

Evaluating the Usability of Synthetic Data for Machine Learning-Based Risk Prediction in Healthcare

Nur Sena Duman Çelikel

Graduate Program in Smart Systems Engineering
Galatasaray University, Istanbul, Türkiye
nur--sena@windowslive.com

Müjde Erol Genevois

Professor, Dept. of Industrial Engineering
Galatasaray University, Istanbul, Türkiye
merol@gsu.edu.tr

Abstract

The limited accessibility of healthcare data due to privacy and data-sharing constraints poses a significant challenge for the development of reliable machine learning-based risk prediction models. In this study, the statistical fidelity and machine learning utility of synthetic tabular data generated by the Conditional Tabular Generative Adversarial Network (CTGAN) are systematically evaluated. The analysis is conducted on three datasets with different feature structures: numerical (Hepatitis C), categorical (Diabetes), and mixed-type (cardiovascular disease). The evaluation framework includes three experimental scenarios: Train on Real Test on Real (TRTR), Train on Synthetic Test on Real (TSTR), and Train on Synthetic Test on Synthetic (TSTS). A diverse set of machine learning models is employed, including Logistic Regression, K-Nearest Neighbors, Support Vector Machines, Random Forest, XGBoost, and LightGBM. Model performance is assessed with particular emphasis on metrics sensitive to class imbalance. The results show that the behavior of CTGAN is strongly dependent on the underlying feature structure of the data. In numerical datasets, macroscopic distributions are largely preserved, while notable degradation is observed in representing decision boundaries of rare classes. In categorical datasets, the model shows difficulty in capturing complex multivariate joint distributions, leading to substantial performance drops, particularly for tree-based methods. In mixed-feature datasets, synthetic data appears to introduce a denoising effect that partially stabilizes learning, but at the cost of reduced boundary precision.

Keywords

Risk Assessment, Artificial Intelligence (AI), Healthcare Application, Machine Learning Applications, Data Analytics

1. Introduction

Artificial intelligence and machine learning techniques have shown significant promise in advancing the early detection and risk prediction of diseases (Dong, et al. 2024; Zheng, et al. 2024; Cedolin, et al. 2024). These methods leverage heterogeneous clinical features to uncover complex relationships between patient characteristics and disease outcomes. Generally, the clinical data used in these predictive models fall into three common categories: objective physiological measurements, behavioral lifestyle factors, and complex mixed datasets combining both (Klassen, et al. 2012). While clinical measurements provide direct quantifiable information on a patient's physiological state, lifestyle factors encompass behaviors such as diet and physical activity. Together, they form the complex tabular datasets commonly used in predictive clinical studies (Lévesque, et al. 2017).

However, developing reliable predictive models is severely bottlenecked by the scarcity of high-quality datasets. Strict privacy and data protection regulations, such as GDPR and HIPAA, limit health data sharing and large-scale data collection. This results in clinical datasets that are often small, imbalanced, or restricted in diversity, which negatively affects the performance and generalization ability of machine learning models.

To address these limitations, this study investigates the effectiveness of Generative Adversarial Network (GAN) based synthetic data generation across healthcare datasets that differ in their underlying feature types. By generating synthetic data and evaluating its impact on downstream classification models, we provide a comparative analysis of how GAN generated data performs across different clinical structures. To ensure reproducibility and transparent benchmarking, the framework is evaluated using publicly available datasets, including Hepatitis C, Cardiovascular Disease, and Diabetes.

2. Literature Review

The application of data mining and predictive modeling in healthcare has expanded rapidly to support risk assessment and clinical decision making (Krittawong, et al. 2020; Zheng, et al. 2024; Dong, et al. 2024). Despite this progress, researchers consistently highlight data scarcity and class imbalance as primary challenges in medical applications (Sidey Gibbons and Sidey Gibbons 2019).

To address these data limitations, various augmentation techniques have been proposed. Traditional oversampling methods, such as SMOTE and ADASYN, have been widely used to handle class imbalance in machine learning tasks (Han, et al. 2005; He, et al. 2008; Douzas, et al. 2018). While effective in certain scenarios, these heuristic approaches often struggle to capture the complex, non linear relationships inherent in structured medical records.

Recently, Generative Adversarial Networks (GANs) have emerged as a powerful alternative for augmenting tabular datasets. By learning the underlying multidimensional distribution of existing datasets, GANs can generate realistic and contextually consistent synthetic samples. In the healthcare domain, GAN generated synthetic Electronic Health Records (EHRs) have been shown to preserve important statistical properties of real datasets while mitigating privacy constraints and data sparsity (Yoon, et al. 2020; Li, et al. 2023; Chen, et al. 2025; Beaulieu Jones, et al. 2019).

Recent studies have further demonstrated that advanced GAN-based tabular generation methods, such as CTGAN (Xu et al. 2019), ConvGen (Schultz et al. 2024), and CTAB-GAN+ (Zhao et al. 2024), can significantly improve downstream classification performance, especially when available data are limited or highly imbalanced. Building on these developments, our study addresses a specific gap in the literature by systematically evaluating how these generative models perform across fundamentally different tabular data architectures, specifically comparing numerical, categorical, and mixed clinical data structures.

3. Methods

The proposed experimental framework evaluates the machine learning utility of synthetic clinical data generated by CTGAN across three distinct healthcare datasets. The workflow (illustrated in Figure 1) consists of data preprocessing, synthetic data generation, and comparative predictive evaluation under multiple testing scenarios.

3.1 Synthetic Data Generation Framework: CTGAN

This study employs the Conditional Tabular Generative Adversarial Network (CTGAN) as the primary framework for synthetic tabular data generation. CTGAN is specifically designed to address tabular data challenges, including complex continuous distributions and severe class imbalance in discrete features (Xu, et al. 2019). Synthetic samples were generated in quantities equal to their corresponding real datasets using the Synthetic Data Vault library (SDV Team 2026).

