A Comparative Study of Generating Synthetic Tabular Data Using CTGAN and DP-CTGAN : Fidelity and Risk Assessment

University of Turku
Department of Computing
Master of Science (Tech) Thesis
Department of Computing
November 2025
Namita Bastola

UNIVERSITY OF TURKU
Department of Computing

Namita Bastola: A Comparative Study of Generating Synthetic Tabular Data Using CTGAN and DP-CTGAN : Fidelity and Risk Assessment

Master of Science (Tech) Thesis, 55 p.
Department of Computing
November 2025

The accelerated increase of digital records in the medical field offers unprecedented opportunities for research advancement and data-driven healthcare analytics. However, the sensitive nature of data raises significant privacy concerns that constrain direct data sharing and collaborative research. Synthetic data generation has been explored as the most practical approach to address these challenges by producing more realistic artificial datasets that closely preserve the key statistical properties while mitigating privacy risk.

This thesis examines the synthetic tabular data generation following the Conditional Tabular Generative Adversarial Networks (CTGAN) proposed by Xu et al., which is designed to generate synthetic tabular data and further integrates differential privacy with varying privacy budgets. For each model, this study implements rigorous model training and executes it across multiple independent runs, producing synthetic datasets and preserving them in a well-structured directory for evaluation. For fidelity assessment, quantitative evaluation metrics—Hellinger distance, pairwise correlation (Spearman and Cramér's V), and SDV metrics such as Total Variation Distance analysis for categorical variables and the Kolmogorov-Smirnov (KS) statistic for continuous variables—have been implemented. Additionally, for privacy exposure measurement, this study adopted a single-out risk—an attack-based framework,that simulates an adversary's likelihood and potential of re-identification for each run-wise synthetic dataset.

Furthermore, the thesis analyzes feature-wise fidelity and run-to-run model variability across multiple trainings and the inherent trade-off between data privacy and fidelity for each model. The results show that CTGAN generates better quality synthetic tabular data compared to DPCTGAN model variants, as privacy control causes fidelity degradation. However, DPCTGAN models are applicable when privacy is a major factor, but there is a trade-off in quality. This work is applicable to collaborative guidelines for generating synthetic data and analyzing healthcare data while maintaining privacy and quality of data.

Keywords: Synthetic Data Generation, Medical Tabular Datasets, Differential Privacy, Generative Adversarial Networks, CTGAN, DPCTGAN, Fidelity,, Single-Out Risk, Model Stability

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background of the Study

Over the last few years, the exponential increase of digital data in the healthcare sector has accelerated scientific research advancement and driven progress in artificial intelligence analytics and machine learning. In this domain, the tabular data consists of structural records with heterogeneous data types such as integer and float and stays in a dominant format in different applications, including the financial, social science, and healthcare fields. However, the sensitivity of data raises significant privacy concerns; direct sharing of such data can result in exposure to confidential individual information that leads to ethical, social, and legal risks. To address these privacy concerns, the concept of synthetic data generation has been explored as the most viable solution; it involves AI-generated data that closely resembles the statistical properties and structure of real data while reducing the risk of re-identification, especially during the sharing of sensitive information [1].

There are several synthetic data generations approaches available, with each offering different trade-offs concerning its complexity, privacy, and realism. Gaussian copulas and Monte Carlo simulation are traditional statical techniques, which simulate the data based on their statistical characteristics [2]; however, they still have difficulties capturing complex and high-dimensional data distributions [3]. In addition, parametric statistical methods, such as multivariate Gaussian distribution,

that generate sample data using the mean and covariance matrix of original datasets [4]. The agent-based model simulates the individual entities' behaviors in existing systems to produce synthetic data that illustrate the complex entities' behaviors and is mainly used in the epidemiology field. Deep learning methods are a type of generative model, including Generative Adversarial Networks (GANs) [5], which are designed to synthesize data by learning complex data distributions using iterative refinement between a generator and the discriminator networks. Along the same lines, Variational Autoencoders (VAEs) synthesize data by learning probabilistic latent representations to generate similarly distributed data [6]. The Conditional Tabular GAN (CTGAN) [7], defined as an expanded version of GAN architecture, was particularly designed to address principal challenges in tabular data synthesis, including mixed data type, imbalanced data distribution, and multi-modal categorical distribution.

