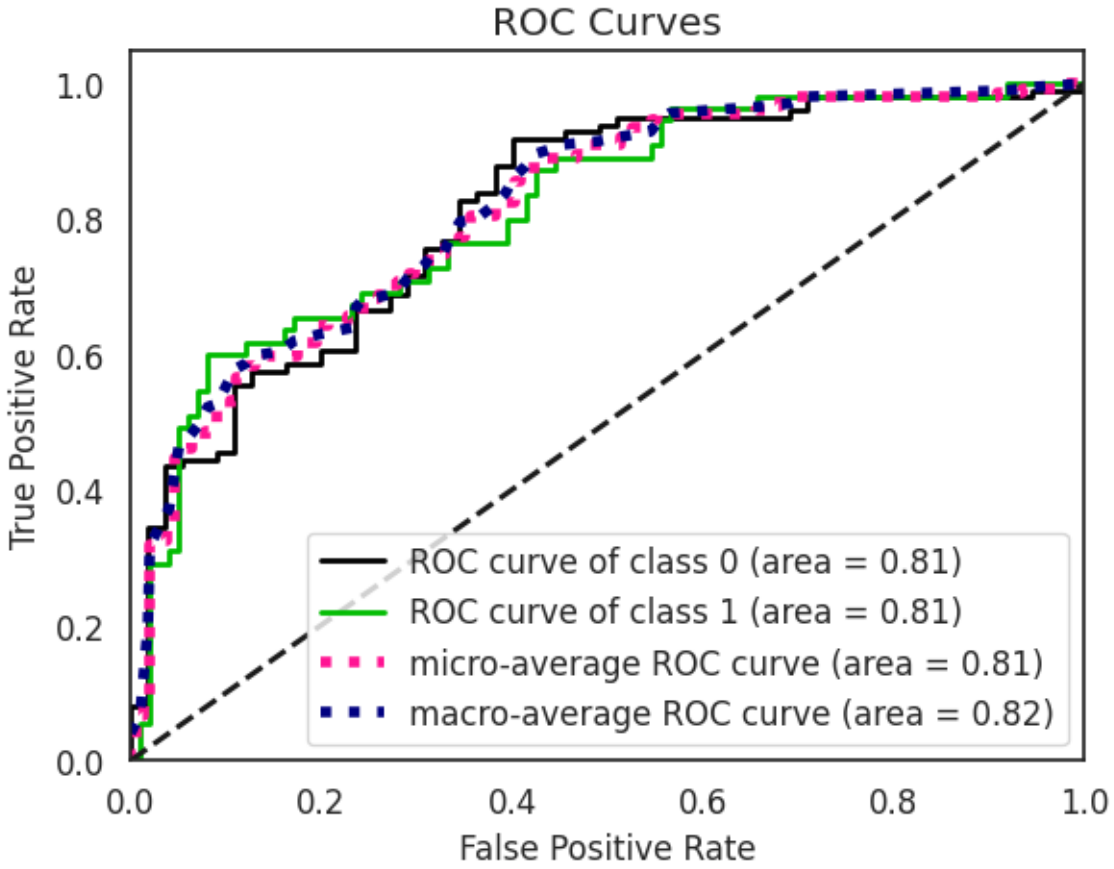


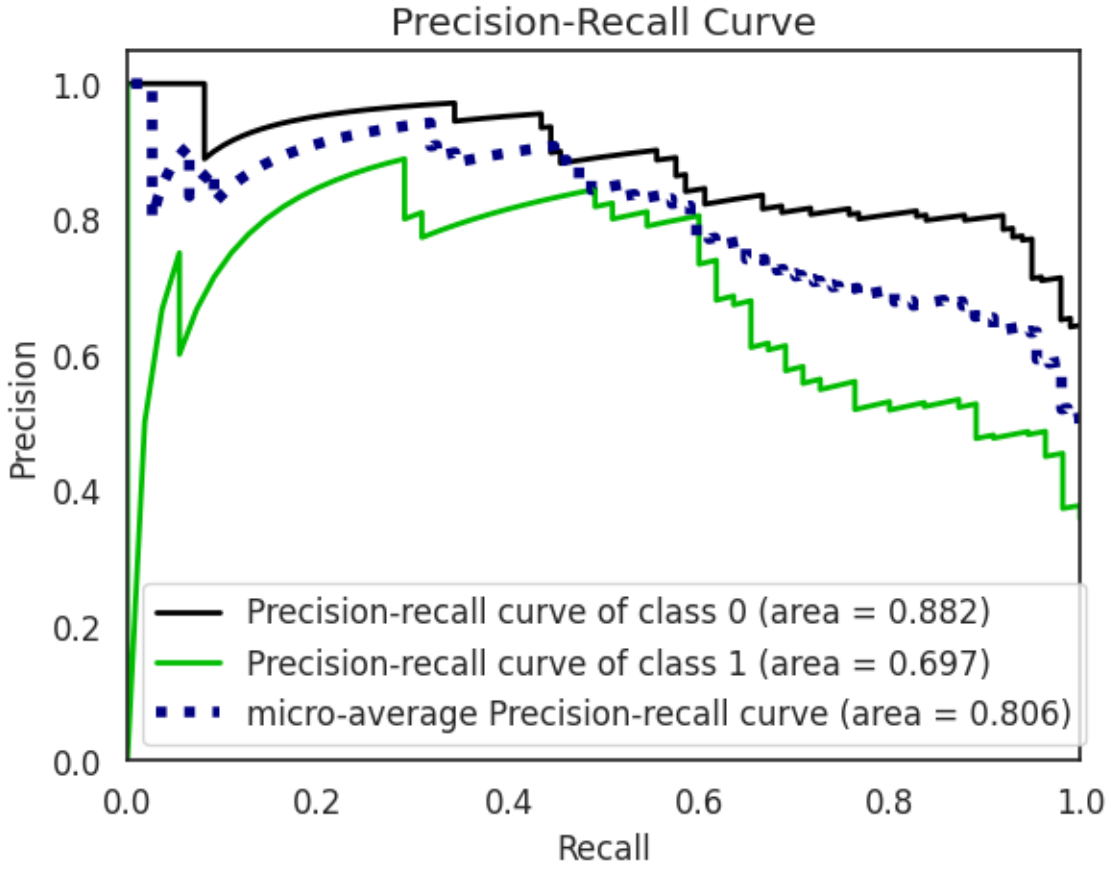
Results (Pima) – [7]
(ROC & PR Curve on Real Data)



