A Comparative Analysis on Synthetic Data Generation of Electronic Health Records using CTGAN, REaLTabFormer and TabDDPM

Arjan Sapkota [a], Girban Adhikari [b], Jivan Acharya [c], Subarna Ghimire [d], Umesh Kanta Ghimire [e]

[a] *Department of Electronics and Computer Engineering, Thapathali Campus, IOE, Tribhuvan University, Nepal*
[a] adgirban1@gmail.com

Abstract

The increasing importance of Electronic Health Records (EHR) for medical research and clinical applications necessitates the generation of high-quality synthetic data that preserves patient privacy. This study evaluates and compares the performance of Conditional Tabular Generative Adversarial Network (CTGAN), Transformers-based models (REaLTabFormer), and Diffusion Models (TabDDPM) across multiple medical datasets. Our findings demonstrate that TabDDPM consistently outperforms other models in generating synthetic data that closely mirrors real-world distributions, effectively preserving statistical properties and feature relationships. Its ability to maintain complex dependencies and capture variations in the data makes it the most reliable choice for synthetic EHR generation. While CTGAN proves to be a strong alternative, particularly excelling in certain datasets, its performance is less stable across different distributions, leading to occasional deviations from real data characteristics. REaLTabFormer, on the other hand, shows potential in specific cases but struggles to maintain statistical integrity and generalization across diverse datasets, limiting its effectiveness in some scenarios.

Keywords

CTGAN, Diffusion Models, GAN, Synthetic Data Generation, Transformers

1. Introduction

Electronic Health Records (EHR) are the backbone of modern healthcare, helping doctors make informed decisions, supporting medical research, and improving patient management. However, strict privacy regulations and security concerns often limit access to this valuable data, making it challenging to use for AI-driven applications. Synthetic data offers a promising solution, it mimics real patient records while protecting privacy, allowing researchers and developers to work with realistic datasets without ethical or legal risks.

Generating high-quality synthetic EHR data isn't easy, though. Medical data is complex, and traditional methods often fall short in capturing its intricate patterns. That's why we explore multiple advanced models, each chosen for a specific reason. CTGAN [1] builds on the progress of Variational Autoencoders (VAE) [2], Generative Adversarial Networks (GANs) [3], and Wasserstein GANs (WGANs)[4], making it one of the best options for generating medical data. ReaLTabFormer [5] leverages the power of Transformers, which is hugely popular in AI research thanks to the "Attention Is All You Need" [6] paper, to better capture relationships in tabular data. Finally, we experiment with Diffusion Models (TabDDPM) [7], which are typically used for image generation, to see how well they perform in creating synthetic EHR data.

The potential applications of synthetic EHR data are vast. It helps improve AI models by providing diverse, scalable datasets for training and validation. It also enables privacy-preserving research, allowing institutions to comply with regulations while still benefiting from large-scale data analysis. Beyond that, synthetic data are a game changer for education and training, giving students and professionals access to realistic datasets without ethical concerns. It even helps test AI models in a safe environment before they're deployed in real-world healthcare settings, ensuring reliability.

2. Related Works

Recent advances in generative modeling have introduced a wide array of techniques for synthesizing complex data, particularly in the field of electronic health records (EHRs). Kingma et al. [2] introduced Auto-Encoding Variational Bayes (AEVB), which led to the development of Variational Autoencoders (VAEs). This framework leverages the Stochastic Gradient Variational Bayes (SGVB) estimator, a differentiable and unbiased method that utilizes ancestral sampling to efficiently optimize recognition models and learn latent representations. Although VAEs provide an efficient means of handling intractable posteriors, they can struggle when confronted with highly complex or multimodal data distributions.

Generative Adversarial Networks (GANs), introduced by Goodfellow et al. [3], employ a competitive process between a generator and a discriminator to produce realistic data samples. Despite their ability to generate visually compelling results, GANs often encounter training instabilities and difficulties in modeling discrete data. Extensions such as Conditional GANs (CGANs) and CTGAN, developed by Xu et al. [1], have been specifically designed to synthesize tabular data by conditioning the generation process to preserve statistical properties. However, these adaptations still inherit some of the inherent challenges of adversarial training.

To address issues of training stability, Arjovsky et al. [4] proposed Wasserstein GANs (WGANs), which replace the traditional Jensen-Shannon divergence with the Wasserstein distance. This modification results in a more stable

convergence and a reduction in mode collapse, though it also brings increased computational costs and the necessity for meticulous tuning of the critic network. Meanwhile, the advent of Transformer-based models, as pioneered by Vaswani et al. [6], has revolutionized the way long-range dependencies are modeled in data. Transformers use self-attention mechanisms that enable parallelizable training, leading to architectures like REaLTabFormer [5], which adapts this approach specifically for relational tabular data by capturing complex inter-feature relationships in EHRs. Despite their powerful capabilities, Transformer-based models are often resource-intensive and require substantial computational power.