Synthetic data is defined as artificial datasets that are produced by model training based on statistical properties or the structure of original data [8], [9]. Additionally, the synthetic data generation approach has been used to mitigate the risk associated with handling and processing personal confidential data in healthcare applications [8]. These approaches are consolidated with many other benefits, including algorithm testing, data augmentation, and resources for advanced data-driven healthcare analytics. In spite of showing promising performance in several application fields, the generative adversarial network model trained on real data still has privacy challenges due to the potential risk for unauthorized disclosure of people's confidential information. These challenges have been empirically shown in tabular data, where the adversary used auxiliary information to match synthetic data with real data. To reduce the privacy risks and improve data security, it is essential to augment GAN-based synthetic data with additional privacy preservation techniques [10]. The foundation of differential privacy was established as a rigorous

mathematical framework that prevents the influence of any single data point on the learned model and, as a result, reduces the vulnerability of people's identities to disclosure in generated synthetic data. Integration of differential privacy often results in a degradation of the utility trade-off due to added noise, causing thorough evaluation of the effectiveness of privacy-preserving GANs.

The focus of this study is to train a generative model using both CTGAN synthesizers and a DPCTGAN synthesizer to generate synthetic tabular healthcare data across multiple runs. Thereafter, the generated healthcare data quality and associated privacy risk are quantified and used to assuage privacy risk concerns.

1.2 Objectives of the Study

The primary goal of this study is to generate synthetic tabular healthcare data by training a GAN-based model and further integrating differentially private stochastic gradient descent with varying privacy budgets. The privacy budget ϵ is the major parameter of the DP mechanism that quantifies the generative model's privacy loss caused by insignificant change in the input data. Thereafter, the study evaluates and compares the resulting quality and privacy risk associated with generated synthetic tabular healthcare data. Furthermore, we also analyze the variability of models across multiple independent training runs for each model. Specifically, it aims to understand how varying differential privacy budgets ϵ values affect the balance between data quality and privacy guarantees.

1. To generate synthetic tabular data using the CTGAN algorithm referenced in [11] and the DP-CTGAN documented in [12] across multiple independent training runs with varying differential privacy levels, ϵ, with the aim of evaluating and comparing the resulting data quality. For data quality, perform statistical analyses, including pairwise correlation, Hellinger distance, total variation distance, and the

Kolmogorov-Smirnov statistic, to assess the quality of the generated data.

2. To evaluate the privacy leakage in synthetic tabular data using both CTGAN with and without differentially private integrated model variants, implement the single-out risk-based privacy framework proposed in [13]. Examine the trade-off between privacy and fidelity of generated synthetic tabular datasets using both with and without integrated DP mechanisms ($\epsilon = 1, 10, 100$) and determine the best balance between data similarity and privacy protection.

1.3 Organization of the Thesis

The thesis work is organized as follows:

Chapter 1, "Introduction," provides the background and objective of the study. Chapter 2, "Literature Review," provides the key theoretical concepts that are associated with this study, an exhaustive review of the GANs-based synthetic data generation methods with their advantages and challenges. It also covers the theory of differential privacy (DP)—a main focus of this work—along with the research gap. Chapter 3, "Methods and Materials," discusses the theoretical methods and materials used in this research, focusing on the CTGAN methods with an explanation of how this model addresses key challenges that occur when handling tabular data within GAN-based models. Additionally, it also describes the thesis's proposed model: the integration of DP into the CTGAN model with the major modifications made to align with the objectives of this study. This chapter also provides an overview of the adopted dataset for the study. Chapter 4, "Model Process and Evaluation," includes data preprocessing, experimental setup for model training, synthetic data generation process, and adopted quantitative evaluation of metrics. These metrics include both generated data quality assessments and privacy risk evaluations. Chapter 5, "Results and Discussion," provides a detailed discussion about