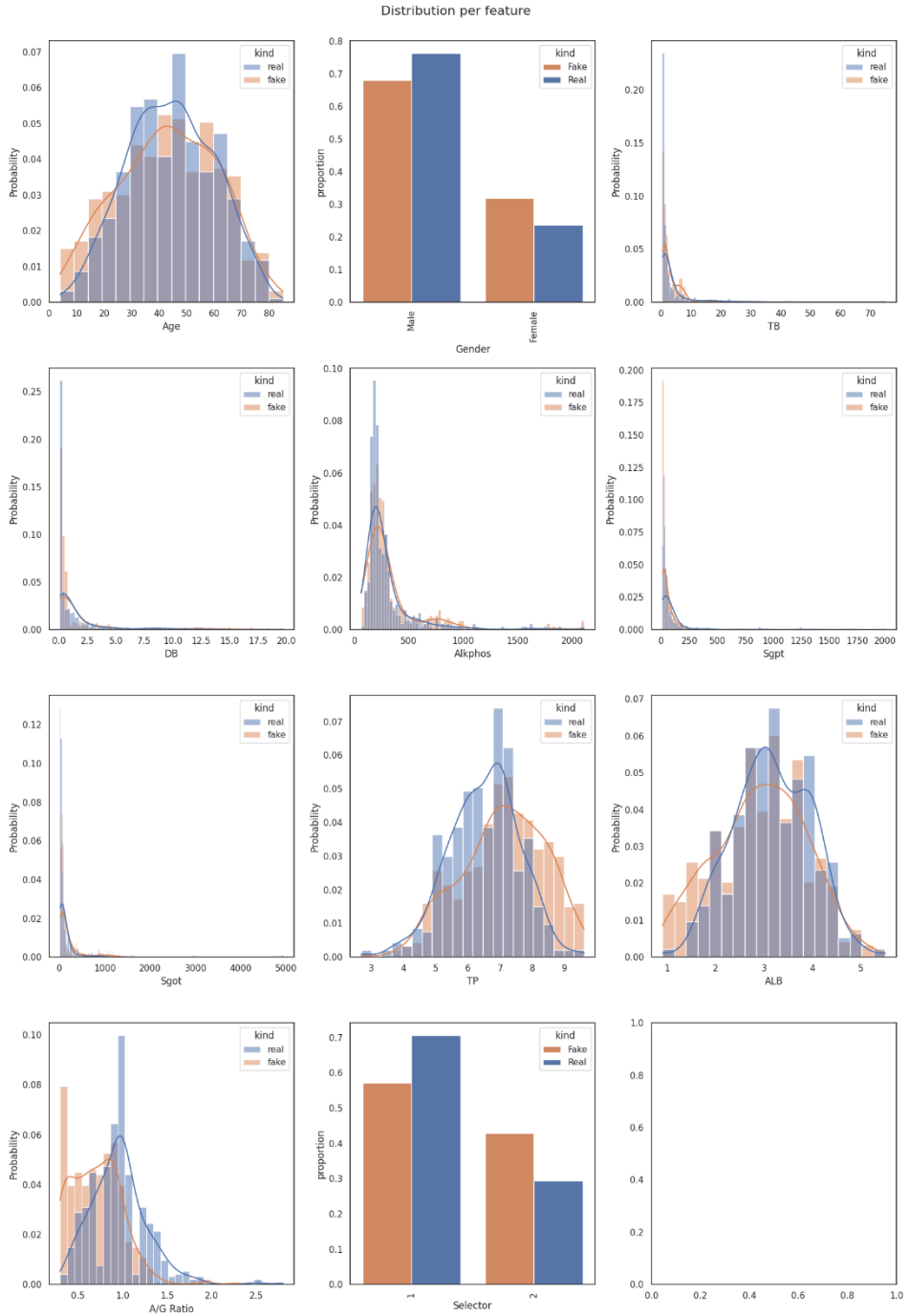


Results (Pima) – [8]
(ROC & PR Curve on Synthetic Data)