In parallel, diffusion models have emerged as a promising alternative for high-fidelity data generation. Denoising Diffusion Probabilistic Models (DDPMs), introduced by Ho et al. [8], use an iterative denoising process inspired by thermodynamics to progressively transform noise into realistic data. This approach, further refined in models such as TabDDPM [7] for tabular data synthesis, delivers robust and high-quality synthetic samples, albeit with slower generation speeds and higher computational demands. Recent work by Ceritli et al. [9] highlights the potential of diffusion-based methods in EHR synthesis, where they have shown to outperform GAN- and VAE-based approaches in terms of both fidelity and privacy preservation.

Additionally, ensemble and tree-based methods like XGBoost, developed by Chen and Guestrin [10], and Random Forests, introduced by Breiman [11], though not directly generative, are recognized for their robustness in predictive tasks. These methods often serve as benchmarks or components within hybrid models that evaluate the quality of synthetic data. In the specific context of healthcare, medGAN proposed by Choi et al. [12] leverages the principles of GANs to generate synthetic EHRs that maintain privacy, demonstrating the practical application and inherent challenges of applying these advanced models in sensitive domains.

In summary, the landscape of synthetic EHR data generation is characterized by diverse approaches, each offering unique strengths and facing distinct challenges. Variational autoencoders and adversarial methods such as GANs provide efficient latent space representations and realistic sample generation, yet they are often hampered by issues like training instability. In contrast, WGANs, Transformer-based architectures, and diffusion models offer improved stability and fidelity at the expense of increased computational requirements. By carefully comparing these methods on dimensions such as training stability, fidelity, and computational efficiency, researchers can better select and tailor generative approaches to meet the stringent demands of privacy-preserving synthetic data generation in healthcare.

3. System Architecture and Methodology

The system architecture shown in Figure 1 begins with an Original Dataset containing real-world EHR data, which is divided into Train and Test Datasets. The Train Dataset is passed through a Data Synthesizer that leverages CTGANs, ReaLTabFormer, and TabDDPM models to generate synthetic data with patterns similar to the original. This synthetic data is then used to train machine learning models including Logistic Regression, XGBoost, Random Forest, and Multi-Layer Perceptron (MLP).

These models are also trained separately on the Original Data for comparison. Both sets of trained models those using real data and those using synthetic data are evaluated based on key performance metrics such as Accuracy, F1-Score, and Area Under the Curve (AUC). This side-by-side comparison helps assess how well the synthetic data reflects the real-world patterns and whether it can effectively be used to train predictive models for healthcare applications.

This project utilizes diverse datasets to enhance synthetic data generation across different domains. The Pima Indian Diabetes Dataset [13] contains variables related to diabetes risk, such as glucose concentration, BMI, and insulin levels, aiding in predictive model development. The Indian Liver Patient Dataset (ILPD) [14] provides key attributes for liver disease risk assessment, including bilirubin levels, enzyme concentrations, and albumin ratios. The Stroke Prediction Dataset [15] includes demographic and clinical factors like age, hypertension, heart disease, and smoking status to support stroke risk modeling. Additionally, MIMIC-III (Medical Information Mart for Intensive Care III) [16] is a comprehensive clinical database featuring patient demographics, vital signs, laboratory test results, medications, and ICU stay details, facilitating research in critical care medicine. By working with these diverse datasets, the project ensures that the synthetic data reflects a wide range of real-world clinical scenarios.

Table 1: Dataset Summary

Dataset	Features	Rows
Indian Liver Patient	10	583
Pima Indian Diabetes	8	768
Stroke Prediction	11	5110
MIMIC-III (Mortality)	19	58976

3.1 Data Preprocessing

Data preprocessing is a critical step in preparing the dataset for machine learning models. This section outlines the detailed preprocessing steps undertaken to ensure the quality and usability of the EHR data.

3.1.1 MIMIC-III

In this dataset, several key tables from the MIMIC-III dataset were used to extract relevant information for modeling and analysis. Some of these tables are:

- **ADMISSIONS:** Contains information about patient admissions, including `SUBJECT_ID`, `HADM_ID`, admission type, marital status, ethnicity, and a flag indicating whether the patient expired in the hospital.
- **PATIENTS:** Includes demographic details of the patients, such as `SUBJECT_ID` and gender.
- **CALLOUT:** Aggregates the number of callout events by

SUBJECT_ID and HADM_ID, resulting in a count column NUMCALLOUT.