CTGAN utilizes a variational Gaussian mixture model to estimate and normalize the underlying modes of continuous features. Additionally, a conditional generation mechanism enforces the generation of specific categories during training, ensuring minority classes are adequately represented and preventing mode collapse. Training durations were adjusted according to dataset scale: 500 epochs for the limited HCV dataset, and 100 epochs for the Diabetes and Cardiovascular datasets to balance learning capacity and computational efficiency. For each dataset (Cardiovascular, Hepatitis C, and Diabetes), an independent CTGAN model is trained using the corresponding preprocessed real data.

The objective of the training process is to learn the joint probability distribution of the real dataset D_{real} and to generate a synthetic dataset D_{syn} that closely resembles D_{real} while preserving the original feature space dimensionality.

3.2 Machine Learning Models for Utility Evaluation

To evaluate the predictive utility of the CTGAN generated data, a diverse ensemble of machine learning classifiers is employed. This diversity ensures that the synthetic data preserves both global linear trends and intricate non linear decision boundaries present in the original datasets. The classifiers, implemented using *scikit learn*, *XGBoost*, and *LightGBM* frameworks, include:

- **Logistic Regression (LR):** Serves as a linear baseline to assess basic class separability.
- **K Nearest Neighbors (KNN):** Evaluates the preservation of local data manifolds and spatial distribution.
- **Support Vector Machine (SVM):** Utilizing a Radial Basis Function kernel to test the replication of complex clinical geometries.
- **Random Forest (RF):** Determines if hierarchical feature interactions are effectively captured.
- **Gradient Boosting Models (XGBoost and LightGBM):** Assesses fine grained utility by detecting subtle patterns and residual errors.

Evaluating performance across this heterogeneous set ensures results represent a general property of the synthetic data rather than an artifact of a specific algorithm.

3.3 Dataset Selection and Preprocessing

To ensure a robust evaluation, three healthcare datasets with varying scales and feature compositions are utilized. Each dataset is formulated as a binary classification problem to distinguish between diseased and non-diseased states.

Table 1. An overview of datasets and feature composition.

Dataset	Total Samples	Healthy	Disease	Features	Numerical	Categorical
<i>HCV</i>	589	533	56	12	11	1
<i>Cardiovascular</i>	70,000	35,021	34,979	11	5	6
<i>Diabetes</i>	253,680	218,334	35,346	21	1	20

The HCV dataset (Lichtinghagen and Hoffmann 2020) consists of laboratory values from blood donors and Hepatitis C patients. While the original data includes categorical demographic variables such as gender, the preprocessing stage focused exclusively on continuous laboratory measurements, such as Albumin (ALB) and Cholesterol (CHOL). Features deemed redundant or non-numerical during the feature engineering phase were excluded to maintain a focus on purely clinical indicators within this data-scarce environment.

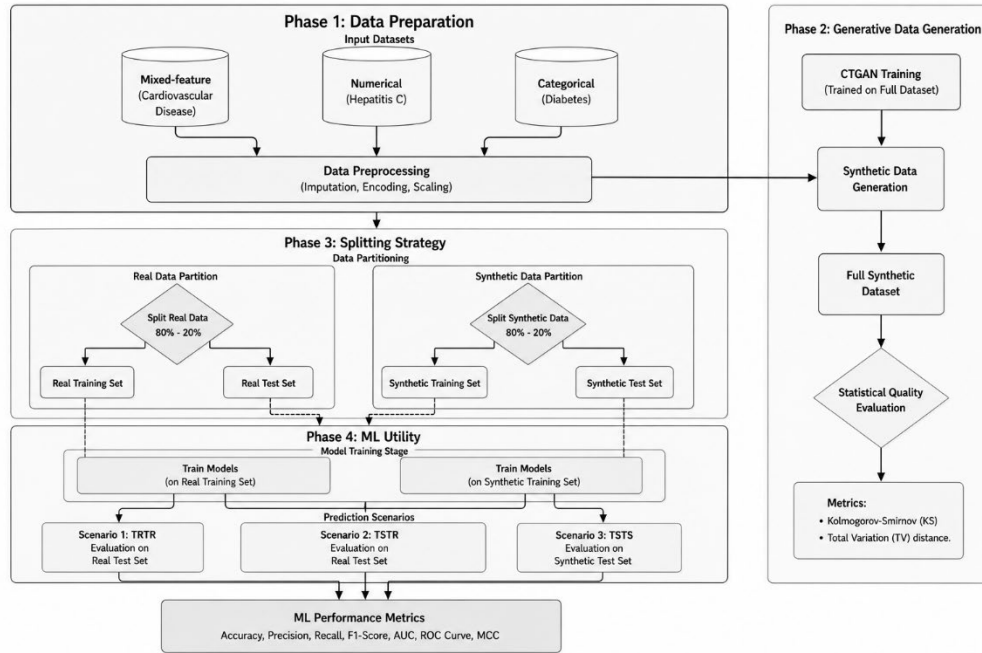


Figure 1. Methodology pipeline for real and synthetic data generation and evaluation.

The Cardiovascular Disease Data Set (Ulianova 2019) contains 70,000 patient records collected during medical examinations. This data set provides a balanced mix of lifestyle features, including numerical, ordinal, and binary variables. The target variable is binary, specifically labeling the presence or absence of cardiovascular disease based on clinical indicators recorded at the time of examination.

The Diabetes data set (Teboul 2021) is a consolidated version of the 2015 Behavioral Risk Factor Surveillance System (BRFSS) (Centers for Disease Control and Prevention (CDC) 2015) survey conducted annually by the Centers for Disease Control and Prevention. While the original BRFSS survey collects health-related risk behaviors and chronic conditions from over 400,000 Americans, this refined version focuses on over 250,000 respondents. The feature set comprises 21 variables, nearly all of which are categorical, with Body Mass Index (BMI) being the sole numerical indicator. The target variable was transformed into a binary classification problem to differentiate between healthy individuals and those diagnosed with pre-diabetes or diabetes.

