Journal of Innovations in Engineering Education
JIEE 2026, Vol. 9, Issue 1.

https://doi.org/10.3126/jiee.v9i1.82519

THAPATHALI CAMPUS
Institute of Engineering
Tribhuvan University

A comparative analysis on synthetic data generation of electronic health records using CTGAN, REaLTabFormer and TabDDPM

Girban Adhikari[a,*], Jivan Acharya[a], Arjan Sapkota[a], Subarna Ghimire[a] and Umesh Kanta Ghimire[a]

[a] *Department of Electronics and Computer Engineering, Thapathali Campus, IOE, Tribhuvan University, Nepal*

ARTICLE INFO

Article history:

Received 30 July 2025
Revised in 24 January 2026
Accepted 30 January 2026

Keywords:

CTGAN
Diffusion Models
GAN
Synthetic data generation
Transformers

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.

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 widely adopted 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 adequately 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. While comparative analyses of synthetic EHR generation methods have been reported in prior work, this study differs in its experimental execution by evaluating GAN-, Transformer-, and Diffusion-based models within a single, controlled pipeline. All models are trained on the same preprocessed datasets, evaluated using identical downstream classifiers, and assessed with consistent performance metrics, enabling

*Corresponding author:
adgirban1@gmail.com (G. Adhikari)

a fair and reproducible comparison across generative paradigms. Beyond that, synthetic data are a significant enabler 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.

Unlike prior studies that evaluate generative models under heterogeneous preprocessing pipelines, task settings, or dataset selections, this work implements a unified experimental framework spanning four medical datasets. Each generative model is trained using dataset-specific but consistent preprocessing steps, and synthetic data utility is evaluated through identical downstream classifiers - Logistic Regression, Random Forest, XGBoost, and Neural Networks, using the same train–test splits and evaluation metrics. This controlled design enables direct cross-paradigm comparison and isolates model behavior from confounding experimental variations, providing evidence on trade-offs between data fidelity, downstream utility, and computational cost.

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 show reduced performance 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.

Recent studies have introduced further advances in synthetic EHR generation. A 2025 comparison of multiple GAN variants including CTGAN showed that adversarial models can effectively preserve statistical properties of medical tabular data while supporting downstream classifiers such as XGBoost and SVM [13]. Diffusion based methods have also progressed, as demonstrated by a 2024 model that combines transformer conditioned denoising and dynamic masking to generate high fidelity mixed type tabular data [14]. In addition, fairness aware approaches such as BtGAN [15] aim to reduce subgroup biases in synthetic health data by enforcing balanced

density representations, highlighting that utility, fidelity and fairness are equally important considerations for synthetic EHR generation.

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 adequately 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 [16] 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) [17] provides key attributes for liver disease risk assessment, including bilirubin levels, enzyme concentrations, and albumin ratios. The Stroke Prediction Dataset [18] 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) [19] 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. Although MIMIC-III is inherently longitudinal and relational, this study deliberately aggregates patient information into a single tabular representation to align with the input requirements of CTGAN, REaLTabFormer, and TabDDPM, which are designed for static tabular data. The aggregation process preserves clinically meaningful summary statistics while enabling a fair and consistent comparison across generative models. Some of the tables in this dataset 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

