Evaluating CTGAN-Generated Synthetic Data for Heart Disease Prediction: Fidelity, Predictive Utility, and Feature Preservation

Wan Aezwani Wan Abu Bakar[1], Nur Laila Najwa Josdi[2]*, Mustafa Man[3]*, Evizal Abdul Kadir[4]
Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Besut Campus, 22200 Besut, Terengganu, Malaysia[1, 2]
Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia[3]
Universitas Islam Riau, Jl. Kaharuddin Nasution 113, Pekanbaru 28284, Riau, Indonesia[4]

***Abstract*—The increasing scarcity and sensitivity of clinical data necessitate the development of high-quality synthetic datasets. This study evaluated the ability of Conditional Tabular GAN (CTGAN) to generate synthetic heart disease data that preserves the statistical properties and predictive patterns of the Cleveland Heart Disease dataset. It assessed the fidelity of numerical and categorical features, preservation of pairwise correlations, and predictive utility using Logistic Regression and Random Forest classifiers. Dimensionality reduction analysis using PCA and t-SNE further measured the global similarity between the real and synthetic datasets. The results obtained show that CTGAN successfully reproduces the general distribution and correlations, especially for key features such as age, talach, and old peak. However, some discrepancies remain in categorical attributes. Predictive modeling shows moderate transferability, indicating that synthetic data captures important patterns without completely replicating the original labels. These findings highlight the potential of CTGAN-generated synthetic data as a privacy-preserving alternative for benchmarking and early algorithm development, while emphasizing the importance of feature-level and prediction validation in synthetic data research.**

Keywords*—*Conditional Tabular GAN (CTGAN); correlation analysis; dimensionality reduction; feature importance; heart disease prediction; predictive utility; synthetic data; tabular data fidelity

I. Introduction

Cardiovascular diseases, particularly heart disease, remain a leading cause of global mortality, accounting for approximately 17.9 million deaths annually according to the World Health Organization. Early identification and timely intervention are essential to reducing morbidity and mortality associated with these conditions. In recent years, machine learning (ML) techniques have attracted significant attention for their potential to support heart disease diagnosis and risk prediction, thereby enabling data-driven clinical decision-making [1].

Despite these advances, the robustness and generalizability of ML models in healthcare are often constrained by limitations inherent to real-world clinical data [2]. One of the primary challenges is the scarcity of large-scale, high-quality, and representative datasets. Publicly available resources, such as the UCI Heart Disease dataset, are typically limited in size and frequently exhibit class imbalance, which restricts their suitability for training reliable predictive models [3]. In addition, patient privacy regulations and ethical considerations impose further restrictions on data sharing and reuse [4]. These challenges collectively hinder reproducibility and limit the practical deployment of ML-based solutions in clinical settings [5].

To address data scarcity while mitigating privacy concerns, synthetic data generation has emerged as a promising alternative. Among existing approaches, Generative Adversarial Networks (GANs) have demonstrated strong capability in modeling complex data distributions. In particular, Conditional Tabular GAN (CTGAN) is designed to handle tabular datasets with both continuous and categorical variables, as well as class imbalance. These characteristics make CTGAN especially suitable for healthcare applications [6], where heterogeneous feature types and skewed data distributions are common.

Although GAN-based synthetic data generation has been widely explored, most existing studies primarily emphasize downstream model performance, such as classification accuracy, while providing limited systematic evaluation of artificial data fidelity. In particular, the preservation of statistical properties and feature-level relationships in synthetic tabular data is often under-examined. Furthermore, there is insufficient empirical evidence regarding whether synthetic tabular data generated for heart disease prediction can offer comparable predictive utility to real clinical data while maintaining essential statistical and structural characteristics.

To address these gaps, this study evaluates the quality and fidelity of CTGAN-generated synthetic heart disease data through a comprehensive assessment framework. The evaluation includes descriptive statistical comparisons, correlation structure analysis, distributional similarity testing, and utility-based evaluation using a logistic regression classifier implemented in Python. Rather than augmenting predictive pipelines, the study assesses whether synthetic data can serve as a statistically reliable substitute for real clinical data in downstream machine learning tasks.

By integrating statistical fidelity analysis with predictive utility evaluation, this work provides empirical insights into the suitability of CTGAN-generated synthetic data for healthcare machine learning applications. The findings contribute to ongoing discussions on the role of synthetic data in clinical

*Corresponding author.

artificial intelligence, particularly in scenarios involving sensitive or limited-access tabular datasets.

The remainder of this paper is organized as follows. Section II reviews related work on synthetic data generation and evaluation in healthcare. Section III describes the dataset, CTGAN configuration, and the methodology for synthetic data generation and evaluation. Section IV presents the experimental results. Section V discusses the findings in relation to existing studies and outlines the limitations. Finally, Section VI concludes the paper and suggests directions for future research, followed by the Acknowledgments and References.

II. Literature Review

Heart disease prediction increasingly relies on machine learning (ML) and deep learning (DL), offering potential for timely diagnosis and improved clinical decision-making. However, challenges such as limited high-quality clinical data and persistent class imbalances drive the use of synthetic data generation. Generative Adversarial Networks (GANs) have emerged as a promising approach to generate realistic tabular datasets while preserving patient privacy. This literature review focuses on three areas: (1) ML and DL approaches for heart disease prediction, (2) GAN-based synthetic data generation in healthcare, and (3) assessment of the quality, utility, and privacy of synthetic data. Across these themes, this review highlights methodological advances, benchmarking strategies, and remaining gaps, thereby providing a foundation for this study.

A. Heart Disease Prediction Using Deep Learning and Machine Learning

Traditional machine learning (ML) models, such as logistic regression, decision trees, support vector machines (SVMs), k-nearest neighbors (KNN), and random forests (RFs), have been widely used for heart disease prediction due to their interpretability and computational simplicity [7,8]. Recent studies demonstrate that classical ML models, particularly tree-based methods, achieve high predictive performance across diverse datasets. For instance, Random Forest and Bagged Trees consistently produced 94–99% accuracy in multi-site validations, highlighting their reliability [15,16]. Nevertheless, these approaches often depend on manual feature engineering and may struggle to capture nonlinear relationships inherent in high-dimensional clinical data.

To address these limitations, deep learning (DL) models, intense neural networks (DNNs), and backpropagation neural networks (BP-NNs) have been increasingly employed. When combined with feature selection techniques such as L1-regularized LinearSVC or minimum redundancy maximum relevance (mRmR), these models consistently demonstrate improved predictive accuracy, often exceeding 98% [9,10]. These findings emphasize the importance of intelligent feature selection to optimize DL performance.

DL models have also been applied to non-tabular data, such as cardiac sound signals. Using Mel-frequency cepstral coefficients (MFCC) extracted from heart sounds, DL classifiers achieved near-perfect test accuracy and F1-scores [11]. Similarly, mobile-based cardiac screening employing DL has shown promising results for accessible, non-invasive diagnosis [12]. These studies illustrate the potential of incorporating alternative data modalities to complement traditional clinical predictors.

Ensemble and hybrid models have further enhanced prediction accuracy by combining multiple algorithms or data sources. For example, a DL ensemble integrating electronic medical record (EMR) and sensor data achieved 98.5% accuracy using information gain weighting [13]. Stacked ensembles of Extra Trees, XGBoost, and RF yielded robust accuracy around 92–95% [14]. Hybrid frameworks combining KNN, XGBoost, and LSTM also outperformed individual classifiers across multiple datasets [20]. These results underscore the effectiveness of multi-model strategies in capturing complementary predictive patterns.