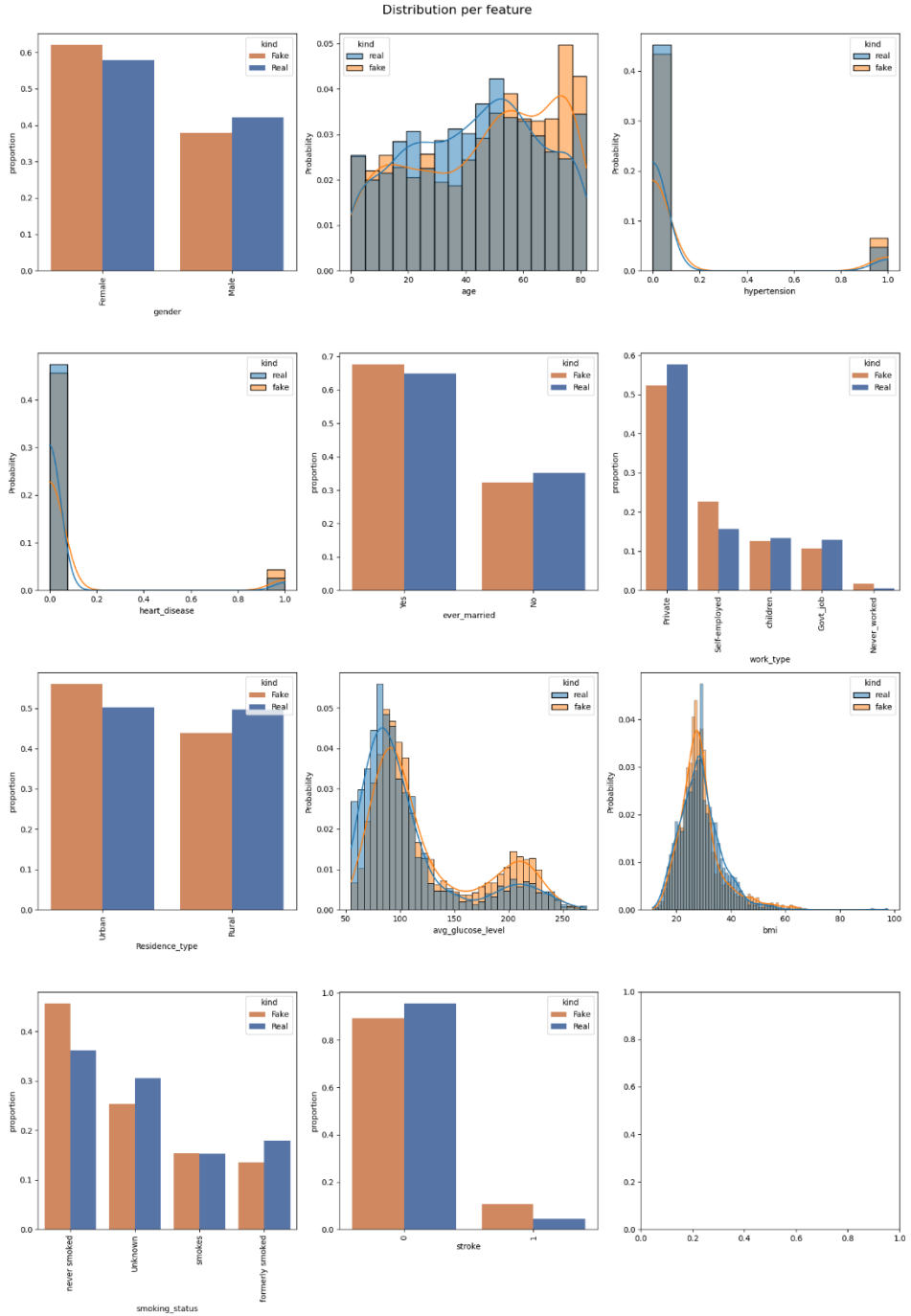