the results obtained from the evaluation metrics and highlights the key outcomes in alignment with the objective of this thesis work outlined in the first chapter. Chapter 6, "Conclusion," highlights the key findings of the study and gives suggestions for potential future research work. Finally, Chapter 7, "Declaration of AI Usage," mentioned the different AI tools with details about how I have used them while working on my thesis work.

2 Literature Review

After defining the objective of the study, this chapter reviews the key fundamental theory associated with this work. Firstly, it analyzes the associated privacy challenges with synthetic data generation. Then there's an overview of synthetic data generative methods with their advantages, and limitations. Furthermore, we explore existing research gaps and provide details of an overview of differential privacy, which is one of the major focuses of this study.

2.1 Privacy Concerns

The sustained increase in the volume of medical data records has simultaneously increased the opportunity of using those data for secondary purposes, especially for research purposes [14]. In addition, sharing medical information leads to many benefits, including evidence-based decision-making, improving research replicability, and reducing clinical trial costs [14]. However, sharing this data often presents a significant challenge in keeping the ideal balance between protecting individual privacy and easing research for potential benefits.

Privacy is a very crucial factor that must be considered when dealing with an individual's sensitive health data, as unauthorized disclosure of patient personal information can lead to unequal treatment, misdiagnoses, loss of integrity, and trust toward healthcare institutions [15], [16]. To govern these unauthorized disclosures, various strict rules and regulations were introduced, such as the Health Insurance

Portability and Accountability Act (HIPAA) [17], the Privacy Rule, and the General Data Protection Regulation (GDPR) [18] in the European Union. They provide an extensive set of frameworks and advisories for private medical data protection. That explains how to collect, handle, and share individual sensitive medical data. For instance, GDPR provides a comprehensive framework for safeguarding individual data rights by imposing strict guidelines and principles, which provide an extra layer of protection.

The organization's guidelines align with these privacy regulatory frameworks to prevent the risk associated with their individual data privacy by integrating legal requirements, secure data transfer protocols, and information consent protocols with the data anonymization technique (i.e., removing or changing personally identifiable information as per the above-mentioned regulation guidelines).

In summary, it is important to consider the privacy factor while dealing with sharing patient information for secondary purposes, as it can lead to algorithm biases and medical treatment inequality. The above-mentioned, existing regulatory guidelines serve as global best policies and a framework for international sensitive data sharing practices and collaboration, as well as further efforts to balance innovation with privacy. In this data-driven era, finding the best solution that strengthens both individual privacy and scientific evolution have become the main challenge.

2.2 Synthetic Data Generation

Synthetic data generation techniques involve generating artificially stimulated data that preserves the statistical distribution of real data without disclosure of people's sensitive information while addressing privacy concerns [19]. In addition, synthetic data generation is an approach that synthesizes data by using the existing prebuilt mathematical model or algorithm [1]. Furthermore, it is defined as statistical models that have been designed to mitigate the privacy risk associated with handling and

unfolding sensitive data, especially in the context of health data applications [8].

There are many popular approaches available that depend on deep generative adversarial models for synthetic data generation, including Variational Autoencoders (VAE) [6] and Generative Adversarial Networks (GAN) [5]. These models learned from real sensitive data and produced artificial data that closely keeps the statistical structure of original data. The primary objective of these models is to guarantee that our produced data closely mimics the statistical properties and structure of real data (i.e., correlation among attributes) and simultaneously reduce the likelihood of people's personal information breaches and risk of re-identification. In this case, it is considered one of the most promising privacy-enhancing technologies, as it does not directly duplicate or lose information from original sensitive records [8].