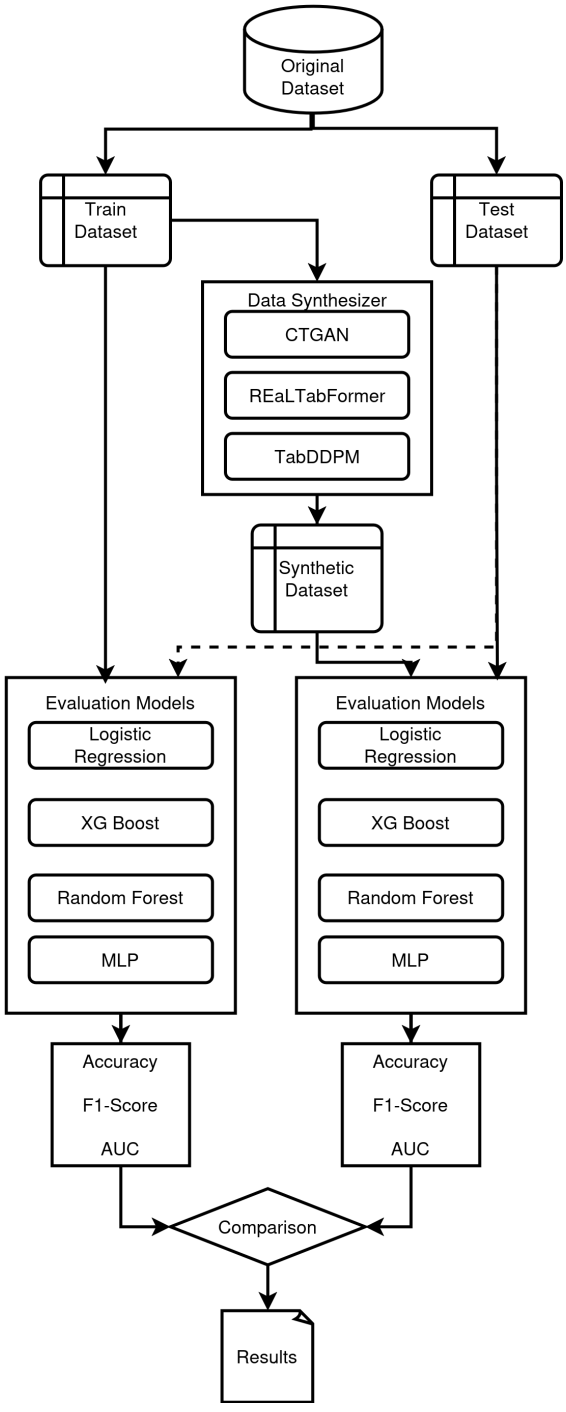


Figure 1: System architecture and workflow

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.

This preprocessing necessarily discards fine-grained temporal dynamics, ICU time-series structure, and event sequences, and therefore does not aim to model full longitudinal EHR generation.

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. This reduces noise and stabilizes model inputs, allowing generative models to focus on clinically relevant signal rather than raw temporal fluctuations.

Temporal alignment Time-series data were synchronized based on timestamps to maintain the sequence of events and ensure consistency across different tables. This alignment ensures that physiological trends are captured correctly and prevents temporal drift that could distort patient trajectories.

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. These derived variables capture clinically meaningful thresholds and contribute to more interpretable synthetic data outputs.

Data loading and inspection The dataset was first loaded and inspected for missing values, variable distributions, and anomalies to inform preprocessing steps. Initial exploratory analysis helps identify potential biases or irregularities early, ensuring that downstream modeling remains robust.

Handling missing values Missing data were imputed using statistical measures (mean, median, mode) or removed if missingness was excessive to maintain dataset integrity. This prevents generative models from learning erroneous patterns and ensures consistent dimensionality across samples.

Removing invalid data Records with unrealistic values (e.g., negative mortality indicators) were identified and excluded to prevent biases. Such filtering preserves data quality and avoids contaminating the generative process with non-physiological artifacts.

Feature engineering Categorical variables were encoded numerically, and additional features were created to capture meaningful patterns, such as mortality flags. Encoding and derived features improve compatibility with neural architectures and enhance the expressiveness of the feature space.

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. This step accelerates convergence during optimization and prevents high-variance features from dominating the learning process.

Each generative model was trained using the training strategy recommended in its original implementation. CTGAN and TabDDPM were trained for a fixed number of epochs to ensure convergence, whereas REaLTabFormer employs early stopping and converges within 20–60 epochs as reported in prior work. As a result, epoch counts are not directly comparable across architectures; however, all models were trained until loss stabilization rather than prematurely terminated. 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. CTGAN required only 2 hours of training with modest memory usage (roughly 5 GB), whereas REaLTabFormer and TabDDPM required 6–12 hours and 10–16 hours respectively with substantially higher GPU memory (roughly 10 GB), highlighting the practical computational gap between GAN, Transformer, and Diffusion-based EHR generators. 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 demonstrates competitive performance in some cases but shows reduced performance 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.

The reported results should be interpreted as a utility-focused evaluation and do not, by themselves, guarantee distributional fidelity or privacy safety of the generated synthetic data. Also, the performance results are presented as single-point estimates and do not include confidence intervals or statistical significance testing, which limits claims about performance variability across runs. Although TabDDPM attains higher scores in several metrics, Table 3 shows overlapping and dataset-dependent performance across models, suggesting comparative strengths rather than universal dominance. Consequently, conclusions drawn from the MIMIC-III experiments should be interpreted as applicable to aggregated tabular EHR representations rather than full temporal or sequential clinical records.