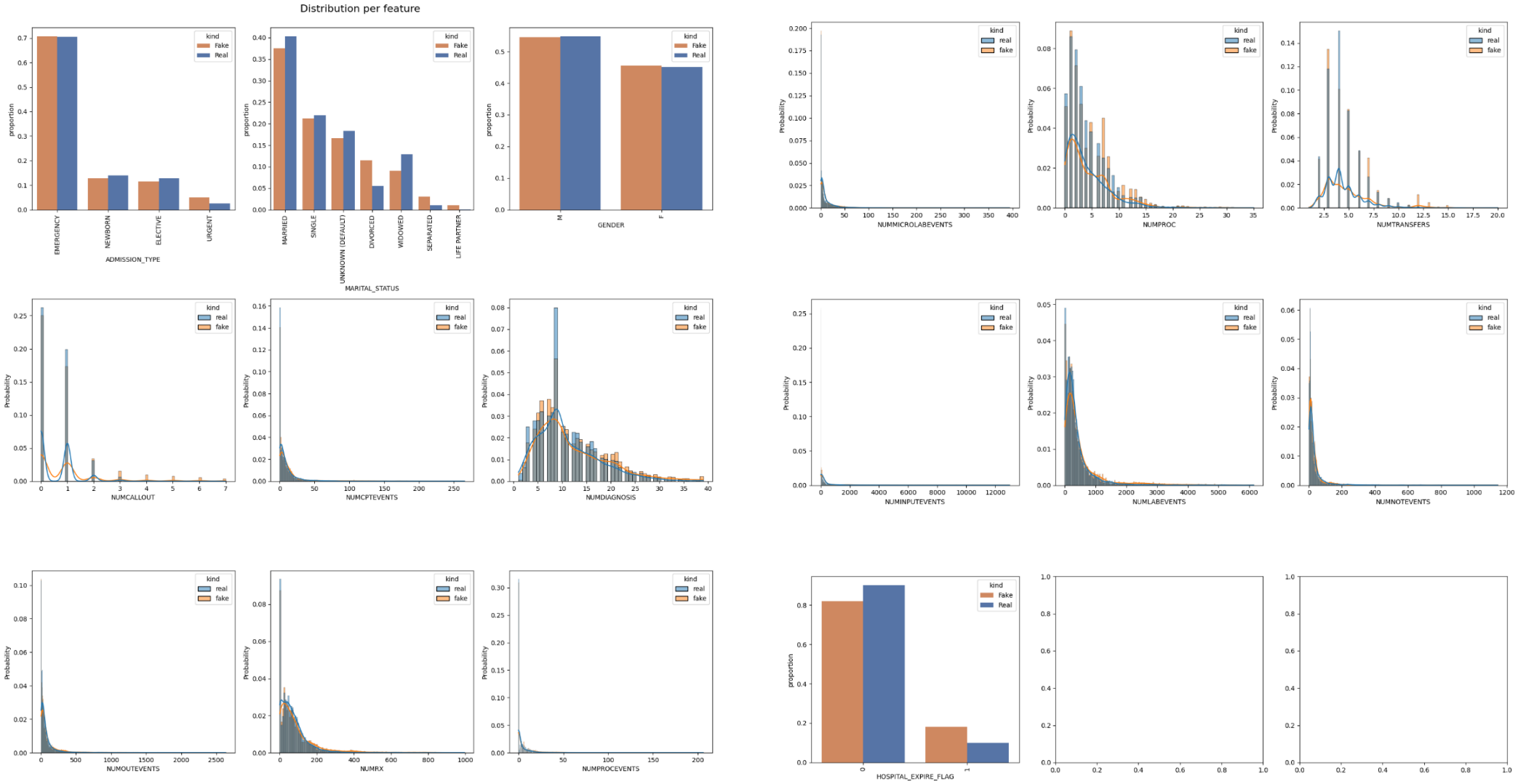
Results (ILPD) – [9]