Recent studies have explored additional predictors beyond standard clinical features. For example, SHAP analysis applied to Random Forest models identified novel biochemical markers, such as calcium and inflammation indicators, as relevant for cardiovascular risk [17]. Similarly, genomic data and psychological variables have been integrated within ML frameworks, improving predictive performance from 71.3% to 85.1% [18, 19]. These findings suggest that incorporating diverse data types can enhance model generalizability and interpretability.

Despite these promising results, generalizability remains limited. A systematic review found that only 10 out of 486 AI-based cardiovascular models underwent meaningful external validation [21]. Common limitations include small sample sizes, class imbalance, and strict privacy regulations, which hinder robust training and evaluation. These challenges highlight the need for data augmentation strategies, such as synthetic data generation via Generative Adversarial Networks (GANs), to address data scarcity while maintaining privacy. The following sections explore recent advances in GAN-based tabular synthesis, benchmarking approaches for evaluating fidelity and fairness, and real-world applications of synthetic clinical data.

B. Synthetic Data Generation in Healthcare using GANs, Disease Prediction using Deep Learning, and Machine Learning

Recent work has introduced enhanced GAN variants tailored for the unique characteristics of clinical tabular datasets. CTAB-GAN+ [22] incorporates task-aware loss functions, mixed-type variable encoders, and optional differential privacy mechanisms, improving F1-score performance under privacy constraints. STNG [23] integrates multiple generative approaches (Gaussian Copula, CopulaGAN, TVAE, Conditional GAN) with downstream validation via AutoML, demonstrating high fidelity across diverse medical datasets. SMOOTH-GAN [24] emphasizes clinically valid synthetic distributions while balancing privacy. In contrast, transformer-based models such as TT GAN [25] and FCT GAN [26] further improve the modeling of complex inter-variable dependencies in heterogeneous datasets. These studies collectively highlight the trend toward domain-aware, privacy-conscious GAN architectures capable of generating realistic clinical data.

To ensure synthetic data is not only realistic but also practical, several benchmarking frameworks have emerged. End-to-end pipelines [27] evaluate fidelity, privacy, and utility,

revealing that GANs remain competitive even compared to diffusion-based models. Comparative studies [28–30] have consistently shown that CTGAN and WGAN-GP preserve inter-variable correlations and downstream classifier performance, while hybrid methods such as CTGAN-ENN [31] and BT-GAN [32] address class imbalance and subgroup fairness. Overall, these evaluations emphasize the need for multi-dimensional assessment (fidelity, utility, fairness) to validate synthetic tabular data for healthcare applications.

Practical deployment of synthetic data requires attention to privacy and regulatory alignment. Studies using Bayesian networks and GANs on large datasets (e.g., CPRD) [33] stress ethical deployment with privacy risk mitigation, such as k-anonymity and differential privacy. Alternative approaches generate synthetic data from summary statistics to avoid direct access to patient-level data [34], while pipelines such as EMR-WGAN [35] provide hands-on workflows for real-world model development. Advanced models such as IGAMT [36] and DP-CGANs [37] further ensure privacy, fairness, and high downstream utility, particularly for heterogeneous and irregular clinical tabular data. These studies highlight a growing emphasis on ethically aligned, privacy-preserving synthetic data generation suitable for clinical AI research.

Despite the development of sophisticated GAN architectures and benchmarking strategies, questions remain regarding how to rigorously evaluate synthetic data quality and ensure it is both valuable and privacy-preserving in clinical contexts. The following section discusses evaluation frameworks, performance metrics, and architectural innovations that address these challenges.

C. Evaluation of Synthetic Data Quality and Utility in Healthcare

The evaluation of synthetic healthcare data revolves around three foundational pillars: fidelity, utility, and privacy. Fidelity ensures that synthetic data accurately replicates real-world distributions, utility reflects its usefulness in downstream analyses, and confidentiality safeguards sensitive patient information [38]. These dimensions were formalized in early frameworks and have since guided multiple evaluation standards. Preserving inter-variable correlations in time-series health data [39] and implementing domain-aware pipelines for EHRs [40] demonstrate practical strategies to maintain these dimensions. Comprehensive benchmarking studies [41,42] combine statistical similarity metrics (e.g., KS tests, Chi-square tests) with privacy assessments, such as membership inference attacks. SynthRO [43], using MIMIC-III/IV datasets, highlights the trade-offs between fidelity, utility, and privacy, emphasizing the need for balanced optimization in synthetic data generation.

Performance of synthetic data varies across clinical domains, making context-aware evaluation essential. In oncology and diabetes datasets, a divide-and-conquer CTGAN approach achieved high fidelity (97.65% shape score) and strong classification metrics [44]. Conditional GANs tailored for cardiovascular data outperformed standard CTGAN and TVAE models in fidelity measures such as the Jaccard index and KS statistic [45]. Modeling survival curves in oncology trials further illustrates the challenge of replicating censored time-to-event data [46]. Comparative analyses [47,48] highlight that no single model universally outperforms others, reinforcing the importance of domain-specific benchmarks. Broader evaluation frameworks [49] also incorporate fairness and environmental cost alongside fidelity and utility, underscoring the increasing complexity of evaluation metrics in clinical applications.

Architectural improvements in GAN-based models have targeted fidelity, usability, and privacy. Divide-and-conquer CTGAN [50] and MargCTGAN [51] enhance downstream classification performance and preserve marginal distributions, particularly for small clinical datasets. DP-CTGAN [52] integrates differential privacy directly into training to ensure privacy-conscious generation. Memory-efficient models such as MeTGAN [53] address high-cardinality categorical features while reducing computational overhead, and OCT-GAN [54] incorporates neural ordinary differential equations (NODEs) to handle class imbalance and multimodal distributions. Collectively, these innovations reflect a trend toward modular, scalable, and ethically aligned generative frameworks that meet the complex demands of healthcare data.

Although Generative Adversarial Networks (GANs) have gained traction as a solution for data scarcity in healthcare, most existing studies focus primarily on augmenting training datasets to improve predictive model performance. While this utility-based approach has achieved notable improvements in classification accuracy and robustness, there remains a lack of systematic assessment of GAN-generated data as a stand-alone substitute for real patient data. Evaluations frequently prioritize high-level utility metrics without thoroughly examining whether synthetic data preserves the intricate statistical and structural patterns inherent in clinical datasets. CTGAN, a GAN variant designed for mixed-type tabular data, shows promise in handling categorical and continuous variables with imbalanced distributions.

However, limited studies have rigorously assessed its ability to maintain clinically meaningful inter-feature relationships, such as those among chest pain, ECG readings, and heart disease diagnosis, which are critical for interpretability and real-world applicability. This study addresses these gaps by conducting a comprehensive evaluation of CTGAN-generated synthetic data for heart disease prediction, focusing on statistical fidelity, multivariate structure, and downstream classification performance. By doing so, it provides deeper insights into the reliability and limitations of CTGAN in generating synthetic clinical data for sensitive domains like cardiovascular healthcare.

III. Methodology

To address the challenges posed by scarce and sensitive clinical datasets, synthetic data generation using Generative Adversarial Networks (GANs) has gained prominence in clinical research. Further leveraging Conditional Tabular GAN (CTGAN) models, this study generates synthetic medical data from benchmark datasets to evaluate heart disease prediction performance. The primary objective is to assess the quality of the artificial data by comparing its statistical properties and predictive utility with those of the original dataset.