4.1. Mortality dataset

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 properly but may introduce minor shifts in synthetic minority class representations, as seen in Figure 2.

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**

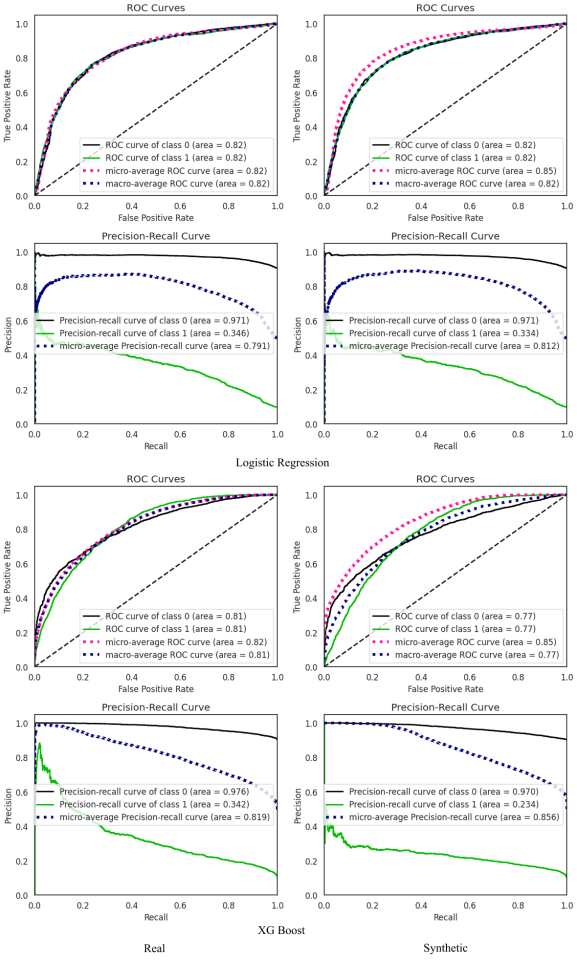


Figure 2: ROC and PR curves on mortality for logistic regression and XGBoost (Using REaLTabFormer)

4.2. 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.

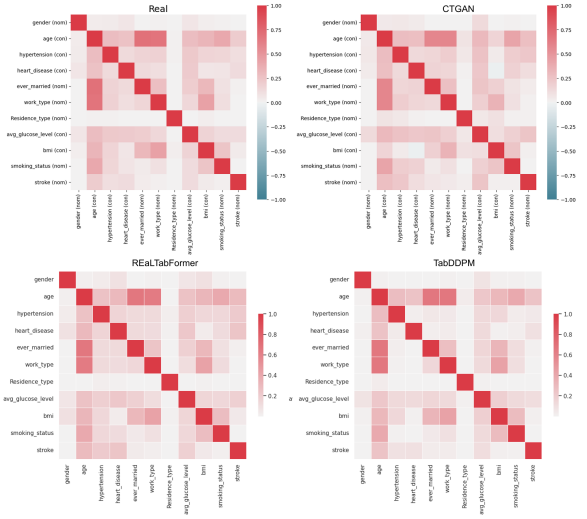


Figure 3: Heatmaps on stroke dataset

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. For the Stroke dataset, REaLTabFormer more effectively preserves certain feature correlations, indicating that Transformer-based models may be preferable when structured dependency retention is prioritized.