- **CPTEVENTS:** Aggregates the number of CPT (Current Procedural Terminology) events by SUBJECT_ID and HADM_ID, resulting in a count column NUMCPTEVENTS.

Rolling Up Data To create a comprehensive dataset suitable for machine learning models, data from these tables were aggregated and rolled up into a single, unified table. The rolling-up process involved merging various sources of patient information to ensure all relevant features were present in a structured manner.

Merging Tables The first step was to merge tables using common keys, such as SUBJECT_ID (patient identifier) and HADM_ID (hospital admission identifier). This allowed the integration of patient demographics, admission details, ICU stay information, and other clinical data. By joining multiple tables, a holistic view of patient records was created, capturing essential features for modeling.

Handling Multiple ICU Stays Some patients had multiple ICU stays during their hospital admission, posing a challenge for data processing. To handle these cases, data were aggregated to create a single record per admission. The earliest and latest timestamps of ICU stays were used to calculate the total length of stay, ensuring that each patient admission had a consolidated representation.

3.1.2 Other Datasets

Feature Aggregation Clinical measurements and lab results were aggregated using summary statistics (mean, median, min, max) for each ICU stay, condensing high-frequency data into meaningful summaries.

Temporal Alignment Time-series data were synchronized based on timestamps to maintain the sequence of events and ensure consistency across different tables.

Creating New Features New features were engineered, such as binary indicators for specific lab results or vital signs, enhancing the predictive power of machine learning models.

Data Loading and Inspection The dataset was first loaded and inspected for missing values, variable distributions, and anomalies to inform preprocessing steps.

Handling Missing Values Missing data were imputed using statistical measures (mean, median, mode) or removed if missingness was excessive to maintain dataset integrity.

Removing Invalid Data Records with unrealistic values (e.g., negative mortality indicators) were identified and excluded to prevent biases.

Feature Engineering Categorical variables were encoded numerically, and additional features were created to capture meaningful patterns, such as mortality flags.

Normalization and Standardization Data were scaled using normalization ([0,1] range) or standardization (zero mean, unit variance) to ensure fair feature contributions in model training.

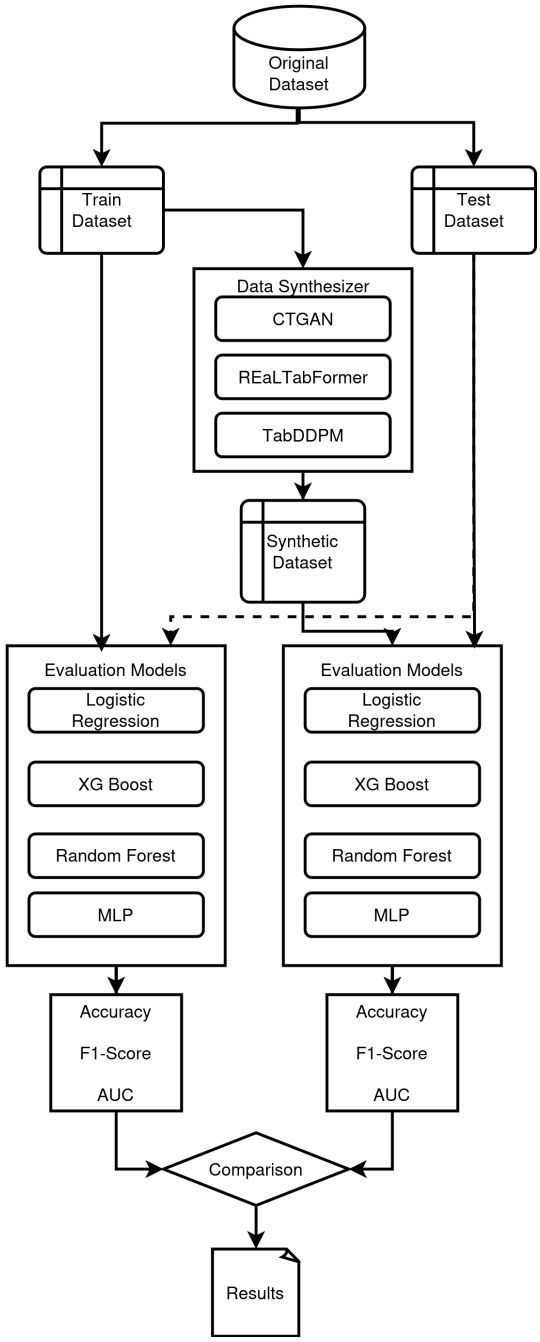


Figure 1: System Architecture and Workflow

ReaLTabFormer utilizes gradient accumulation steps of 4. Early stopping, as implemented in the ReaLTabFormer paper, allows training to stop between 20-60 epochs depending on the dataset. Other hyperparameters are mentioned in Table 2.