A. Data Source and Preprocessing

The Cleveland Heart Disease dataset was retrieved from the UCI Machine Learning Repository using the ucimlrepo Python package. This dataset [55] contains 303 patient records with 14 attributes (13 features and one target). The target column, `num`, indicates heart disease diagnosis:

- 0: Absence.
- 1-4: Presence with increasing severity.

The dataset contains 13 features and one target variable relevant to cardiac health (Table I).

TABLE I CLEVELAND HEART DISEASE DATASET FEATURES

Feature	Type	Description
age	Integer	Age in years
sex	Categorical	Sex (1=male, 0=female)
cp	Categorical	Chest pain type (1–4)
trestbps	Integer	Resting blood pressure (mm Hg)
chol	Integer	Serum cholesterol (mg/dl)
fbs	Categorical	Fasting blood sugar >120 mg/dl
restecg	Categorical	Resting ECG results (0–2)
thalach	Integer	Maximum heart rate achieved
exang	Categorical	Exercise-induced angina (1=yes, 0=no)
oldpeak	Integer	ST depression induced by exercise
slope	Categorical	Slope of peak exercise ST segment
ca	Integer	Number of major vessels (0–3)
thal	Categorical	Thalassemia (3=normal, 6=fixed defect, 7=reversible defect)
num	Target Integer	Diagnosis of heart disease

Two columns, ca and thal, contained missing values, which were imputed using the mode of each column. Numerical features (age, trestbps, chol, thalach, oldpeak) were preserved in their original scale. In contrast, categorical features (sex, cp, fbs, restecg, exang, slope, ca, thal) were encoded to facilitate synthetic data generation. The missing values were handled as follows:

```
df['ca'].fillna(df['ca'].mode()[0], inplace=True)
df['thal'].fillna(df['thal'].mode()[0], inplace=True)
```

B. Synthetic Data Generation Using CTGAN

The CTGAN model was employed to generate synthetic medical records from the cleaned and preprocessed dataset (`df_v2`). This dataset consisted of 303 patient instances, 13 predictive features, and one target variable. All preprocessing steps, including missing value imputation, categorical encoding, and 1st-to-99th percentile clipping, were completed before training.

1) CTGAN configuration: Fig. 1 shows a screenshot of the Jupyter Notebook code used to train the CTGAN model on the Cleveland Heart Disease dataset. The figure illustrates the setup of model parameters, categorical feature specification, and the output progress of generator and discriminator losses during training.

```
from ctgan import CTGAN
import pandas as pd

df_v2 = pd.read_csv("heart_clean_v2_clipped.csv")

categorical_columns = ['sex','cp','fbs','restecg','exang','slope','ca','thal']

ctgan = CTGAN(
    epochs=1200,
    batch_size=256,
    embedding_dim=256,
    generator_lr=2e-4,
    discriminator_lr=2e-4,
    verbose=True,
    pac=1  # fixes the AssertionError for small datasets
)

ctgan.fit(df_v2)
```

```
Gen. (-0.85) | Discrim. (-0.04): 100%|██████████| 1200/1200 [00:58<00:00, 20.57it/s]
```

Fig. 1. Screenshot of CTGAN training code and output progress for the cleveland heart disease dataset.

2) Model training: The CTGAN was trained directly on the cleaned dataset. Since categorical variables in `df_v2` were encoded as integers, CTGAN automatically inferred their discrete nature during training. Training completed successfully, with stable Wasserstein losses for both the generator (≈ –0.85) and discriminator (≈ –0.04), indicating balanced adversarial optimization without signs of divergence or mode collapse.

3) Synthetic data generation: After training, a synthetic dataset of equal size to the original (`n = 303`) was generated. Matching the sample size allowed direct pairwise comparison of statistical properties and predictive utility between real and synthetic datasets. The synthetic data was generated as follows:

```
num_samples = df_v2.shape[0]
synthetic_data = ctgan.sample(num_samples)
synthetic_df = pd.DataFrame(synthetic_data, columns=df_v2.columns)
```

To ensure structural and semantic consistency with the real dataset, categorical variables were rounded and cast to integer types. At the same time, numeric features (e.g., chol, oldpeak) were clipped to medically valid ranges to prevent unrealistic values. These adjustments ensured that the synthetic dataset adhered to clinically plausible ranges while preserving statistical fidelity.

C. Statistical Evaluation of Synthetic Data

To assess the fidelity of the synthetic dataset, multiple statistical analyses were performed to compare distributions and inter-feature relationships between the real and synthetic datasets, as detailed below.

Kolmogorov–Smirnov (KS) tests and Wasserstein distances were computed for continuous features (age, trestbps, chol, thalach, oldpeak) to quantify distributional differences between real and synthetic datasets. Features with significant KS p-values (<0.05) or elevated Wasserstein distances were flagged as exhibiting moderate deviations.

Correlation analysis was performed to compare pairwise feature relationships between the real and synthetic datasets. Features showing substantial deviations in Pearson correlation coefficients were highlighted to identify potential discrepancies in the underlying structure.

Exploratory Data Analysis (EDA) was conducted to identify features with moderate or significant deviations based on distributional and proportional differences.

These analyses provide a quantitative foundation for evaluating whether the synthetic data preserves the statistical characteristics of the original dataset.

D. Utility Evaluation of Synthetic Data

To assess the practical utility of the synthetic dataset, predictive models were trained and evaluated on both the real and synthetic datasets. This evaluation measures whether the synthetic data preserves the predictive patterns present in the real dataset.

1) Restoring target column for predictive modeling: The target column (num) was restored in both the real and synthetic datasets to enable predictive evaluation. Features and target were separated as follows:

$$X\_real, y\_real \text{ and } X\_synth, y\_synth$$

Where *X_real* contains all feature columns from the real dataset, *y_real* contains the target column (num) from the real dataset, *X_synth* contains all feature columns from the synthetic dataset, and *y_synth* contains the target column (num) from the artificial dataset.

2) Dataset splitting: Both datasets were split into training and testing subsets using a 70:30 ratio to enable unbiased evaluation. Stratified sampling was applied to maintain the proportion of patients with and without heart disease (num) across subsets. The training data was used to fit the classifiers, and the testing data was used to evaluate predictive performance. The dataset splitting was implemented as follows:

```
From sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(
X, y, test_size=0.3, random_state=42, stratify=y)
```

3) Predictive modeling: Two classifiers were employed:

a) Logistic Regression (LR): A linear baseline model for binary classification.

a) *Random Forest (RF):* an ensemble tree-based model capable of capturing non-linear relationships. Each model was incorporated into a pipeline with preprocessing for categorical and numerical features. Models were trained on the training subset and evaluated on the testing subset to ensure reproducible performance assessment.

4) Performance metrics: Model performance was assessed using standard classification metrics: Accuracy, Precision, Recall (Sensitivity), F1-score, ROC-AUC, and the Confusion Matrix. This procedure provides a reproducible framework to quantify the predictive utility of the synthetic data. Results and corresponding visualizations are presented in the Results section.

IV. Result

A. Numerical Feature Fidelity

The CTGAN-generated synthetic dataset was first evaluated for the fidelity of continuous features relative to the original Cleveland Heart Disease dataset. Table II presents a comparison of numeric features, including mean, standard deviation, percentage difference, and overall fidelity rating for each feature.