Additionally, artificial data generated from these methods serve as valuable resources to support research in many fields, such as advanced scientific research, artificial intelligence analytics, machine learning analysis, and predictive models. Nonetheless, it is crucial to consider that producing high-utility statistical distribution data obliges specialized and robust methods, and that findings are consistent with those obtained from real-world data remains and continues to be a subject of active investigation.

2.2.1 GAN-based Synthetic Data Generation Techniques

Over the last few years, synthetic data generation methods have attracted significant attention in advanced research and machine learning model training across various sectors. As a result, this method has been used in a diverse set of data formats, including tabular data, video records, image data, speech generation, and so on.

Generative Adversarial Networking (GANs)

Following a study conducted by Goodfellow et al.,((2014) [5], the General Adversarial Network (GAN) was introduced as a type of generative model that works by integrating through an adversarial process between two neural networks known as a generator and a discriminator. Both neural networks are trained at the same time with different objectives. The generator aims to learn the estimated distribution of data represented by p_{mode} by using a training data sample obtained from the underlying distribution of p_{data}. In contrast, the discriminator functions to figure out the likelihood that a given sample of data is derived from the training samples rather than being generated by the generator.

To understand model concepts in detail, the target details distribution p_{mode}, the generator G produces the sample denoted as G(z) by transforming the noise vector 'z' from the predefined distribution. Meanwhile, the discriminator D function calculates the probability that a given input x is genuine data or a synthetic example generated by G(z). These generative models use different structures, including autoencoders, Convolutional Neural Networks (CNN) [20], and Recurrent Neural Networks (RNN) [21], and others for both the generator and discriminator. As an output, both models G and D update their weights by minimizing two different loss functions referred to as $Loss_{\text{G}}$ and $Loss_{\text{D}}$, which guide the adjustment of network parameters.

At the time of the model training process, the generator changes its sample solely depending on backpropagation from the generated samples, i.e., G(z). In contrast, the discriminator updates its weights based on feedback from both generated and real data [22]. The primary goal of training overall GANs is to express them as a two-player min-max game with the value function V (D, G), where both the generator and discriminator work to optimize their own objectives [5].

Here, the GANs model frameworks show the continuous interaction between both the discriminator and generator. As a result, they guide the system towards

equilibrium, where the generator produces sample data that closely emulates the statistical distribution of real data.

Conditional Tabular Generative Adversarial Network (CTGAN) Techniques

The GAN-based models have shown notable results in the generation of tabular synthetic data, particularly valuable in healthcare, finance, and social science domains, as their privacy and scarcity of data are the most significant concerns. However, generating tabular synthetic data has become relatively more challenging than generating image data due to the heterogeneous mix of data types and the inherent characteristics of tabular data, including multi-modal, non-normal distribution with each continuous and categorical column, and imbalanced categorical columns.

To address the principal challenges faced by GANs while dealing with tabular data, Xu et al.,(2019) [7], introduced the Conditional Tabular Generative Adversarial Network for generating synthetic tabular data. The model has introduced several new techniques, such as incorporating mode-specific normalization into the training pipeline, making architectural redesigns, and using conditional generators to handle imbalanced class distribution and training via data sampling [7]. In this way, the GAN-based model can be proposed as one of the promising ways to handle inherent characteristic issues of tabular data and help researchers to generate more balanced distributed data, which would enhance model efficiency and analytical soundness. Furthermore, an in-depth discussion of this method has been provided in 3.1.

Conditional Generative Adversarial Networking (cGANs)

Even after showing significant performance in various fields of application, the GANs model was still unable to control the generation process under specific conditions. To address this issue, the concept of cGANs framework was developed as an expanded

version of the standard GAN, which conditions both neural networks on additional information, denoted as y, as class labels or other modalities [23]. The conditioning is applied by providing the auxiliary variables y as additional inputs into both the generator and discriminator, to handle the lack of mode control in the original unconditional GANs [23]. This technique enhance the versatility of original GANs by enabling synthetic data in a multi-modal setting and supporting different fields of application.