Table 2: Model Training Parameters

Model	Epochs	Batch Size	Learning Rate
CTGAN	1000	32, 256	2.10^{-4}
ReaLTabFormer	20-60	8	-
TabDDPM	1000	32, 256, 512	0.0001

4. Results and Discussion

After the training of the synthetic data generation models on all the datasets, synthetic data was sampled with equal number of rows as in the original trainset size. The Table 3 presents the performance of synthetic data generated from all three data generation models across four datasets using multiple evaluation metrics. TabDDPM consistently achieves higher F1-scores and ROC-AUC values, indicating better synthetic data quality compared to CTGAN and REaLTabFormer. In the Mortality and Stroke datasets, Random Forest and Neural Networks trained on TabDDPM synthetic data show strong predictive performance, closely matching real data results. CTGAN performs well in some cases but struggles with stability across datasets. REaLTabFormer shows competitive results but lags in key metrics for certain models. Overall, TabDDPM proves to be a more reliable approach for generating high-utility synthetic medical data.

4.1 Mortality Dataset

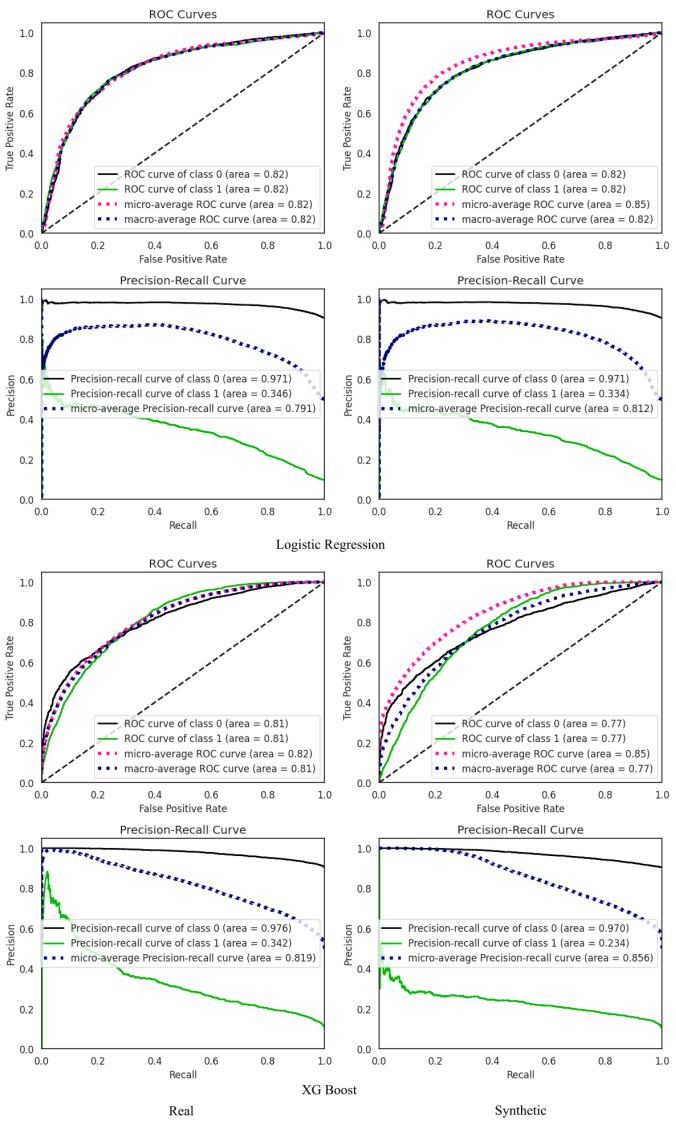


Figure 2: ROC and PR curves on Mortality for Logistic Regression and XGBoost (Using REaLTabFormer)

The ROC and Precision-Recall (PR) curves compare the classification performance of real (left) and synthetic (right) data generated by REaLTabFormer using Logistic Regression and XGBoost. The ROC curves indicate that the synthetic data closely follows the real data distribution, with AUC scores remaining consistent across both evaluation models. However, slight variations in the decision boundaries can be observed, particularly in XGBoost, which shows a lower AUC for the synthetic data.