Results (Stroke) – [10]



Results (MIMIC) – [11]



Discussion of Results (Pima) – [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 (ILPD) – [2]

Model	Dataset Type	Accuracy	F1-Score	ROC-AUC
Logistic Regression	Real	0.64	0.66	0.82
	Synthetic	0.58	0.6	0.82
XG Boost	Real	0.62	0.65	0.72
	Synthetic	0.68	0.7	0.8
Neural Network	Real	0.72	0.73	0.8
	Synthetic	0.73	0.68	0.73
Random Forest	Real	0.74	0.64	0.76
	Synthetic	0.68	0.7	0.82

Discussion of Results (Stroke) – [3]

Model	Dataset Type	Accuracy	F1 Score	ROC-AUC
Logistic Regression	Real	0.74	0.8	0.85
	Synthetic	0.79	0.84	0.82
XG Boost	Real	0.92	0.91	0.79
	Synthetic	0.87	0.89	0.74
Neural Network	Real	0.8	0.84	0.76
	Synthetic	0.85	0.87	0.72
Random Forest	Real	0.91	0.9	0.82
	Synthetic	0.92	0.91	0.79

Discussion of Results (MIMIC) – [4]

Model	Dataset Type	Accuracy	F1-Score	ROC-AUC
Logistic Regression	Real	0.76	0.8	0.82
	Synthetic	0.72	0.77	0.8
XG Boost	Real	0.58	0.66	0.81
	Synthetic	0.7	0.76	0.78
Neural Network	Real	0.83	0.86	0.88
	Synthetic	0.78	0.82	0.77
Random Forest	Real	0.92	0.91	0.87
	Synthetic	0.9	0.89	0.83

Discussion of Results – [5]

- Real datasets shows higher accuracy and F1-scores across all models and datasets
- Synthetic datasets performs comparably, with some variations

- Logistic Regression
  - Consistent performance between real and synthetic datasets
- XGBoost
  - But in case of Pima and ILPD shows greater accuracy and F1-scores for synthetic datasets

Discussion of Results – [6]

- Neural Network
  - In case of Pima and Stroke shows greater accuracy, F1-scores and ROC for synthetic datasets
- Random Forest
  - Exceptional performance on MIMIC and Stroke Datasets

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] E. Choi, S. Biswal and B. Malin, "Generating Multi-label Discrete Patient Records using Generative Adversarial Networks," in Proceedings of Machine Learning Research, 2017

References – [2]

[4] 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.

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

[6] A. Vaswani, N. Shazeer and N. Parmar, "Attention is All You Need," in Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, 2017.

References – [3]

[7] J. C. L. Borges, R. M. Lima and C. R. M. A. C. Silva, "Deep Unsupervised Learning using Nonequilibrium Thermodynamics," in Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 2018.

[8] J. Ho and X. Jiang, "Denoising Diffusion Probabilistic Models," in Proceedings of the 34th Conference on Neural Information Processing Systems , Vancouver, BC, Canada, 2020.

[9] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016.

References – [4]

[10] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[11] O. Ceritli, J. Doe and A. Smith, "Synthesizing Mixed-type Electronic Health Records using Diffusion Models," IEEE Transactions on Medical Informatics, vol. 22, no. 4, pp. 123-135, 2024.

[12] J. Solatorio, M. Green and P. Lee, "REaLTabFormer: Generating Realistic Relational and Tabular Data using Transformers," IEEE Transactions on Artificial Intelligence, vol. 10, no. 3, pp. 456-468, 2024.

References – [5]

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

[14] R. Bendi and V. N, "Indian Liver Patient Dataset," UCI Machine Learning Repository, 2012. [Online]. Available: https://doi.org/10.24432/C5D02C. [Accessed June 2024].

[15] F. Soriano, "Stroke Prediction Dataset," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset. [Accessed June 2024].

References – [6]

[16] A. E. Johnson, T. J. Pollard and L. Shen, "MIMIC-III, a freely accessible critical care database," MIT Laboratory for Computational Physiology, 2016. [Online]. Available: https://physionet.org/content/mimiciii/1.4/. [Accessed June 2024].