TABLE II Comparison of Numeric Features Between Real and Synthetic Data

Feature	Real μ[a]	Synth μ[a]	Real σ[a]	Synth σ[a]	Δ%[b]	Fidelity[c]
age	54.42	51.68	8.88	8.55	5.04%	M
trestbps	131.62	124.78	17.17	20.99	5.20%	M
chol	246.29	207.42	48.82	46.49	15.78%	W
thalach	149.68	162.95	22.44	19.96	8.87%	M
oldpeak	1.03	0.99	1.12	1.06	3.53%	S

[a]μ and σ denote the mean and standard deviation of the respective datasets.
[b]Δ% represents the percentage difference between the real and synthetic mean values.
[c]Fidelity ratings are categorized as Strong (S), Moderate (M), or Weak (W).

Fig. 2 presents kernel density estimates (KDEs) of numeric features, comparing real and synthetic datasets.

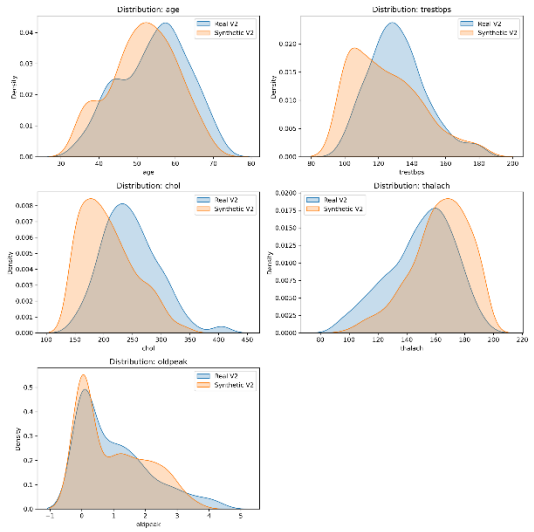

Fig. 2. KDE comparison of numeric feature distributions (real vs. synthetic).

B. Categorical Feature Fidelity

Categorical features were evaluated for fidelity by calculating the maximum proportional differences between the real and synthetic datasets. Table III highlights the top features with their corresponding fidelity assessments, indicating which variables were well-preserved and which exhibited larger deviations.

TABLE III Categorical Feature Fidelity Between Real and Synthetic Datasets

Feature	Max % Difference	Fidelity
sex	3.96%	Strong
cp	7.59%	Moderate
fbs	22.11%	Weak
restecg	10.56%	Weak
exang	0.33%	Strong
slope	17.16%	Weak
ca	11.22%	Weak
thal	11.88%	Weak

C. Distributional Similarity

To further evaluate the fidelity of numeric features, Kolmogorov–Smirnov (KS) tests and Wasserstein distances were calculated, providing quantitative measures of distributional similarity between the real and synthetic datasets. Table IV summarizes these results, highlighting features with the most significant discrepancies.

TABLE IV DISTRIBUTIONAL SIMILARITY OF NUMERIC FEATURES BETWEEN REAL AND SYNTHETIC DATASETS

Feature	KS Statistic	KS p-value	Wasserstein Distance
age	0.1782	0.0001	2.74
trestbps	0.2937	0.0000	7.17
chol	0.3597	0.0000	38.87
thalach	0.2607	0.0000	13.27

D. Feature Correlation Analysis

Pairwise feature correlations were analyzed to assess whether the CTGAN-generated synthetic dataset preserved the relationships observed in the original dataset. Table V lists the top five feature pairs exhibiting the most considerable absolute differences in correlation.

TABLE V COMPARISON OF PAIRWISE FEATURE CORRELATIONS BETWEEN REAL AND SYNTHETIC DATASETS

Feature 1	Feature 2	Real Corr	Synthetic Corr	|Δ|[a]
age	ca	0.365	-0.042	0.41
oldpeak	slope	0.575	0.273	0.30
restecg	chol	0.166	0.442	0.28
trestbps	fbs	0.170	-0.104	0.27
ca	num	0.521	0.288	0.23

[a]|Δ| = absolute difference in means

Fig. 3 visualizes these differences in a heatmap, providing an intuitive overview of how inter-feature relationships are maintained or altered in the synthetic dataset.

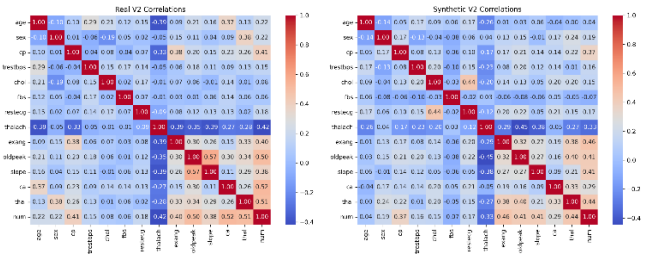

Fig. 3. Heatmap of correlation differences between real and synthetic datasets.

E. Predictive Utility Evaluation

To evaluate the predictive utility of the CTGAN-generated synthetic dataset, Logistic Regression and Random Forest classifiers were trained on one dataset and tested on the other. This setup assesses whether the synthetic data preserves the predictive patterns of the original dataset. Table VI summarizes the classification performance metrics (Accuracy, Precision, Recall, F1-score) for each training-testing scenario.

TABLE VI PREDICTIVE PERFORMANCE OF CLASSIFIERS TRAINED ON SYNTHETIC VS. REAL DATASETS

Train → Test	Model	Accuracy	Precision	Recall	F1
Synthetic → Real	Logistic Regression	0.560	0.566	0.560	0.561
Synthetic → Real	Random Forest	0.560	0.608	0.560	0.561
Real → Synthetic	Logistic Regression	0.615	0.557	0.615	0.571
Real → Synthetic	Random Forest	0.582	0.464	0.582	0.500
Train → Test	Model	Accuracy	Precision	Recall	F1

To demonstrate the evaluation procedures used in this study, confusion matrices were generated programmatically in Python within a Jupyter Notebook environment. The implementation includes model predictions, confusion matrix computation, and visualization using Seaborn heatmaps. Fig. 4 provides a snapshot of the evaluation code, illustrating the end-to-end process used to assess classifier performance across real and synthetic datasets.

```python
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import confusion_matrix

# Function to plot and save confusion matrix as heatmap
def plot_confusion_matrix(y_true, y_pred, classes, title, filename):
    cm = confusion_matrix(y_true, y_pred, labels=classes)
    plt.figure(figsize=(6,5))
    sns.heatmap(cm, annot=True, fmt="d", cmap="Blues", xticklabels=classes, yticklabels=classes)
    plt.xlabel("Predicted Label")
    plt.ylabel("True Label")
    plt.title(title)
    plt.tight_layout()
    plt.savefig(filename, dpi=300)
    plt.show()

# Define class labels
classes = sorted(y_real.unique())

# Dictionary of models
models = {
    "Logistic Regression": pipe_lr,
    "Random Forest": pipe_rf
}

# 🔁 Train on Synthetic → Test on Real
for name, model in models.items():
    y_pred = model.predict(X_test_real)
    plot_confusion_matrix(
        y_test_real,
        y_pred,
        classes=classes,
        title=f"{name}: Train on Synthetic → Test on Real",
        filename=f"cm_{name.replace(' ','_')}_syn2real.png"
    )

# 🔁 Train on Real → Test on Synthetic
for name, model in models.items():
    y_pred = model.predict(X_test_syn)
    plot_confusion_matrix(
        y_test_syn,
        y_pred,
        classes=classes,
        title=f"{name}: Train on Real → Test on Synthetic",
        filename=f"cm_{name.replace(' ','_')}_real2syn.png"
    )
```

Fig. 4. Code snippet used to generate confusion matrices for real and synthetic dataset evaluation.