The example data sets can be summarized as followed:

- **HCV (Small-scale):** Focused on continuous laboratory measurements (589 samples).
- **Cardiovascular (Medium-scale):** A mix of numerical, ordinal, and lifestyle features (70,000 samples).
- **Diabetes (Large-scale):** Primarily categorical and binary indicators derived from population-wide surveys (253,680 samples).

A summary of the datasets after cleaning and binary class mapping is presented in Table 1. All datasets were partitioned using a stratified 80/20 train-test split to preserve class distributions. Additionally, the table presents the categorical and numerical features available in the respective datasets.

3.4 Experimental Design and Evaluation Strategy

The core evaluation of this study relies on the Train-on-Synthetic, Test-on-Real (TSTR) methodology (Esteban, et al. 2017), which serves as a method for assessing the clinical utility of generative models. In the TSTR framework, the original dataset R is partitioned into a training set R_{train} and a hold-out test set R_{test} . A generative model is then trained on R_{train} to produce a corresponding synthetic dataset S .

To quantify the effectiveness of the synthetic data, two separate classification models are developed using the same architectural hyperparameters:

- 1) M_R : A baseline model trained exclusively on the real training set R_{train} .
- 2) M_S : A model trained exclusively on the synthetic dataset S .

Both models are subsequently evaluated on the same unseen real test set, R_{test} . By comparing the performance metrics of M_S and M_R on real-world data, we can measure the degree to which the synthetic data captures the underlying distribution and predictive signals of the original source. The relative discrepancy in performance is quantified using the following formula:

$$\text{Performance Drop} = \frac{\text{Perf}(M_R, R_{\text{test}}) - \text{Perf}(M_S, R_{\text{test}})}{\text{Perf}(M_R, R_{\text{test}})} \times 100\%$$

The evaluation of the synthetic data is further extended through the Train-on-Synthetic, Test-on-Synthetic (TSTS) methodology, which serves as a diagnostic tool for assessing the internal statistical cohesion and distributional fidelity of the generative process. In this framework, a classification model M_S is both trained and evaluated using independent partitions of the synthetic dataset S , effectively treating the synthetic data as its own closed-loop environment. By measuring the performance of M_S on a synthetic test set S_{test} , we can determine whether the generator has captured a consistent and learnable joint distribution rather than producing stochastic noise.

Furthermore, TSTS is critical for detecting mode collapse; a scenario where the TSTS performance significantly exceeds the TRTR baseline often indicates that the generator has overfitted to a narrow, simplified subset of the real distribution, whereas an alignment between TSTS and TRTR suggests that the synthetic data successfully replicates the complexity and class separability of the original clinical manifold (Xu, et al. 2019). This approach ultimately allows us to verify that the generated samples are internally logical and possess the necessary diversity to represent the underlying patterns of the source data.

In summary, three distinct configurations are analysed:

- **TRTR (Train on Real, Test on Real):** The gold-standard performance, representing the upper bound of the real data.
- **TSTR (Train on Synthetic, Test on Real):** Directly measures the ML Utility. A minimal gap between TRTR and TSTR indicates high fidelity of synthetic data.
- **TSTS (Train on Synthetic, Test on Synthetic):** Evaluates internal consistency and potential overfitting of the generative model.

3.5 Performance Metrics

Predictive utility is quantified by benchmarking models trained under the Train on Synthetic, Test on Real (TSTR) scenario against the gold standard Train on Real, Test on Real (TRTR) baseline. Minimal performance degradation in TSTR relative to TRTR indicates that the synthetic data effectively preserves the statistical distributions, feature dependencies, and predictive signals of the original source.

Furthermore, internal consistency is assessed by jointly analyzing Train on Synthetic, Test on Synthetic (TSTS) and TSTR results. While TSTS evaluates the structural coherence and learnability of the generated data, a substantial performance gap where TSTS markedly exceeds TSTR may indicate generative overfitting. Such a discrepancy suggests the synthetic data contains simplified patterns or artifacts that fail to generalize to real world samples. High quality synthetic data are thus characterized by TSTR performance closely approaching the TRTR benchmark, aligned with TSTS results that demonstrate reliable generalization.

To address class imbalance in the Hepatitis C and Diabetes datasets, model evaluation extends beyond accuracy to include precision, recall, F1 score, and AUC ROC. Additionally, the Matthews Correlation Coefficient (MCC) is employed as a robust metric that incorporates all confusion matrix components, providing a balanced measure of classification quality even under severe imbalance (Hicks, et al. 2022). The evaluation prioritizes these class sensitive and threshold independent metrics to ensure a rigorous assessment of predictive performance across varying clinical data structures.

3.6 Conclusion

In conclusion, the established experimental framework provides a robust and transparent pipeline for quantifying the fidelity of GAN based synthetic data. By integrating datasets of varying complexity, utilizing the specialized architectural advantages of CTGAN, and employing a diverse array of machine learning classifiers, this study ensures that the evaluation of machine learning utility and synthetic data generation is comprehensive and agnostic to ML models. The systematic interplay between the TRTR, TSTR, and TSTS scenarios allows for a granular differentiation between genuine statistical learning and potential generative overfitting. This holistic methodological approach sets a rigorous foundation for the subsequent analysis of results, aiming to determine the practical feasibility of using synthetic surrogates in high stakes clinical decision-making environments.