4.3. ILPD dataset

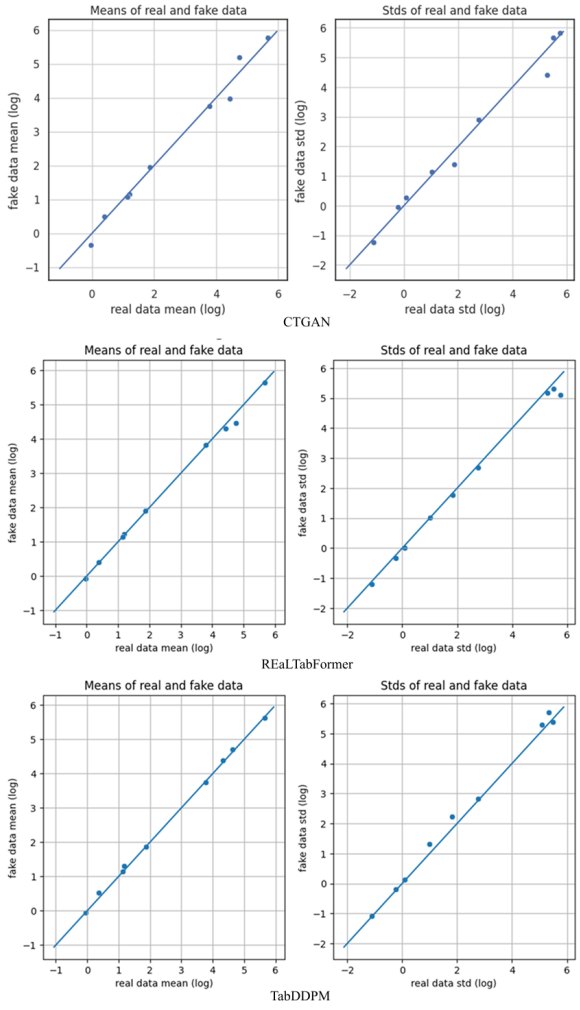


Figure 4: Logmeans of ILPD dataset

Figure 4 evaluates the statistical consistency of the real and synthetic ILPD datasets by comparing their feature-wise means and standard deviations on a log scale. The identity line (y = x) represents perfect agreement, providing a visual benchmark for how faithfully each model reproduces the underlying distributional characteristics of the real data. Among the three generative models, TabDDPM shows the strongest correspondence with this ideal, reflecting its superior ability to preserve both central tendencies and dispersion across features.

REaLTabFormer also achieves a generally strong alignment, though slight deviations indicate that certain features are not captured with complete fidelity. In contrast, CTGAN displays more pronounced discrepancies, particularly in replicating the standard deviations of several features, suggesting limitations in modeling the full range of variability present in the ILPD dataset. These observations collectively highlight the relative advantage of diffusion-based and transformer-based approaches over traditional GAN-based models for this dataset.

4.4. PIMA dataset

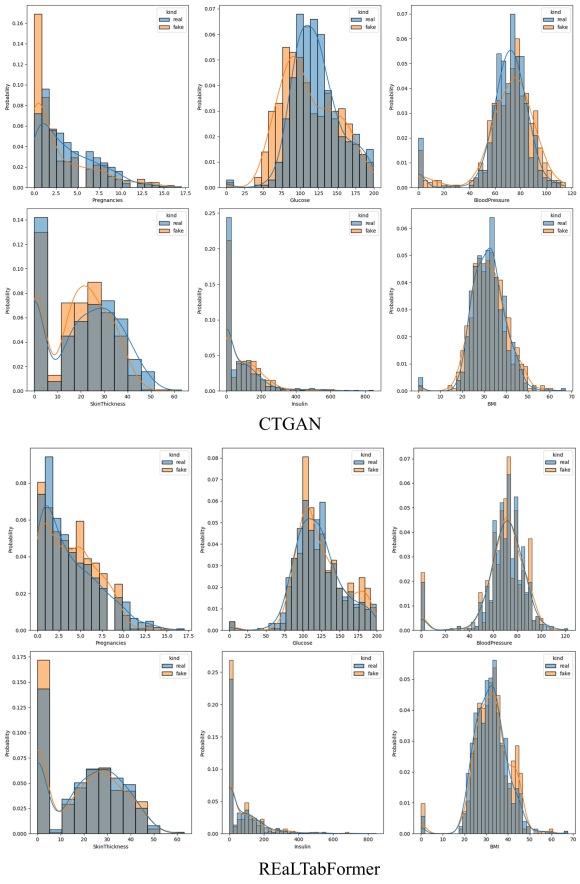
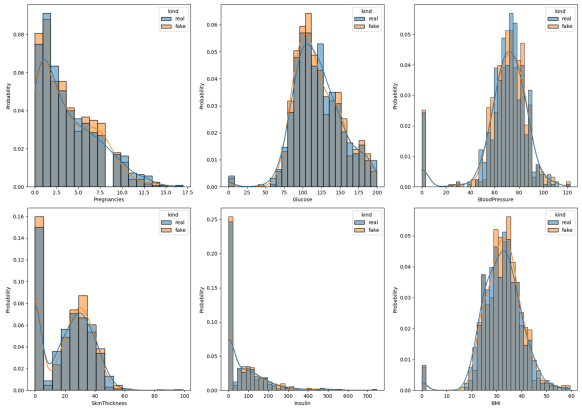


Figure 5: Distribution per feature of PIMA dataset for CTGAN and REaLTabformer

Figure 5 displays the distribution of six key features, which are 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 shows reduced performance 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.