The PR curves reveal class-wise predictive performance differences. For the majority class (class 0), both real and synthetic data achieve high precision, but for the minority class (class 1), the synthetic data exhibits a drop in recall and precision, more pronounced in XGBoost. Logistic Regression maintains a more stable performance across both real and synthetic datasets, suggesting that REaLTabFormer preserves general patterns well but may introduce minor shifts in synthetic minority class representations, as seen in Figure 2.

4.2 Stroke Dataset

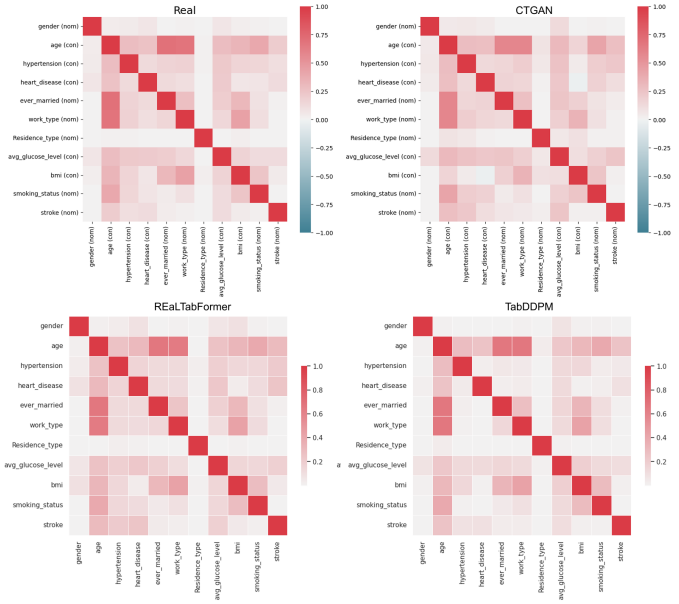


Figure 3: Heatmaps on Stroke Dataset

Figure 3 presents a comparative heatmap analysis of real and synthetic stroke datasets, illustrating the correlation between key features such as age, hypertension, heart disease, and smoking status. The heatmaps include the real dataset (top-left) alongside those generated by CTGAN, REaLTabFormer, and TabDDPM.

From the comparison, it is evident that for this dataset REaLTabFormer captures the underlying correlation patterns more effectively than CTGAN and TabDDPM. While CTGAN generates data with some degree of alignment to the real dataset, it introduces some distortions. REaLTabFormer demonstrates a stronger resemblance to real-world patterns, whereas in this case TabDDPM has struggled in mapping correlation between features such as work_type, ever_married, heart_disease and Residence_type. It shows that in some cases REaLTabFormer has maintained structured data consistency.

Table 3: Performance comparison of different models on synthetic and real EHR datasets