The resulting confusion matrices for both Logistic Regression and Random Forest classifiers, trained and evaluated on real and synthetic datasets, are presented in Fig. 5 and Fig. 6. These matrices provide a clear visualization of correct versus incorrect predictions for each class, allowing detailed inspection of model behavior and cross-domain generalization (synthetic→real and real→synthetic). They serve as direct evidence of the evaluation phase and illustrates how well predictive patterns are preserved in the synthetic data.

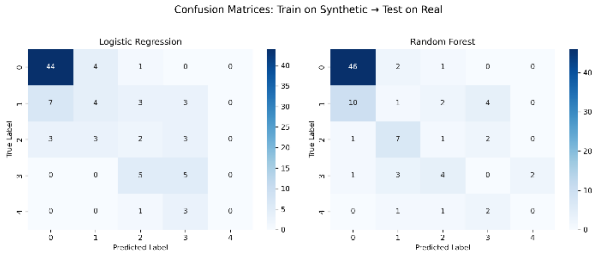

Fig. 5. Confusion matrices for Logistic Regression and Random Forest classifiers (TSTR).

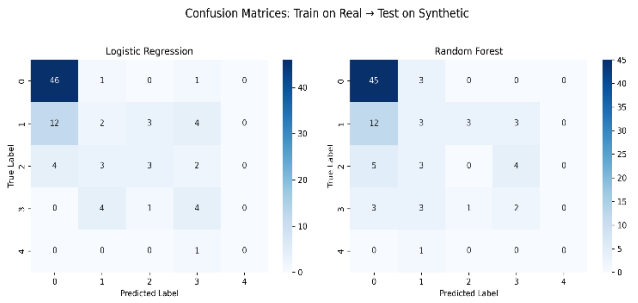

Fig. 6. Confusion matrices for Logistic Regression and Random Forest classifiers (TRTS).

E. Feature Importance Comparison

Random Forest feature importance was computed independently for the real and synthetic datasets to evaluate whether CTGAN preserved the relative influence of key predictive features. Fig. 7 and Fig. 8 depict the comparison, allowing assessment of how closely the synthetic data replicates feature-level patterns learned from the real dataset.

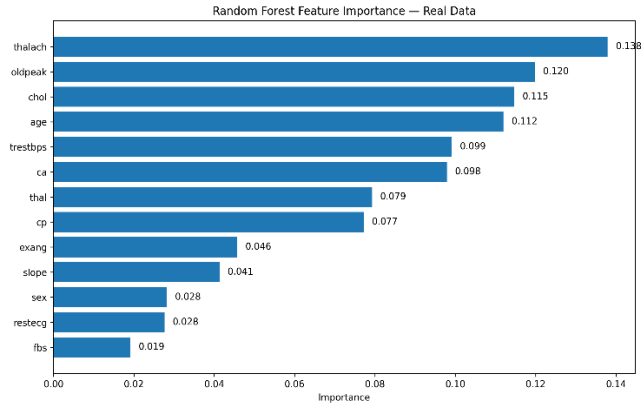

Fig. 7. Random forest feature importance on real data.

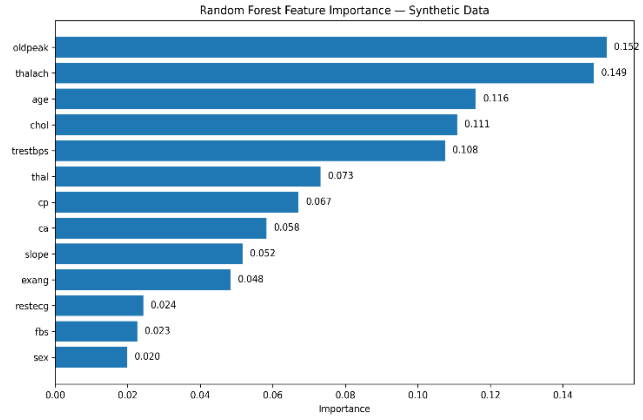

Fig. 8. Random forest feature importance on synthetic data.

F. Dimensionality Reduction Analysis

Table VII summarizes quantitative distance metrics derived from PCA and t-SNE embeddings, while Fig. 9 and Fig. 10 present scatter plots illustrating the spatial overlap between synthetic and real samples in the reduced feature space.

TABLE VII DIMENSIONALITY REDUCTION-BASED SIMILARITY BETWEEN REAL AND SYNTHETIC DATASETS

Method	Distance
PCA	1.013
t-SNE	6.952
Avg min dist (synthetic → real)	2.022

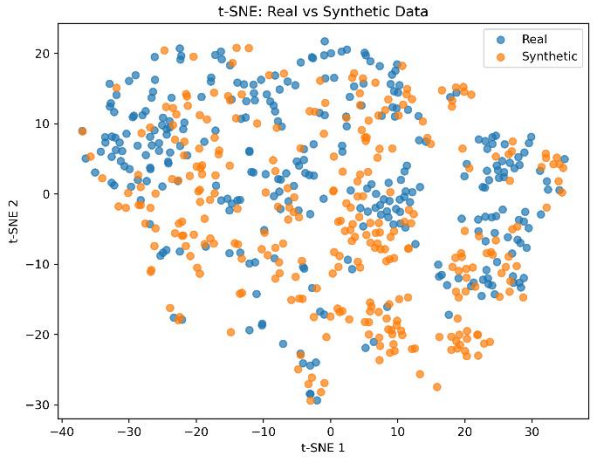

Fig. 9. t-SNE scatter plots showing synthetic and real samples in feature space.

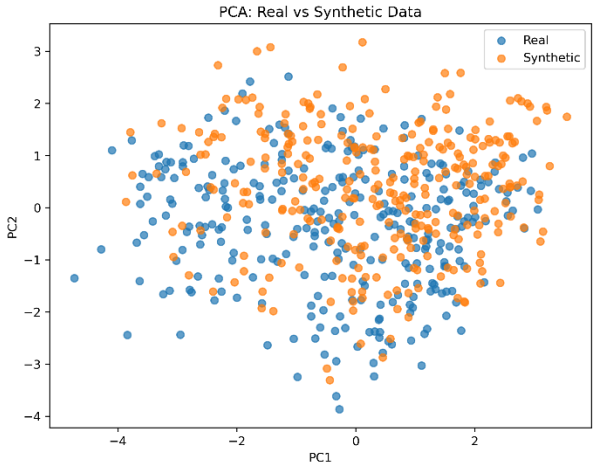

Fig. 10. PCA scatter plots showing synthetic and real samples in feature space.

V. Discussion

The present study evaluated how effectively CTGAN-generated synthetic data replicates the statistical structure, feature relationships, and predictive utility of the Cleveland Heart Disease dataset. Findings indicate that, although CTGAN preserves several overarching patterns, particularly in continuous variables, it does not fully reproduce finer-grained distributional characteristics or certain categorical feature relationships. These discrepancies affect downstream model performance and indicate areas where synthetic data diverges most from clinical signals.

Numerical features such as age, trestbps, and oldpeak exhibited moderate to strong fidelity, with percentage

differences in means generally below 6%, whereas features such as cholesterol (chol) demonstrated weaker fidelity. Categorical features, including fbs, slope, ca, and thal, exhibited higher variability, reflecting the challenges of modeling sparse or complex discrete distributions in tabular GANs. Similar challenges in preserving categorical distributions and marginal fidelity have been reported in CTGAN-based studies on clinical tabular data in papers 44 and 45, suggesting that these limitations are not dataset-specific but intrinsic to current GAN-based tabular synthesis approaches.