2.2.2 Advantages

An important advantage of the synthetic data generation method is its ability to scale data volume for model training. That helps address data scarcity issues and disclosure risk related to real data. For example, the study in [19] demonstrated how the implementation of synthetic data generation approaches, like CTGAN for tabular EHR radiomics data, directly addresses the limited sample issue occurring in the clinical model training by generating large-scale synthetic data.

The second advantage is its capability to analyze and construct a model framework without relying on the original datasets. It helps to mitigate the risk of leading to the revealing of individuals' sensitive personal information. Synthetic data does not secure the privacy guarantee itself, so most of the time additional techniques are needed to ensure the protection against privacy risks. To illustrate, a systematic review in [1] highlighted that ensuring privacy in synthetic data itself stands for a sophisticated challenge that typically needs the blend of an additional approach, such as k-anonymity [24] and differential privacy [25].

The third advantage is its strength in cross-substitution collaboration without original data sharing. This approach creates a universal dataset that enables multiple institutions to perform model development without shared real data and privacy waivers. For example, the study conducted in [19] shows how five hospitals have

conducted EHR consortium research by using a synthetic pool to replace real data.

Another advantage of the GANs-based synthetic data generation methods is that they offer several ways to generate customized datasets in different domains, such as health and finance. This enables precise control over the feature distribution, overcoming data quality issues such as noise, missing values, and unbalanced classes inherent in real tabular data [7]. Moreover, it can augment limited data sets by producing diverse, high-fidelity samples, enhancing the robustness and generalization of AI models [26]. Moreover, the use of generated data eases the training, validation, and benchmarking of machine learning models, whereas real data is more scarce, restricted, or costly to get [27].

2.2.3 Challenges and Limitations

As already discussed in previous sections, synthetic data generation has the potential to be the most promising tool for supporting several advanced research and innovation areas and robust privacy preservation. It helps reduce bias, address issues arising from limited or incomplete real datasets, and make it easier to model samples that are more accurate and reliable. However, many studies have indicated that it's not automatically secure, and in some cases, it can still lead to exposing the confidential information [28], [29].

A major drawback of synthetic data generation is that the generative models, especially the ones that are trained on real-world data, are more vulnerable to inference attacks; they can potentially leak the underlying individual sensitive information, particularly in the field of domains where the dataset contains distinctive, identifiable patterns [30]. In addition, the study conducted in [31] also discussed the potential risk of data disclosure and adversary attack based on synthetic data. First, disclosure of personal identifiers does not look informative; just imagine the significant risk posed by discovering that a well-known public figure, whose most

of the personal information has already been publicly available, is included in a dataset related to a sensitive issue—such as a financial fraud investigation or a leak of private medical records. These incidents increase the re-identification of attack privacy concern for synthetic data. That is why it is so important to apply robust privacy measures carefully to ensure that generated data stays useful without putting anyone's privacy at risk.

A further imitation of synthetic data generation lies in that it often fails to extract inherent statistical properties of original datasets. There are several unanticipated events known as outliers that are quite common in real life and often make it extremely hard for a generative model to represent those extreme cases correctly and accurately, at least not confidently [29], [32]. As a consequence, the models developed using only synthetic data may exhibit reduced performance on real-world senior datasets where outliers are included. Consider the example provided in [32], a situation involving a rare event—such as the inclusion of a multi billionaire within a wealth dataset. If so, the generator might not accurately preserve the statistical relationship and distribution of original data outliers nor expose the potential of rare outliers that disclose sensitive information about these individuals.

While synthetic data generation has been introduced to mitigate biases found in real-world datasets, as previously mentioned in the above advantages section. However, it can paradoxically introduce new biases if there is a full diversion of real dataset distribution. As a result, inaccuracies in generated synthetic data have the potential to propagate these biases into evaluation models, thereby adversely influencing their performance and reliability.