Data Generation Models	Evaluation Models	PIMA			ILPD			MORTALITY			STROKE		
		Accuracy	F1	ROC-AUC	Accuracy	F1	ROC-AUC	Accuracy	F1	ROC-AUC	Accuracy	F1	ROC-AUC
CTGAN	Logistic Regression	0.71 / 0.73	0.71 / **0.73**	0.81 / 0.81	0.64 / 0.58	0.66 / 0.60	0.82 / 0.82	0.76 / 0.72	0.80 / 0.77	0.82 / 0.80	0.74 / **0.79**	0.80 / **0.84**	0.85 / 0.82
	XG Boost	0.75 / 0.72	0.75 / 0.73	0.79 / **0.82**	0.62 / **0.68**	0.65 / **0.70**	0.72 / **0.80**	0.58 / **0.70**	0.66 / **0.76**	0.81 / 0.78	0.92 / **0.87**	0.91 / 0.89	0.79 / 0.74
	Neural Network	0.73 / **0.74**	0.73 / 0.72	0.77 / 0.80	0.72 / **0.73**	0.73 / 0.68	0.80 / 0.73	0.83 / 0.78	0.86 / 0.82	0.88 / 0.77	0.80 / **0.85**	0.84 / 0.87	0.76 / 0.72
	Random Forest	0.77 / 0.73	0.77 / 0.73	0.83 / 0.81	0.74 / 0.68	0.64 / **0.70**	0.76 / **0.82**	0.92 / 0.90	0.91 / 0.89	0.87 / 0.83	0.91 / **0.92**	0.90 / 0.91	0.82 / 0.79
REaLTabFormer	Logistic Regression	0.70 / **0.71**	0.71 / **0.72**	0.81 / **0.82**	0.64 / **0.66**	0.66 / **0.68**	0.82 / 0.79	0.76 / **0.79**	0.80 / **0.83**	0.82 / 0.82	0.74 / **0.74**	0.80 / **0.81**	0.85 / 0.84
	XG Boost	0.69 / 0.64	0.69 / 0.65	0.77 / **0.78**	0.62 / **0.78**	0.65 / **0.79**	0.72 / **0.82**	0.73 / **0.75**	0.78 / **0.80**	0.81 / 0.77	0.91 / **0.89**	0.91 / 0.90	0.80 / 0.80
	Neural Network	0.76 / 0.63	0.74 / 0.63	0.78 / 0.79	0.74 / 0.74	0.74 / 0.63	0.79 / 0.64	0.86 / 0.85	0.88 / 0.87	0.87 / 0.84	0.83 / **0.84**	0.86 / 0.87	0.76 / 0.77
	Random Forest	0.77 / 0.70	0.77 / 0.71	0.82 / 0.79	0.71 / **0.74**	0.72 / 0.72	0.77 / **0.81**	0.92 / 0.92	0.91 / 0.90	0.90 / 0.88	0.93 / **0.90**	0.92 / 0.90	0.82 / 0.82
TabDDPM	Logistic Regression	0.75 / **0.76**	0.75 / **0.76**	0.86 / **0.86**	0.62 / 0.62	0.63 / 0.63	0.72 / **0.75**	0.76 / **0.77**	0.80 / **0.81**	0.82 / 0.82	0.75 / **0.77**	0.82 / **0.84**	0.79 / **0.80**
	XG Boost	0.73 / **0.76**	0.74 / **0.76**	0.84 / **0.83**	0.65 / **0.68**	0.66 / **0.69**	0.68 / **0.70**	0.67 / 0.66	0.75 / 0.73	0.84 / 0.78	0.89 / **0.91**	0.91 / 0.92	0.78 / **0.79**
	Neural Network	0.69 / 0.73	0.69 / 0.70	0.79 / 0.80	0.65 / 0.66	0.60 / **0.67**	0.67 / 0.66	0.83 / 0.78	0.86 / 0.83	0.87 / 0.85	0.88 / **0.87**	0.90 / 0.89	0.68 / **0.71**
	Random Forest	0.76 / **0.75**	0.74 / 0.74	0.81 / **0.83**	0.67 / 0.69	0.55 / **0.67**	0.72 / **0.73**	0.90 / 0.90	0.86 / 0.87	0.76 / **0.81**	0.96 / **0.95**	0.93 / 0.93	0.71 / **0.86**