Correlation analyses further indicated that, although significant pairwise relationships were preserved, some feature interactions deviated in synthetic data, particularly those involving high-variance or multi-level categorical variables. This finding aligns with prior comparative evaluations, which show that no single synthetic data generator consistently preserves all inter-feature dependencies across domains and datasets in papers 47 and 48. Importantly, this contextualizes the observed correlation drift as a known trade-off rather than an implementation flaw.

Predictive evaluations demonstrated that classifiers trained on synthetic data partially generalized to real data, and vice versa. Logistic Regression and Random Forest classifiers achieved moderate accuracy (≈56–61%) when cross-tested between real and synthetic datasets. Although predictive performance did not fully match that of models trained and tested on the original dataset, these results indicate that synthetic data retains essential structure relevant to downstream modeling. Similar observations have been reported in clinical benchmarking studies, where high fidelity does not always translate into equivalent downstream performance, as reported in paper 44. Confusion matrices highlight that synthetic data remains somewhat distinguishable from real data, reflecting subtle discrepancies that warrant further improvement.

PCA and t-SNE analyses suggests that synthetic samples occupy a feature space broadly similar to real data, although differences remain. Random Forest feature importance comparisons indicate that the most influential predictors, thalach, oldpeak, age, and cholesterol (chol), are consistently identified in both datasets. This consistency suggests that CTGAN preserves dominant predictive signals even when fine-grained distributions differ, supporting its utility for exploratory analysis and preliminary modeling tasks.

Despite these encouraging results, several limitations were noted. Discrete or sparse categorical features were reproduced less faithfully, resulting in reduced predictive alignment for some combinations of features. Additionally, multi-class ROC AUC metrics were not consistently computable, limiting certain aspects of discrimination evaluation. These challenges reflect broader issues reported in synthetic healthcare data evaluation, where fidelity and utility must be balanced and remain sensitive to feature type and dataset complexity in paper 49.

These limitations also highlight potential avenues for future improvement. Hybrid architectures, conditional GAN variants, and feature-specific regularization approaches may enhance fidelity for challenging features. Integrating domain knowledge into the generation process may further reduce discrepancies, particularly for clinically relevant variables.

This study provides a systematic, multi-faceted evaluation of CTGAN-generated synthetic heart disease data, encompassing feature-level fidelity, correlation preservation, predictive utility, and global feature-space similarity. The findings provide practical guidance for researchers aiming to leverage synthetic data for privacy-preserving modeling or for augmenting limited clinical datasets. By situating the observed strengths and limitations within established comparative findings in papers 44, 45, 47, 48, and 49, this work offers practical guidance for researchers considering synthetic tabular data for privacy-preserving modeling or data augmentation, while maintaining transparency about current methodological constraints.

VI. Conclusion

CTGAN demonstrated potential for generating synthetic tabular data suitable for preliminary modeling and exploratory analysis in privacy-sensitive contexts. The synthetic dataset retained many essential characteristics of the original data and supported predictive performance within a comparable range, indicating its usefulness for low-risk, non-diagnostic applications. Despite observable limitations, the results suggest that CTGAN can serve as a supportive tool for data augmentation, early-stage model development, and constrained-data environments.

While certain categorical features and multi-feature interactions exhibited lower fidelity, the synthetic data preserved sufficient structure to support exploratory modeling, feature importance analysis, and preliminary predictive tasks, cross-dataset evaluations using Logistic Regression and Random Forest indicated that synthetic samples retained meaningful relationships with the target variable, offering practical utility when access to real clinical data is limited or restricted.

Importantly, this work presents a transparent and reproducible framework for evaluating synthetic tabular data in healthcare that integrates feature-level fidelity analysis, correlation preservation, predictive utility assessment, and dimensionality reduction techniques. The findings highlight the role of CTGAN-generated data as a complementary resource for dataset augmentation, algorithm benchmarking, and exploratory research under privacy constraints.

Future work may further enhance data fidelity through hybrid GAN architectures, domain-informed conditional generation, and targeted improvements for sparse or complex categorical features. Overall, this study demonstrates that GAN-based synthetic data can provide a practical balance between data utility and privacy considerations for healthcare analytics.

Acknowledgment

A sincere thanks to the Center of Research and Innovation Management (CREIM) at UniSZA for facilitating this publication. Heartfelt appreciation to the UniSZA team, especially Miss Nur Laila Najwa Josdi, PhD candidate, for her contributions to the project's technical setup, and to Dr. Wan Aezwani Wan Abu Bakar, her supervisor, for valuable guidance, conceptual input, and meticulous review of the manuscript. Special acknowledgment to Assoc. Prof. Ts. Dr. Mustafa Man from UMT for his assistance with network integration and helpful recommendations during the review process.

Additionally, the team expresses gratitude to the international collaborator and the insights provided by Dr. Evizal Abdul Kadir, research fellow at Universitas Islam Riau, Indonesia, which significantly enhanced this work.

REFERENCES

[1] F. Mesquita, G. Marques, "An explainable machine learning approach for automated medical decision support of heart disease," *Data & Knowledge Engineering*, *153*, 102339, 2024, https://doi.org/10.1016/j.datak.2024.102339

[2] G. Gulati, J. Upshaw, B. S. Wessler, R. J. Brazil, J. Nelson, D. van Klaveren, C. M. Lundquist, J. G. Park, H. McGinnes, E. W. Steyerberg, B. Van Calster, and D. M. Kent, "Generalizability of Cardiovascular Disease Clinical Prediction Models: 158 Independent External Validations of 104 Unique Models," *Circulation. Cardiovascular quality and outcomes*, *15*(4), e008487, 2022. https://doi.org/10.1161/CIRCOUTCOMES.121.008487

[3] R. Hagan, C. J. Gillan, and F. Mallett, "Comparison of machine learning methods for the classification of cardiovascular disease," *Informatics in Medicine Unlocked*, *24*, 100606, 2021. https://doi.org/10.1016/j.imu.2021.100606

[4] R. Kumar, S. Garg, R. Kaur, M. G. M. Johar, S. Singh, S. V. Menon, P. Kumar, A. M. Hadi, S. A. Hasson, and J. Lozanović, "A comprehensive review of machine learning for heart disease prediction: challenges, trends, ethical considerations, and future directions," *Frontiers in artificial intelligence*, *8*, 1583459, 2025. https://doi.org/10.3389/frai.2025.1583459

[5] T. N. Arvanitis, S. White, S. Harrison, R. Chaplin, and G. Despotou, "A method for machine learning generation of realistic synthetic datasets for validating healthcare applications," *Health Informatics Journal*, *28*(2), 14604582221077000, 2022. https://doi.org/10.1177/14604582221077000

[6] W. Wang and T. W. Pai, "Enhancing Small Tabular Clinical Trial Dataset through Hybrid Data Augmentation: Combining SMOTE and WCGAN-GP," *Data*, *8*(9), 135. 2023. https://doi.org/10.3390/data8090135

[7] S. N. Pasha, D. Ramesh, S. Mohmmad, and A. Harshavardhan, "Cardiovascular disease prediction using deep learning techniques," In *IOP conference series: materials science and engineering* (Vol. 981, No. 2, p. 022006), 2020, IOP Publishing. https://doi.org/10.1088/1757-899X/981/2/022006

[8] A. Tiwari, A. Chugh, and A. Sharma, "Ensemble framework for cardiovascular disease prediction," *Computers in Biology and Medicine*, *146*, 105624, 2022. https://doi.org/10.1016/j.compbiomed.2022.105624

[9] M. Çakmak, "*Heart disease classification using Random Forest machine learning*," 4th International Artificial Intelligence and Data Science Congress, İzmir, Turkey. Published in *All Science Proceedings*. 2024.