In conclusion, these methods are promoted as a potential solution for advanced research work and machine learning model training across various sectors by providing accessible and workable data sets. However, these artificially generated data should not be viewed as a full substitute for original data, because the implementa-

tion of privacy-preserving techniques inevitably introduces some level of distortion [32].

2.3 Research Gap

While reviewing different papers and articles, it has been noted that most research on generative models are mainly focuses on producing synthetic data—particularly in image-related applications—rather than for tabular data. However, in some studies this gap has been addressed by using generative models in the healthcare application to generate synthetic tabular data with and without the differential privacy [33], [34]. The authors in [7] introduced the CTGAN methods, which address the challenges of existing generative models for tabular data modeling. In addition, some studies show the adoption of the CTGAN model integrated with Differential Privacy (DP) to generate synthetic data, perform a risk-based attack privacy framework, and evaluate their success on synthetic data [13].

Throughout this review, a clear research gap has been found: systematic assessment of the model robustness and run-to-run variability of generated synthetic datasets across multiple training instances. Specifically, it focused on the impact of the principal privacy budget parameter that controls the extent of privacy loss. As a result, the study aims to address this gap by conducting a series of independent model trainings for both CTGAN and DPCTGAN with varying privacy budgets, generating tabular synthetic datasets by fitting run-wise synthesizers for each model, and conducting a comprehensive analysis of their statistical data fidelity and privacy guarantee.

2.4 Differential Privacy

Synthetic data is considered one of the most promising techniques to protect privacy. However, traditional methods are not sufficient to fully prevent inadvertently leaking identifiable patterns, whereas the differential privacy approach is grounded in a rigorous mathematical definition of privacy that enables it to be more robust in sensitive domains like finance, healthcare, and social science.

Differential privacy (DP) has been defined as a formal guarantee made by a data holder to a data subject, i.e., the individual whose data are included in datasets, that their participation in any study or analysis will not cause any adverse effect, irrespective of others' studies, datasets, or information sources [25]. More formally, it provides a rigorous mathematical framework that limits the distinguishability between different datasets of one individual, thereby guaranteeing individual privacy while enabling useful population-level insight [25].

As defined by Dwork and Roth (2014) [25], it ensures the possible outcomes of a randomized algorithm $\mathcal{A}$ are private if for any pair of datasets with only one element differing, (D) and (D'), and for any subset $S \subseteq \text{Range}(\mathcal{A})$ of possible results defined as:

$$\Pr[\mathcal{A}(D) \in S] \leq e^{\epsilon} \Pr[\mathcal{A}(D') \in S]$$

In this context, parameter ϵ acts as an upper bound limit controlling the extent of privacy leakage, and parameter δ defines the likelihood of the mechanism failing to provide the intended privacy protection.

Epsilon (ϵ) , often referred to as the "privacy budget," plays a key vital role in the DP framework by regulating how much privacy can be compromised. Furthermore, ϵ especially defines how similar the results of a given algorithm are when it processes two adjacent subsets. This measure captures the extent of the model's privacy loss

caused by insignificant changes in the input data.

By the definition of epsilon, a smaller value of (ϵ) defines a strong privacy guarantee; as a result, it becomes more alike when comparing datasets differing by a single individual. However, it can become more difficult for an adversary who is analyzing only the outputs to identify whether any person's data is part of the dataset or not. As a result, this enhanced privacy guarantee typically comes with the cost of reducing the usefulness or accuracy of generated synthetic data.

It is always very difficult to establish an acceptable range for the privacy parameter ϵ, as it varies on the application case and the underlying data properties. As mentioned in [25], there is no clear agreed rule or process for selecting the privacy parameter ϵ, with no unified strategy for dealing with this and other practical decisions. Because of the importance of these, it is essential to have ongoing collaborative learning within the differential privacy community.

Delta (δ) was added to expand upon the original definition of differential privacy [35], providing a relaxation of the strict constraint that measures change in output of an algorithm when a small modification is made in individuals' data.