4.3 ILPD Dataset

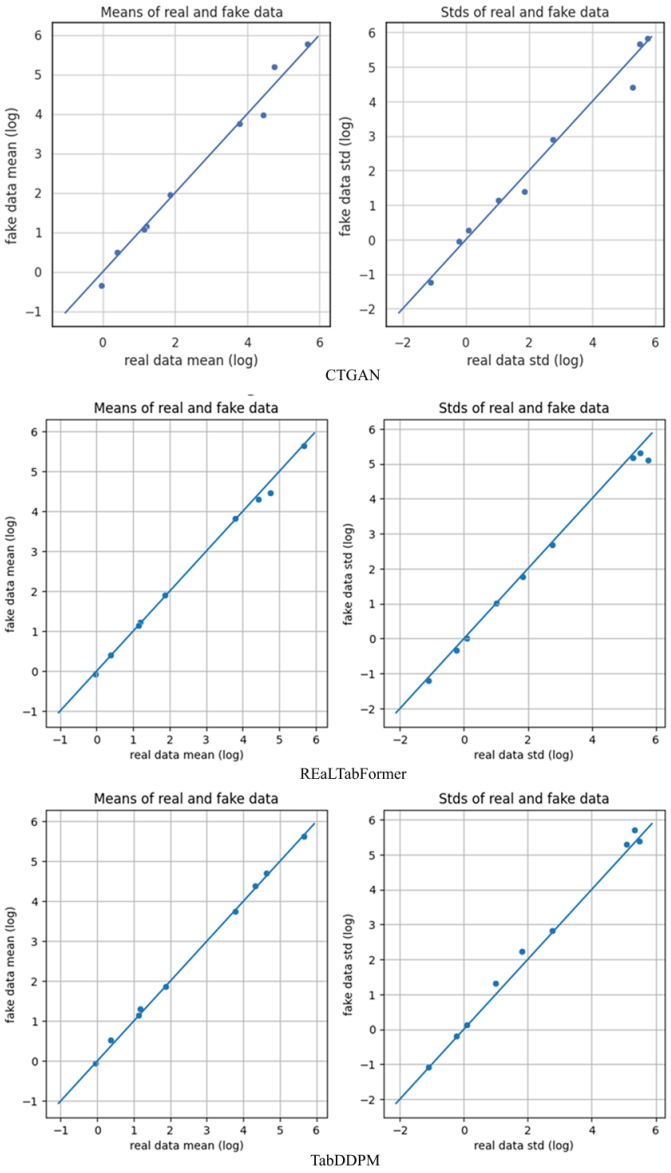


Figure 4: Logmeans of ILPD Dataset

Figure 4 evaluates the statistical consistency of real and synthetic ILPD dataset by plotting their means and standard deviations on a log scale. The identity line (y = x) serves as an ideal that the synthetic data faithfully reproduces the statistical properties of the real dataset. Among the three models, TabDDPM shows the closest alignment with the identity line, indicating its high accuracy in preserving the statistical structure of the real data. REaLTabFormer also demonstrates strong alignment, though with minor deviations. In contrast, CTGAN exhibits noticeable inconsistencies, particularly in replicating standard deviations of some of the features as shown by the deviations, suggesting that it struggles to capture the full variability of the ILPD dataset.

4.4 PIMA Dataset

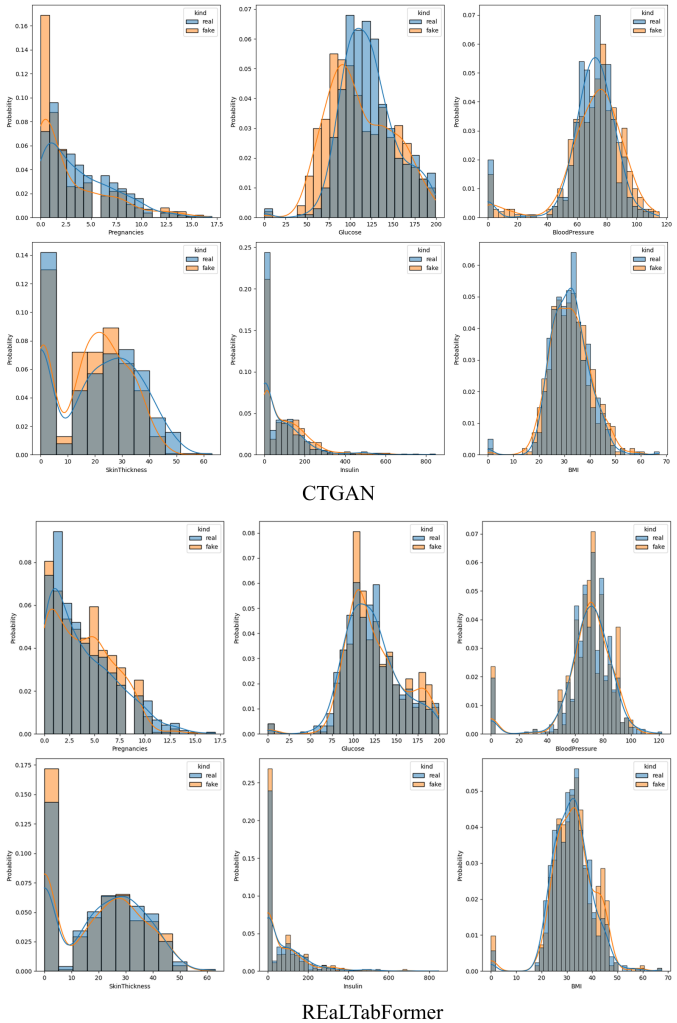
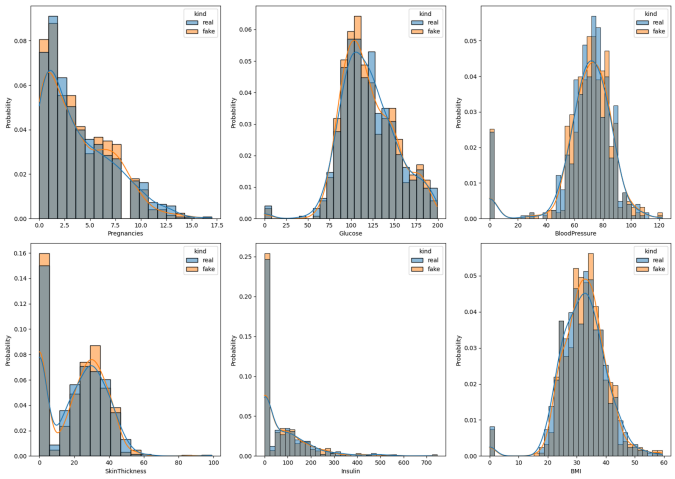


Figure 5: Distribution per feature of PIMA Dataset for CTGAN and REaLTabformer



TabDDPM

Figure 6: Distribution per feature of PIMA Dataset for TabDDPM

Figure 5 displays the distribution of six key features—Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, and BMI—across real and synthetic datasets for the Pima Indian Diabetes dataset. This figure is instrumental in evaluating how closely synthetic data follows real-world distributions, a critical factor in determining a model's effectiveness for generating high-quality synthetic medical data.