Table 2. Comparison of synthetic data quality scores across datasets

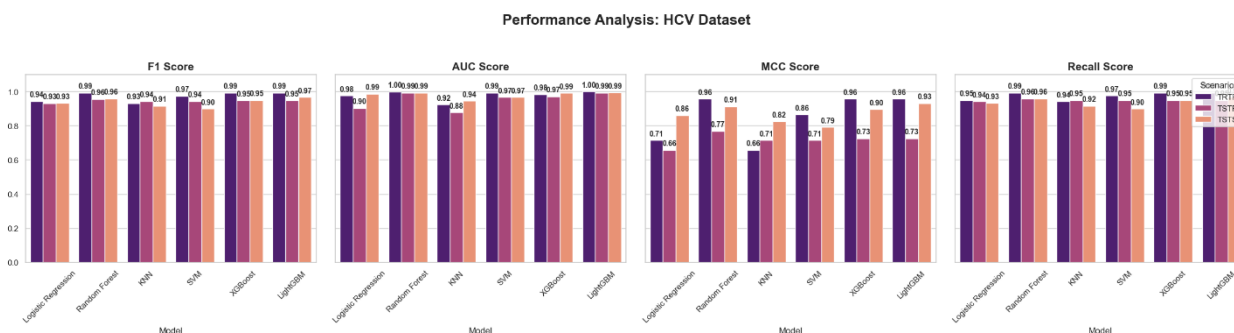
Dataset	Column Shapes (%)	Column Pair Trends (%)	Overall Score (%)
Hepatitis C	70.33	85.23	77.78
Cardiovascular	88.26	70.86	79.56
Diabetes	88.89	78.17	83.53

4. Results and Discussion

The experimental evaluation of the CTGAN-generated synthetic data is structured to systematically investigate both the statistical integrity and the practical applicability of the synthesized cohorts. This section is organized into three analytical stages: First, a Statistical Similarity Analysis is conducted to quantify the degree to which the synthetic data preserves the marginal distributions and feature correlations of the original datasets. Second, an ML utility assessment is performed using the TRTR (Train on Real, Test on Real) and TSTR (Train on Synthetic, Test on Real) paradigms to measure the predictive performance gap across different feature architectures. Finally, an in-depth analysis of General Trends and Model Robustness incorporates the TSTS (Train on Synthetic, Test on Synthetic) methodology to diagnose potential generative mode collapse and explores how specific data typologies—continuous, categorical, and mixed—uniquely constrain the fidelity of synthetic transfer across diverse machine learning algorithms.

4.1 Statistical Similarity Analysis

The statistical quality of the synthetic datasets is quantified using Column Shapes (distributional similarity) and Column Pair Trends (correlation preservation). As summarized in Table 2, the Diabetes dataset achieved the highest overall quality score (83.53%), benefiting from its large sample size and structured categorical features. The Cardiovascular dataset followed with 79.56%, showing high fidelity in marginal distributions (88.26% Column Shapes). The HCV dataset, despite its limited scale, reached a quality score of 77.78%. Notably, categorical variables such as *Sex* and the target *Category* achieved near-perfect *TVComplement* scores (a metric that evaluates the distributional overlap of discrete features, where values closer to 1 indicate higher similarity). In contrast, continuous biochemical indicators in the HCV and Cardio datasets exhibited varying *KSComplement* scores (a statistical measure comparing the cumulative distributions of numerical variables) ranging from 0.47 to 0.86, reflecting the inherent complexity of modeling multi-modal clinical distributions (Figure 2).



(a) Comprehensive scenario performance comparison across six models for the continuous numerical HCV dataset.

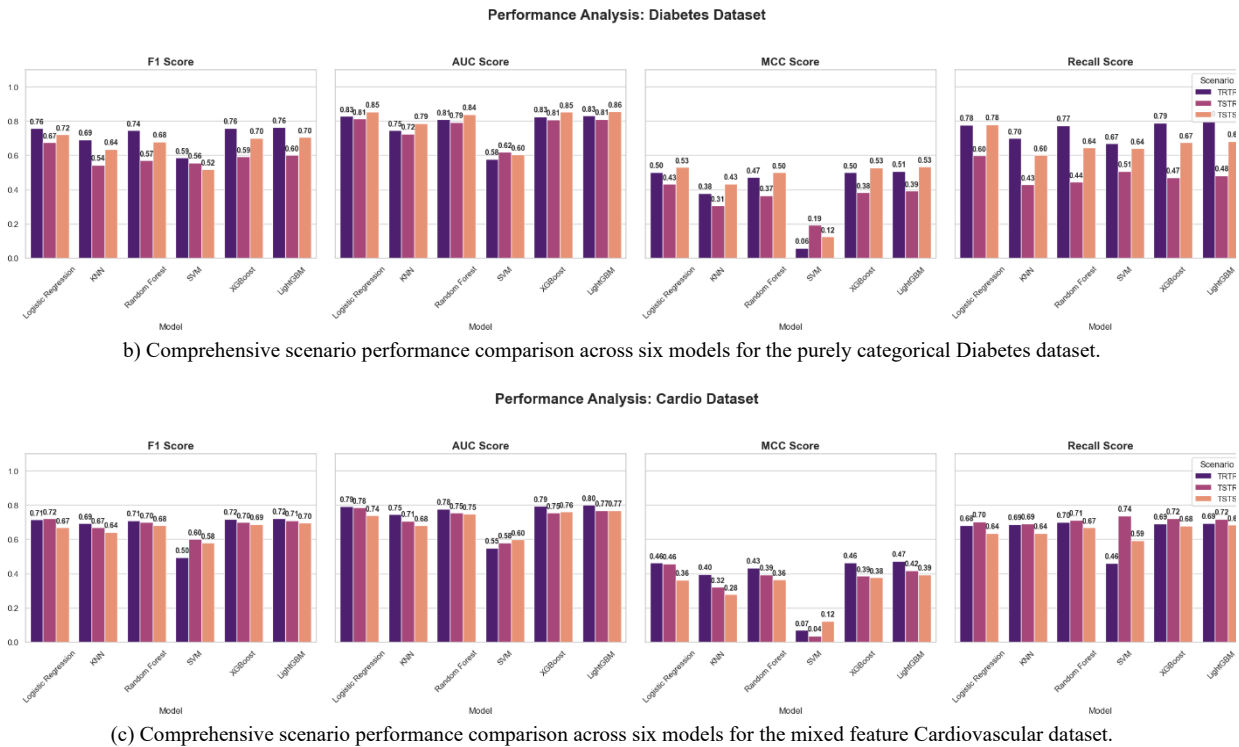


Figure 2. Complete performance evaluation of the TRTR, TSTR, and TSTS scenarios across all datasets, evaluating five key metrics simultaneously.