[10] M. D. Teja, and G. M. Rayalu, "Optimizing heart disease diagnosis with advanced machine learning models: a comparison of predictive performance," *BMC Cardiovasc Disord* **25**, 212, 2025. https://doi.org/10.1186/s12872-025-04627-6

[11] D. Zhang, Y. Chen, Y. Chen, S. Ye, W. Cai, J. Jiang, and M. Chen, "Heart Disease Prediction Based on the Embedded Feature Selection Method and Deep Neural Network," *Journal of Healthcare Engineering*, *2021*(1), 6260022. https://doi.org/10.1155/2021/6260022

[12] S. M. Vincent Paul, S. Balasubramaniam, P. Panchatcharam, P. Malarvizhi Kumar, and A. Mubarakali, "Intelligent framework for prediction of heart disease using deep learning," *Arabian Journal for Science and Engineering*, *47*(2), 2159-2169, 2022. https://doi.org/10.1007/s13369-021-06058-9

[13] M. K. Alnajjar and S. S. Abu-Naser, "*Heart sounds analysis and classification for cardiovascular diseases diagnosis using deep learning*," International Journal of Academic Engineering Research (IJAER), 6(1), 7–23, 2022.

[14] L. Brunese, F. Martinelli, F. Mercaldo, and A. Santone, "Deep learning for heart disease detection through cardiac sounds," *Procedia Computer Science*, *176*, 2202-2211, 2020. https://doi.org/10.1016/j.procs.2020.09.257

[15] F. Ali, S. El-Sappagh, S. R. Islam, D. Kwak, A. Ali, A. M. Imran, and K. S. Kwak, "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion," *Information Fusion*, *63*, 208-222, 2020. https://doi.org/10.1016/j.inffus.2020.06.008

[16] M. M. R. Khan Mamun and T. Elfouly, "Detection of cardiovascular disease from clinical parameters using a one-dimensional convolutional neural network," *Bioengineering*, *10*(7), 796, 2023. https://doi.org/10.3390/bioengineering10070796

[17] M. Dorraki, Z. Liao, D. Abbott, P. J. Psaltis, E. Baker, N. Bidargaddi, H. R. Wardill, A. van den Hengel, J. Narula, and J. W. Verjans, "Improving Cardiovascular Disease Prediction With Machine Learning Using Mental Health Data: A Prospective UK Biobank Study," JACC. Advances, 3(9), 101180, 2024. https://doi.org/10.1016/j.jacadv.2024.101180

[18] T. Vu, Y. Kokubo, M. Inoue, M. Yamamoto, A. Mohsen, A. Martin-Morales, R. Dawadi, T. Inoue, J. T. Tay, M. Yoshizaki, N. Watanabe, Y. Kuriya, C. Matsumoto, A. Arafa, Y. M. Nakao, Y. Kato, M. Teramoto, and M. Araki, "Machine Learning Model for Predicting Coronary Heart Disease Risk: Development and Validation Using Insights From a Japanese Population-Based Study," JMIR cardio, 9, e68066, 2025. https://doi.org/10.2196/68066

[19] Z. Alireza, M. Maleeha, M. Kaikkonen, and V. Fortino, "Enhancing prediction accuracy of coronary artery disease through machine learning-driven genomic variant selection," *Journal of Translational Medicine*, *22*(1), 356, 2024. https://doi.org/10.1186/s12967-024-05090-1

[20] H. Sadr, A. Salari, M. T. Ashoobi, and M. Nazari, "Cardiovascular disease diagnosis: a holistic approach using the integration of machine learning and deep learning models," *European Journal of Medical Research*, *29*(1), p.455, 2024. https://doi.org/10.1186/s40001-024-02044-7

[21] Y. Cai, Y.Q. Cai, L. Y. Tang, Y. H. Wang, M. Gong, T. C. Jing, H. J. Li, J. Li-Ling, W. Hu, Z. Yin, and D. X. Gong, "Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: a systematic review," *BMC medicine*, *22*(1), p.56, 2024. https://doi.org/10.1186/s12916-024-03273-7

[22] Z. Zhao, A. Kunar, R. Birke, H. Van der Scheer, and L. Y. Chen, "CTAB-GAN+: enhancing tabular data synthesis," *Frontiers in Big Data*, *6*, 1296508, 2024. https://doi.org/10.3389/fdata.2023.1296508

[23] H. H. Rashidi, S. Albahra, B. P. Rubin, and B. Hu, "A novel and fully automated platform for synthetic tabular data generation and validation," *Scientific Reports*, *14*(1), p.23312, 2024. https://doi.org/10.1038/s41598-024-73608-0

[24] T. Hyrup, A. D. Lautrup, A. Zimek, and P. Schneider-Kamp, "A systematic review of privacy-preserving techniques for synthetic tabular health data," *Discover Data*, *3*(1), pp.1-32, 2025. https://doi.org/10.1007/s44248-025-00022-w

[25] H. Y. J. Kang, M. Ko, and K. S. Ryu, "Tabular transformer generative adversarial network for heterogeneous distribution in healthcare," *Scientific Reports*, *15*(1), 10254, 2025. https://doi.org/10.1038/s41598-025-93077-3

[26] Z. Zhao, R. Birke, and L. Y. Chen, "FCT-GAN: enhancing global correlation of table synthesis via Fourier transform," In *Proceedings of the 32nd ACM international conference on information and knowledge management* (pp. 4450-4454), 2023. https://doi.org/10.1145/3583780.3615202

[27] M. Hernandez, P. A. Osorio-Marulanda, M. Catalina, L. Loinaz, G. Epelde, and N. Aginako, "Comprehensive evaluation framework for synthetic tabular data in health: fidelity, utility, and privacy analysis of generative models with and without privacy guarantees," *Frontiers in Digital Health*, *7*, 1576290, 2025. https://doi.org/10.3389/fdgth.2025.1576290

[28] M. Abedi, L. Hempel, S. Sadeghi, and T. Kirsten, "GAN-Based Approaches for Generating Structured Data in the Medical Domain, " *Applied Sciences*, *12*(14), p. 7075, 2025. https://doi.org/10.3390/app12147075

[29] A. X. Wang, S. S. Chukova, C. R. Simpson, and B. P. Nguyen, "Challenges and opportunities of generative models on tabular data," *Applied Soft Computing*, *166*, 112223, 2024. https://doi.org/10.1016/j.asoc.2024.112223

[30] H. A. Ahmed, J. A. Nepomuceno, B. Vega-Márquez, and I. A. Nepomuceno-Chamorro, "Synthetic Data Generation for Healthcare: Exploring Generative Adversarial Networks Variants for Medical Tabular Data," *International Journal of Data Science and Analytics*, pp.1-16, 2025. https://doi.org/10.1007/s41060-025-00816-w

[31] I. N. M. Adiputra, and P. Wanchai, "CTGAN-ENN: a tabular GAN-based hybrid sampling method for imbalanced and overlapped data in customer churn prediction," *J Big Data* **11**, 121, 2024. https://doi.org/10.1186/s40537-024-00982-x