CTGAN shows reasonable alignment with real distributions but struggles with features like Glucose and Pregnancies, where deviations are more pronounced. REaLTabFormer achieves better distribution matching, particularly in capturing the spread of Insulin and BMI values, but it has also shown slight deviation in replicating the Pregnancies feature. However, TabDDPM as seen in 6 consistently outperforms the other models, closely approximating the real data distributions across most features.

5. Conclusion

CTGAN performs well across datasets, particularly in the Mortality and Stroke datasets, where it achieves high scores in accuracy, F1, and ROC-AUC. However, its performance varies in the ILPD dataset, showing a drop in ROC-AUC compared to other datasets. The model maintains consistent results across evaluation models but shows slight variations when compared to other data generation models.

REaLTabFormer struggles in the ILPD dataset, yielding the lowest accuracy and ROC-AUC among all models. It performs moderately in the PIMA dataset but shows competitive results in the Mortality and Stroke datasets, where it achieves high accuracy and F1 scores. TabDDPM demonstrates overall robustness, consistently achieving strong accuracy and F1 scores, especially in the PIMA and Stroke datasets. It surpasses other models in key metrics, making it the most reliable choice for synthetic data generation.

Future Enhancements

While the core objectives of this project have been successfully achieved, several additional analyses remain to ensure a comprehensive evaluation of synthetic data generation techniques. These future enhancements aim to refine our understanding of TabDDPM's performance and its comparison to CTGAN and REaLTabFormer across diverse EHR datasets.

Domain-Specific Feature Validation Assess whether synthetic data preserves critical domain-specific patterns, such as temporal trends and rare disease occurrences, to ensure its applicability in real-world healthcare settings.

Enhancing Data Synthesis Techniques Explore hybrid models that combine generative approaches (e.g., diffusion models with adversarial training) to improve fidelity and diversity. For example, integrating Variational Autoencoders (VAEs) with TabDDPM could enhance feature representation.

Alternative Evaluation Metrics Incorporate domain-expert validation and statistical measures such as Kernel Density Estimation (KDE) for distribution comparison and the Kolmogorov-Smirnov (KS) test to assess differences between real and synthetic data distributions.

Acknowledgments

The authors are grateful to the Department of Electronics and Computer Engineering, Thapathali Campus, for providing the necessary financial support and resources for this project. The authors appreciate the guidance and encouragement of the faculty members, whose valuable insights have greatly contributed to this research.

References

[1] Leyi Xu, Mijung Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[2] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2013.

[3] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014.

[4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, Sydney, Australia, 2017.

[5] Aivin V Solatorio and Olivier Dupriez. Realtabformer: Generating realistic relational and tabular data using transformers. *IEEE Transactions on Artificial Intelligence*, 10(3):456–468, 2024.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In

Proceedings of the 31st International Conference on Neural Information Processing Systems, pages 6000–6010, Long Beach, California, 2017.

[7] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. *arXiv preprint arXiv:2209.15421*, 2022.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th Conference on Neural Information Processing Systems*, Vancouver, BC, Canada, 2020.

[9] Taha Ceritli, Ghadeer O Ghosheh, Vinod Kumar Chauhan, Tingting Zhu, Andrew P Creagh, and David A Clifton. Synthesizing mixed-type electronic health records using diffusion models. *IEEE Transactions on Medical Informatics*, 22(4):123–135, 2024.

[10] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, San Francisco, CA, USA, 2016.

[11] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[12] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In *Proceedings of Machine Learning Research*, 2017.

[13] Kaggle. Pima indians diabetes database, 2024. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

[14] Ramana Bendi and N Venkateswarlu. Indian liver patient dataset, 2012. https://doi.org/10.24432/C5D02C.

[15] Federico Soriano. Stroke prediction dataset, 2021. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.

[16] Alistair EW Johnson, Tom J Pollard, and Lu Shen. Mimic-iii, a freely accessible critical care database, 2016. https://physionet.org/content/mimiciii/1.4/.