4.2 Machine Learning Utility Analysis

The core of the evaluation focuses on the performance gap between models trained on real data (TRTR) and those trained on synthetic data (TSTR) (Figure 3- Figure 5).

- **Diabetes Dataset (Categorical Data):** Exhibited the highest stability. Models like XGBoost and LightGBM achieved TSTR performance nearly identical to TRTR, suggesting that the synthetic data effectively captured the predictive signals required for diabetes risk assessment.
- **Cardiovascular Dataset (Mixed Data):** Showed a moderate and controlled performance degradation (4-7% in accuracy). Ensemble models (RF, XGBoost) demonstrated superior robustness, maintaining stable AUC-ROC values despite the synthetic nature of the training set.
- **HCV Dataset (Numerical Data):** Proved to be the most challenging scenario. While TRTR results were near-perfect, a significant drop was observed in TSTR, particularly for SVM and KNN. This indicates that while CTGAN preserves global statistical properties, the fine-grained decision boundaries of small-scale biochemical datasets are harder to replicate.

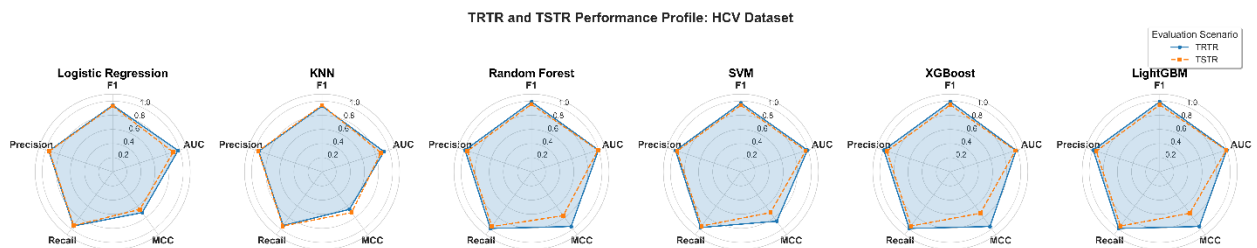


Figure 3. This radar chart compares the baseline and synthetic data performance metrics across six machine learning models for the continuous numerical HCV dataset.

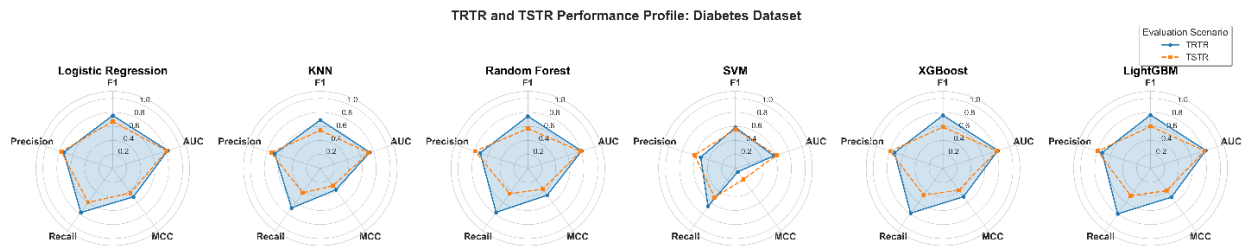


Figure 4. This radar chart compares the baseline and synthetic data performance metrics across six machine learning models for the purely categorical Diabetes dataset.

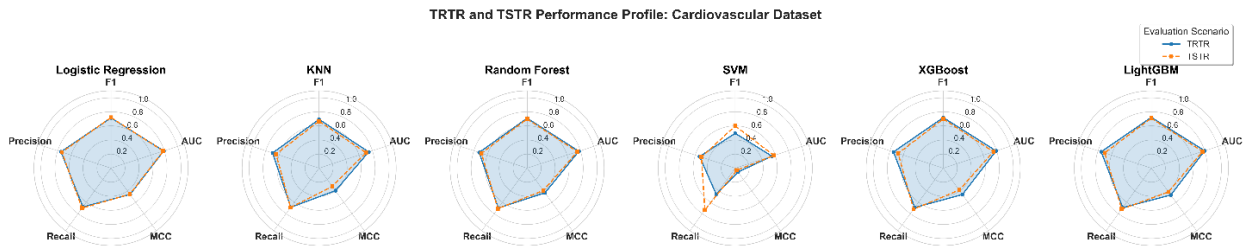


Figure 5. This radar chart compares the baseline and synthetic data performance metrics across six machine learning models for the highly noisy, mixed-feature Cardiovascular dataset.

4.3 General Trends and Model Robustness

Before evaluating synthetic utility, the baseline performance on original datasets (TRTR) was analyzed. The results indicate that each dataset's feature architecture fundamentally dictates its class separability. The HCV dataset, dominated by continuous numerical features, facilitated high separability with an average F1 score of 0.970, AUC of 0.978, and MCC of 0.850 (Figure 2a). The high MCC confirms precise decision boundaries despite class imbalance.

Table 3. Complete Multidimensional Performance Evaluation: TRTR vs. TSTR. The Δ (%) column represents the percentage degradation, where negative values denote an improvement when trained on synthetic data.