TabDDPM

Figure 6: Distribution per feature of PIMA dataset for TabDDPM

5. Conclusion

As seen from Table 3 CTGAN demonstrates competitive performance 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 shows reduced performance 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. Overall, TabDDPM demonstrates strong and often superior performance in terms of distributional fidelity and downstream utility; however, no single model is uniformly optimal across all datasets and evaluation criteria.

In a small number of cases, models trained on synthetic data marginally outperform those trained on real data. This behavior may arise from the smoothing and implicit regularization effects of generative models, which can reduce noise and class imbalance present in real-world clinical data. Nevertheless, such results warrant caution, as they may also indicate overfitting or unintended data leakage, which were not explicitly tested in this study.

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.

Comprehensive synthetic data quality evaluation The evaluation in this study primarily emphasizes downstream task utility, assessed through Accuracy, F1-score, and ROC-AUC, which reflects the practical usability of synthetic EHR data for predictive modeling. While this approach provides meaningful insight into model utility, it does not fully capture all dimensions of synthetic data quality that are critical in healthcare applications.

Privacy and memorization risk assessment Although privacy preservation is a central motivation for synthetic data generation, this study does not explicitly evaluate privacy leakage or memorization risk. Future extensions should therefore include privacy-focused evaluations such as membership inference attacks, nearest-neighbor distance analysis, and attribute disclosure risk assessments to ensure that synthetic samples do not unintentionally reveal sensitive patient information.

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 worth Rs.10,000 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] Xu L, Skoularidou M, Cuesta-Infante A, et al. Modeling tabular data using conditional gan[C]// Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI). 2019.

[2] Kingma D P, Welling M. Auto-encoding variational bayes[C]// International Conference on Learning Representations (ICLR). 2013.

[3] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[C]// Advances in Neural Information Processing Systems (NIPS). 2014: 2672-2680.

[4] Arjovsky M, Chintala S, Bottou L. Wasserstein gan[C]// Proceedings of the 34th International Conference on Machine Learning (ICML). Sydney, Australia, 2017.
[5] Solatorio A V, Dupriez O. Realtabformer: Generating realistic relational and tabular data using transformers[J]. IEEE Transactions on Artificial Intelligence, 2024, 10(3): 456-468.
[6] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, 2017: 6000-6010.
[7] Kotelnikov A, Baranchuk D, Rubachev I, et al. Tabddpm: Modelling tabular data with diffusion models[J/OL]. arXiv preprint arXiv:2209.15421, 2022. https://arxiv.org/abs/2209.15421.
[8] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[C]// Proceedings of the 34th Conference on Neural Information Processing Systems. Vancouver, BC, Canada, 2020.
[9] Ceritli T, Ghosheh G O, Chauhan V K, et al. Synthesizing mixed-type electronic health records using diffusion models[J]. IEEE Transactions on Medical Informatics, 2024, 22(4): 123-135.
[10] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA, 2016: 785-794.
[11] Breiman L. Random forests[J]. Machine Learning, 2001, 45 (1): 5-32.
[12] Choi E, Biswal S, Malin B, et al. Generating multi-label discrete patient records using generative adversarial networks[C]// Proceedings of Machine Learning Research. 2017.
[13] Ahmed H A, Nepomuceno J A, Vega-Márquez B, et al. Synthetic data generation for healthcare: Exploring generative adversarial networks variants for medical tabular data[J/OL]. International Journal of Data Science and Analytics, 2025, 20: 5739-5754. DOI: 10.1007/s41060-025-00816-w.
[14] Villaizán-Vallelado M, Salvatori M, Segura C, et al. Diffusion models for tabular data imputation and synthetic data generation[J]. ACM Transactions on Knowledge Discovery from Data, 2025, 19(6): 1-32.
[15] RamachandranPillai R, Sikder M F, Bergström D, et al. Bt-gan: Generating fair synthetic healthdata via bias transforming generative adversarial networks[J/OL]. CoRR, 2024, abs/2404.13634. https://arxiv.org/abs/2404.13634.
[16] Kaggle. Pima indians diabetes database[EB/OL]. 2024. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.
[17] Bendi R, Venkateswarlu N. Indian liver patient dataset[EB/OL]. 2012. https://doi.org/10.24432/C5D02C.
[18] Soriano F. Stroke prediction dataset[EB/OL]. 2021. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.
[19] Johnson A E, Pollard T J, Shen L. Mimic-iii, a freely accessible critical care database[Z]. 2016.