[32] R. Ramachandranpillai, M. F. Sikder, D. Bergström, and F. Heintz, "Bt-GAN: generating fair synthetic healthdata via bias-transforming generative adversarial networks," *Journal of Artificial Intelligence Research*, *79*, 1313-1341, 2024. https://doi.org/10.1613/jair.1.15317

[33] P. Myles, C. Mitchell, E. Redrup Hill, L. Foschini, and Z. Wang, "High-fidelity synthetic patient data applications and privacy considerations," *Journal of Data Protection & Privacy*, *6*(4), 344-354, 2024. https://doi.org/10.69554/LQOM5698

[34] R. Nasimov, N. Nasimova, S. Mirzakhalilov, G. Tokdemir, M. Rizwan, A. Abdusalomov, Y. I. Cho, "GAN-Based Novel Approach for Generating Synthetic Medical Tabular Data," *Bioengineering*, 11(12):1288, 2024. https://doi.org/10.3390/bioengineering11121288\

[35] C. Yan, Z. Zhang, S. Nyemba, and Z. Li, "Generating synthetic electronic health record data using generative adversarial networks: Tutorial," *Jmir Ai*, *3*, e52615, 2024. https://doi.org/10.2196/52615

[36] E. Wang, K. Mott, H. Zhang, S. Gazit, G. Chodick, and M. Burcu," Validation Assessment of Privacy-Preserving Synthetic Electronic Health Record Data: Comparison of Original Versus Synthetic Data on Real-World COVID-19 Vaccine Effectiveness. *Pharmacoepidemiology and drug safety*, *33*(10), e70019, 2024. https://doi.org/10.1002/pds.70019

[37] C. Sun, J. van Soest, and M. Dumontier, "Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy," *Journal of biomedical informatics*, *143*, 104404, 2023. https://doi.org/10.1016/j.jbi.2023.104404

[38] A. Goncalves, P. Ray, B. Soper, J. Stevens, L. Coyle, and A. P. Sales, "Generation and evaluation of synthetic patient data," *BMC medical research methodology*, *20*(1), p.108, 2020. https://doi.org/10.1186/s12874-020-00977-1

[39] I. Isasa, M. Hernandez, G. Epelde, F. Londoño, A. Beristain, X. Larrea, A. Alberdi, P. Bamidis, and E. Konstantinidis, "Comparative assessment of synthetic time series generation approaches in healthcare: leveraging patient metadata for accurate data synthesis," *BMC Medical Informatics and Decision Making*, *24*(1), p.27, 2024. https://doi.org/10.1186/s12911-024-02427-0

[40] J. L. Achterberg, M. R. Haas, and M.R. Spruit, "On the evaluation of synthetic longitudinal electronic health records," *BMC Med Res Methodol* 24, 181, 2024. https://doi.org/10.1186/s12874-024-02304-4

[41] C. Umesh, M. Mahendra, S. Bej, O. Wolkenhauer, and M. Wolfien, "Challenges and applications in generative AI for clinical tabular data in physiology," *Pflügers Archiv-European Journal of Physiology*, *477*(4), pp.531-542, 2025. https://doi.org/10.1007/s00424-024-03024-w

[42] G. Santangelo, G. Nicora, R. Bellazzi, and A. Dagliati, "How good is your synthetic data? SynthRO, a dashboard to evaluate and benchmark synthetic tabular data," *BMC Medical Informatics and Decision Making*, *25*(1), p.89, 2025. https://doi.org/10.1186/s12911-024-02731-9

[43] X. Chen, Z. Wu, X. Shi, H. Cho, and B. Mukherjee, "Generating synthetic electronic health record data: a methodological scoping review with benchmarking on phenotype data and open-source software," *Journal of the American Medical Informatics Association*, *32*(7), 1227-1240, 2025. https://doi.org/10.1093/jamia/ocaf082

[44] H. Y. J. Kang, E. Batbaatar, D. W. Choi, K. S. Choi, M. Ko, and K. S. Ryu, "Synthetic tabular data based on generative adversarial networks in health care: generation and validation using the divide-and-conquer strategy," *JMIR Medical Informatics*, *11*, e47859, 2023. https://doi.org/10.2196/47859

[45] M. Alqulaity and P. Yang, "Enhanced conditional GAN for high-quality synthetic tabular data generation in mobile-based cardiovascular healthcare," *Sensors*, *24*(23), 7673, 2024. https://doi.org/10.3390/s24237673

[46] I. Akiya, T. Ishihara, and K. Yamamoto, "Comparison of synthetic data generation techniques for control group survival data in oncology clinical trials: simulation study," *JMIR medical informatics*, *12*(1), e55118, 2024. https://doi.org/10.2196/55118

[47] C. Yan, Y. Yan, Z. Wan, Z. Zhang, L. Omberg, J. Guinney, S. D. Mooney, and B. A. Malin, "A multifaceted benchmarking of synthetic electronic health record generation models. *Nature Communications*, *13*(1), p.7609, 2022. https://doi.org/10.1038/s41467-022-35295-1

[48] M. Miletic and M. Sariyar, "Challenges of Using Synthetic Data Generation Methods for Tabular Microdata," *Applied Sciences*, *14*(14), 5975, 2024. https://doi.org/10.3390/app14145975

[49] V. B. Vallevik, A. Babic, S. E. Marshall, S. Elvatun, H. M. Brøgger, S. Alagaratnam, and J. F. Nygård, "Can I trust my fake data–a comprehensive quality assessment framework for synthetic tabular data in healthcare," *International Journal of Medical Informatics*, *185*, 105413, 2024. https://doi.org/10.1016/j.ijmedinf.2024.105413

[50] K. El Emam, L. Mosquera, X. Fang, and A. El-Hussuna, "Utility Metrics for Evaluating Synthetic Health Data Generation Methods: Validation Study," *JMIR Medical Informatics*, *10*(4), e35734, 2022. https://doi.org/10.2196/35734

[51] T. Afonja, D. Chen, and M. Fritz, "MargCTGAN: A 'Marginally' Better CTGAN for the Low Sample Regime," In: Köthe, U., Rother, C. (eds) Pattern Recognition. DAGM GCPR 2023. Lecture Notes in Computer Science, vol 14264. Springer, Cham, 2024. https://doi.org/10.1007/978-3-031-54605-1_34

[52] M. L. Fang, D. S. Dhami, and K. Kersting, "Dp-ctgan: Differentially private medical data generation using CTGANS," In *International conference on artificial intelligence in medicine* (pp. 178-188). Cham: Springer International Publishing, 2022. https://doi.org/10.1007/978-3-031-09342-5_17

[53] S. Singh, K. Kayathwal, H. Wadhwa, and G. Dhama, "MeTGAN: Memory Efficient Tabular GAN for High Cardinality Categorical Datasets," In: Mantoro, T., Lee, M., Ayu, M.A., Wong, K.W., Hidayanto, A.N. (eds) Neural Information Processing. ICONIP 2021. Communications in Computer and Information Science, vol 1517. Springer, Cham, 2021. https://doi.org/10.1007/978-3-030-92310-5_60

[54] J. Kim, J. Jeon, J. Lee, J. Hyeong, and N. Park, "Oct-gan: Neural ode-based conditional tabular gans," In *Proceedings of the Web Conference 2021*, pp. 1506-1515, 2021. https://doi.org/10.1145/3442381.3449999

[55] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "Heart Disease [Dataset]," UCI Machine Learning Repository, 1989. https://doi.org/10.24432/C52P4X