Dataset	Model	Accuracy			F1 Score			MCC			Recall			Precision		
		TRTR	TSTR	Δ (%)	TRTR	TSTR	Δ (%)	TRTR	TSTR	Δ (%)	TRTR	TSTR	Δ (%)	TRTR	TSTR	Δ (%)
Cardiovascular	KNN	0.75	0.71	5.31	0.69	0.67	3.63	0.40	0.32	19.18	0.69	0.69	-0.49	0.70	0.65	7.49
	Light-GBM	0.80	0.77	4.07	0.72	0.71	1.71	0.47	0.42	11.56	0.69	0.72	-3.73	0.75	0.70	7.00
	LR	0.79	0.79	0.88	0.72	0.72	-0.70	0.46	0.46	1.15	0.68	0.70	-3.27	0.75	0.74	1.99
	RF	0.78	0.75	2.86	0.71	0.70	1.61	0.43	0.39	9.47	0.70	0.71	-1.62	0.72	0.69	4.73
	SVM	0.55	0.58	-5.50	0.49	0.60	-21.21	0.07	0.04	45.34	0.46	0.74	-59.72	0.54	0.51	5.28
	XG-Boost	0.79	0.75	4.92	0.72	0.70	2.76	0.46	0.39	16.95	0.69	0.72	-4.23	0.75	0.68	9.33
Dia	KNN	0.75	0.72	2.82	0.69	0.54	21.45	0.38	0.31	18.76	0.70	0.43	38.55	0.69	0.74	-8.00

	Light-GBM	0.83	0.81	2.36	0.76	0.60	21.31	0.51	0.39	22.81	0.80	0.48	39.79	0.73	0.80	-9.22
	LR	0.83	0.81	1.92	0.76	0.67	10.98	0.50	0.43	13.89	0.78	0.60	23.04	0.74	0.77	-4.57
	RF	0.81	0.79	2.02	0.74	0.57	23.56	0.47	0.37	22.37	0.77	0.44	42.64	0.72	0.79	-10.54
	SVM	0.58	0.62	-7.09	0.59	0.56	5.13	0.06	0.19	-227.44	0.67	0.51	24.05	0.52	0.61	-17.84
	XG-Boost	0.83	0.81	2.32	0.76	0.59	22.06	0.50	0.38	23.15	0.79	0.47	40.43	0.73	0.80	-9.07
HCV	KNN	0.92	0.88	5.00	0.93	0.94	-1.24	0.66	0.71	-8.50	0.94	0.95	-0.90	0.94	0.95	-0.80
	Light-GBM	1.00	0.99	0.81	0.99	0.95	4.45	0.96	0.73	24.18	0.99	0.95	4.27	0.99	0.95	4.53
	LR	0.98	0.90	7.44	0.94	0.93	1.22	0.71	0.66	7.83	0.95	0.94	0.89	0.95	0.94	0.79
	RF	1.00	0.99	0.73	0.99	0.96	3.65	0.96	0.77	19.73	0.99	0.96	3.42	0.99	0.96	3.59
	SVM	0.99	0.97	2.66	0.97	0.94	3.17	0.87	0.71	17.47	0.98	0.95	2.61	0.98	0.95	2.40
	XG-Boost	0.98	0.97	1.19	0.99	0.95	4.45	0.96	0.73	24.18	0.99	0.95	4.27	0.99	0.95	4.53

In contrast, the Diabetes dataset's discrete categorical architecture resulted in moderate separability, yielding an average F1 score of 0.716 and MCC of 0.402 (Figure 2b). Here, tree-based ensembles outperformed distance-based classifiers, effectively navigating the fragmented decision spaces. The Cardiovascular dataset, featuring a mixed architecture, showed the lowest baseline performance with an average F1 of 0.675 and MCC of 0.383 (Figure 2c).

To validate generative integrity and rule out mode collapse or memorization, models were evaluated exclusively on synthetic data (TSTS). Across all datasets, TSTS metrics showed a slight reduction compared to TRTR baselines (e.g., HCV F1 decreased from 0.970 to 0.936). This reduction suggests that the generated samples are not simple direct replications of the original data, although further privacy or memorization tests would be required for a more rigorous confirmation. Notably, the synthesis process exhibited an idealization effect; the GAN captured core marginal distributions but smoothed over real-world outliers. Consequently, the synthetic datasets represent an idealized statistical environment lacking the complex edge cases inherent to authentic clinical data.

The critical evaluation of synthetic data viability lies in the Train on Synthetic, Test on Real (TSTR) scenario. The divergence between TRTR and TSTR performance reveals how different feature architectures uniquely constrain the CTGAN's ability to transfer predictive utility to real-world applications.

In the HCV dataset, CTGAN showed relatively strong performance in modeling continuous numerical distributions, which appears to offer an advantage for this dataset. This advantage is reflected in the minimal degradation of the F1 score (dropping slightly from 0.970 to 0.944), as presented in Figure 3. However, the MCC indicates a more substantial degradation (dropping from 0.850 to 0.717). Because the HCV dataset is highly imbalanced, the CTGAN model appears to have preserved the macro-level trends of the majority class more effectively than the precise, microscopic geometric cut-offs defining the rare minority classes. The models learned the generalized shape of the numerical distributions but misclassified real-world edge cases, suggesting that numerical synthesis may blur rigid class boundaries in imbalanced domains.

The transition to synthetic data was associated with a substantial loss of predictive information within the categorical Diabetes dataset, as displayed in Figure 4. The average F1 score decreased from 0.716 to 0.589. Predictive modeling on categorical data relies on rigid, multi-variable logic (e.g., specific joint distributions across several discrete categories). While the CTGAN appeared to preserve broad marginal categorical distributions (allowing simpler models like Logistic Regression to maintain utility), it showed difficulty in synthesizing the complex, multidimensional micro-patterns. Consequently, advanced algorithms such as Random Forest and XGBoost, which depend on these intricate categorical interactions experienced substantial performance degradation.

In contrast, the models demonstrated high resilience when trained on the synthetic mixed-feature dataset, which can be found on the Figure 5 presenting the comparison of Cardiovascular data set results. The TSTR F1 score (0.682) marginally exceeded the TRTR baseline (0.675), while the MCC remained statistically stable (Figure 3). This phenomenon highlights a distinct advantage of GANs in highly noisy environments: the inherent "smoothing" behavior of the generative model may have acted as a form of data regularization. By smoothing the complex overlap between continuous and categorical variables, the GAN provided the predictive models with a cleaner, denoised signal, potentially improving the apparent decision boundaries for real-world testing.

An evaluation of performance drop metrics, which we established in the previous chapter, across the three datasets reveals how different feature architectures uniquely constrain synthetic data fidelity, specifically highlighting the divergence between preservation of F1 and MCC. The obtained results and the performance drops are included in Table 3.

In the Cardiovascular data set, the mixed-feature architecture produced paradoxical degradation metrics. The dataset experienced a negative drop in F1 (-2.03% average), indicating that synthetic data was associated with an improvement in the baseline F1 score—most notably for SVM, which saw a 21.2% improvement in F1. However, this was accompanied by a substantial 17.2% degradation in MCC. This suggests that the generative model may have acted as a denoising mechanism: it smoothed over the complex overlap of continuous and categorical variables, increasing the F1 score by optimizing majority-class predictions, but appeared to reduce the precise boundary fidelity required for true, symmetric classification (as penalized by the MCC).

In the Diabetes data set, which has predominantly categorical features, the categorical architecture appeared to be more challenging for high-fidelity synthesis. It showed substantial F1-score degradation, averaging a 17.4% drop. Crucially, the breakdown was model-dependent: advanced hierarchical models (Random Forest, XGBoost, LightGBM) experienced marked F1-score degradation of 21% to 23%, while Logistic Regression degraded by only 10.9%. This suggests that while the GAN preserved broad marginal categories (sufficient for linear boundaries), it showed limitations in synthesizing the rigid, multi-variable categorical logic required by advanced decision trees.

Finally, in the HCV dataset, the continuous numerical architecture showed relatively high resilience to synthetic transfer. The average F1 drop was limited to 2.6%, suggesting that the GAN captured the smooth mathematical gradients of the feature space to some extent. However, an accompanying MCC drop of 14.1% indicates a subtle limitation: because the dataset is highly imbalanced, the generative model appeared to preserve the broad continuous distributions of the healthy majority more effectively than the sharp geometric cut-offs of the rare minority classes, leading to increased classification errors at the extreme edges of the data.

5. Conclusion

This study systematically evaluated the Machine Learning Utility and statistical fidelity of synthetic clinical data generated by the Conditional Tabular Generative Adversarial Network (CTGAN). By employing an evaluation framework encompassing Train on Real Test on Real (TRTR), Train on Synthetic Test on Real (TSTR), and Train on Synthetic Test on Synthetic (TSTS) scenarios, we assessed the transferability of predictive signals from synthetically produced data to real world clinical environments. The empirical results suggest that the efficacy of CTGAN is not universal; rather, it is closely related to the underlying feature architecture and data types of the source datasets.

The comparative analysis across three distinct healthcare datasets highlights notable qualitative differences in how generative models process feature varieties. For continuous numerical domains, as observed in the HCV dataset, CTGAN showed relatively high resilience, preserving smooth mathematical gradients and macroscopic feature distributions to some extent. However, the accompanying degradation in the MCC indicates that CTGAN may have

difficulty maintaining the sharp, microscopic geometric thresholds required to accurately classify rare minority classes in highly imbalanced numerical data.

On the other hand, predominantly categorical architectures, represented by the Diabetes dataset, revealed an important limitation in the generation of discrete data spaces. While CTGAN appeared to preserve broad marginal distributions, allowing simpler linear models like Logistic Regression to retain predictive utility, it showed limitations in capturing the rigid, multivariate joint distributions and intricate microscopic patterns inherent to categorical logic. Consequently, advanced hierarchical models, such as Random Forest and XGBoost, experienced substantial performance degradation when attempting to extract complex decision rules from the categorical synthetic data.

In heterogeneous environments with mixed features, such as the Cardiovascular dataset, the synthesis process appeared to introduce a regularization-like effect. The overlap of continuous and categorical variables in the original data created significant statistical noise, which CTGAN appeared to smooth during generation. While this denoising behavior increased baseline metrics such as the F1 score by optimizing majority class predictions, it simultaneously reduced true boundary symmetry (as penalized by MCC), indicating a possible loss of precise boundary fidelity.

Ultimately, the general usability of CTGAN for clinical data synthesis presents a complex compromise. CTGAN can be a useful tool for privacy preserving data publication when the primary objective is to maintain global statistical properties, structural coherence, and basic linear relationships. In the evaluated datasets, it appeared more suitable for continuous numerical data and showed a possible denoising effect in mixed-feature datasets. However, researchers and practitioners must exercise caution when utilizing data generated by CTGAN for high stakes, highly detailed predictive modeling, particularly in environments dominated by complex categorical logic or highly imbalanced rare diseases, where the generative idealization process may blur critical clinical edge cases. Future work in synthetic tabular generation may prioritize architectural advancements that preserve complex categorical joint distributions without sacrificing minority-class boundary sharpness. Future studies could also extend the evaluation to additional healthcare datasets, compare CTGAN with alternative generative models, and incorporate repeated runs and statistical validation to provide a more robust assessment of performance differences.

References

- Cedolin, M., Erol Genevois, M. and Canbulat, Z., Hepatitis C diagnosis using computational intelligence techniques, *International Conference on Intelligent and Fuzzy Systems*, pp. 29-36, Cham, Switzerland, 2024.
- Centers for Disease Control and Prevention (CDC), Behavioral Risk Factor Surveillance System (BRFSS) 2015 Annual Survey Data, U.S. Department of Health and Human Services, 2015.
- Chen, X., Wu, Z., Shi, X., Cho, H. and Mukherjee, B., 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*, vol. 32, pp. 1227-1240, 2025.
- Dong, W., Wan, E. Y. F., Fong, D. Y. T., Tan, K. C., Tsui, W. W., Hui, E. M., Chan, K. H., Fung, C. S. C. and Lam, C. L. K., Development and validation of 10-year risk prediction models of cardiovascular disease in Chinese type 2 diabetes mellitus patients in primary care using interpretable machine learning-based methods, *Diabetes, Obesity and Metabolism*, vol. 26, pp. 3969-3987, 2024.
- Douzias, G., Bacao, F. and Last, F., Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE, *Information Sciences*, vol. 465, pp. 1-20, 2018.
- Esteban, C., Hyland, S. L. and Rätsch, G., Real-valued (medical) time series generation with recurrent conditional GANs, 2017.
- Han, H., Wang, W. and Mao, B., Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning, *Proceedings of the International Conference on Intelligent Computing*, pp. 878-887, Hefei, China, August 23-26, 2005.
- He, H., Bai, Y., Garcia, E. A. and Li, S., ADASYN: adaptive synthetic sampling approach for imbalanced learning, *Proceedings of the 2008 IEEE International Joint Conference on Neural Networks*, pp. 1322-1328, Hong Kong, China, June 1-8, 2008.
- Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P. and Parasa, S., On evaluation metrics for medical applications of artificial intelligence, *Scientific Reports*, vol. 12, pp. 5979, 2022.
- Klassen, A. C., Creswell, J., Plano Clark, V. L., Smith, K. and Meissner, H., Best practices in mixed methods for quality of life research, *Quality of Life Research*, vol. 21, pp. 377-380, 2012.

- Krittanawong, C., Virk, H. U. H., Bangalore, S., Wang, Z., Johnson, K. W., Pinotti, R., Zhang, H., Kaplin, S., Narasimhan, B., Kitai, T., Baber, U., Halperin, J. L. and Tang, W. H. W., Machine learning prediction in cardiovascular diseases: a meta-analysis, *Scientific Reports*, vol. 10, pp. 16057, 2020.
- Lévesque, V., Poirier, P., Despres, J. and Alméras, N., Relation between a simple lifestyle risk score and established biological risk factors for cardiovascular disease, *The American Journal of Cardiology*, vol. 120, no. 11, pp. 1939-1946, 2017.
- Li, J., Cairns, B. J., Li, J. and Zhu, T., Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications, *npj Digital Medicine*, vol. 6, pp. 98, 2023.
- Lichtinghagen, R., Klawonn, F. and Hoffmann, G., HCV data, 2020.
- Schultz, K., Hahn, W., Wolfien, M., Srivastava, P., Wolkenhauer, O. and Bej, S., Convgen: a convex space learning approach for deep-generative oversampling and imbalanced classification of small tabular datasets, *Pattern Recognition*, pp. 110138, 2024.
- SDV Team, The Synthetic Data Vault (SDV), 2026.
- Sidey-Gibbons, J. A. M. and Sidey-Gibbons, C. J., Machine learning in medicine: a practical introduction, *BMC Medical Research Methodology*, vol. 19, pp. 1-18, 2019.
- Teboul, A., Diabetes Health Indicators Dataset, Kaggle, 2021.
- Ulianova, S., Cardiovascular Disease Dataset, Kaggle, 2019.
- Xu, L., Skoularidou, M., Cuesta-Infante, A. and Veeramachaneni, K., Modeling tabular data using conditional GAN, *Advances in Neural Information Processing Systems*, 2019.
- Yoon, J., Drumright, L. N. and Van Der Schaar, M., Anonymization through data synthesis using generative adversarial networks (ads-gan), *IEEE Journal of Biomedical and Health Informatics*, vol. 24, pp. 2378-2388, 2020.
- Zhao, Z., Kunar, A., Birke, R., Van der Scheer, H. and Chen, L. Y., Ctab-gan+: enhancing tabular data synthesis, *Frontiers in Big Data*, vol. 6, pp. 1296508, 2024.
- Zheng, J.-X., Li, X., Zhu, J., Guan, S.-Y., Zhang, S.-X. and Wang, W.-M., Interpretable machine learning for predicting chronic kidney disease progression risk, *Digital Health*, vol. 10, 2024.

Biographies

Nur Sena Duman Çelikel is an M.Sc. candidate in Smart Systems Engineering at Galatasaray University and holds a B.Sc. degree in Industrial Engineering from the same institution. During her undergraduate studies, she conducted a capstone project on analyzing themes and sentiment in user comments in mobile application markets, proposing a methodology that integrates text preprocessing, topic extraction, and sentiment analysis to extract insights from large-scale unstructured data. She is currently pursuing her master's thesis on evaluating the usability of synthetic data for machine learning-based risk prediction in healthcare, focusing on the extent to which synthetic data can support predictive modeling while preserving data utility. Alongside her academic work, she has experience in digital transformation and data-driven product development. She currently works at Garanti BBVA Fleet as a Product Owner and Business & Process Development Specialist. Previously, she held roles at Hepsiburada, TEB Faktoring, and KPMG. She participated in the Erasmus+ program at Polytech Angers, France. She also contributed to social impact initiatives as a volunteer with Emmaüs Solidarité in France, supporting community-based social assistance activities.

Müjde Erol Genevois is a Professor in the Department of Industrial Engineering at Galatasaray University, specializing in production management, operations research, and project management methodologies. Her research primarily explores decision-making processes under uncertainty and the application of bootstrap methods in production planning. In addition to her scholarly work, she has extensive experience in higher education leadership, having served as the Secretary General of Galatasaray University, where she spearheaded strategic initiatives to institutionalize university-industry collaborations and professional networks. Her standing in the academic community is recognized by the French government with the "Ordre des Palmes Académiques," and she continues to be actively involved in international academic networks and strategic sector-university cooperation projects.