It was mentioned that allowing a nonzero δ shows that there is a minor adjustment change that the differential privacy guarantee can breach. However, δ should be optimally smaller than the reciprocal of the data samples [36]. Furthermore, allowing the existence of a non-zero for the delta parameter allows an insignificant change that could compromise the DP guarantee until delta remains slightly less than the reciprocal of the dataset size.

3 Methods and Materials

In this chapter, we briefly overview the fundamental methods that we implemented in our thesis work, such as CTGAN and DPCTGAN for model training. Then there's a comprehensive discussion about the thesis's proposed model and a description of the adopted dataset.

3.1 Conditional Tabular Generative Adversarial Network

The GAN-based model has shown outstanding performance in training generative models in diverse application domains, such as image synthesis and text generation. Nonetheless, the study conducted by Xu et al. (2019) [7] highlights several difficulties that the GAN-based model faced while designing a model for real-world tabular data. Which includes a mixture of heterogeneous categorical and continuous features. Standard GAN-based architectures struggle with handling these mixed data types, non-Gaussian and multimodal distributions, and sparse one-hot-encoded vectors. In addition, highly imbalanced categorical columns can lead to mode collapse, where minor categories are poorly represented or completely missed [7].

These limitations led to the introduction of the CTGAN model, which is an extended version of conditional GANs. The CTGAN model is designed to address the limitation faced by original GANs while handling real-world tabular data by

incorporating new features such as model-specific normalization to effectively guide the distributions that deviate from Gaussian and present multiple modes [7]. Additionally, it also integrates a conditional generator and training via data sampling to deal with the continuous features and unbalanced discrete column.

The CTGAN model used in this thesis work has adopted notations from the work present in [7], where it was originally designed as:

- $x_1 \oplus x_2 \oplus \ldots$: as concatenation of vectors represent as $x_1, x_2, \ldots$
- $\text{gumbel}_\tau(x)$: application of the Gumbel softmax [37] with temperature parameter τ on vector x
- $\text{leaky}_\gamma(x)$: leaky ReLU activation applied to x with leakiness coefficient γ
- $FC_{u \to v}(x)$: as linear transformation mapping a u-dimensional input vector x to a v-dimensional output

Here, the CTGAN model notation used `tanh`, `ReLU`, and `softmax` activation functions, batch normalization [38], and dropout regularization [39].

The CTGAN model has implements a novel multi-specific normalization by fitting a Variational Gaussian mixture model (VGM) [40], with m_i components applied independently across all continuous columns C_i to get its potentially multi-modal distribution. Here, binary vectors are used to encode categorical variables, and missing values are handled by introducing an added separate class, effectively adding a distinct bit to the one-hot vectors. Additionally, to handle the imbalance in categorical columns, this model employs conditional training by sampling records with the probability proportional to the logarithm of category frequency.

This model also incorporates conditioning variables for both the generator (G) and discriminator (D), which allows the model to produce synthetic data based on specific categorical attributes. This helps the model preserve complex dependencies and ensure adequate representation of minority categories during generation.

Conditional vectors define mask vectors for all discrete columns in datasets. The conditioning vector is constructed by the concatenation of all mask vectors of discrete columns. The generator loss function tries to generate more realistic sample datasets that fool the discriminator classified as real while also ensuring the generated sample datasets respect the condition consistency via the conditional loss term.

Further, this model also innovates the use of a specialized sampling technique to balance the representation of discrete and categorical columns while training the model. Columns are randomly selected, and categories are sampled according to the logarithm of their frequency, ensuring the generator learns to produce rare categories as well as common ones. The steps of training-by-sampling to generate synthetic rows in the model are defined as follows:

1. First initialize zero mask vectors $m_i = \mathbf{0}$ for each categorical column of data D_i, where $i = 1, 2, \ldots, N_d$.

2. Uniformly sample a discrete column D_{i^*} from all discrete columns.

3. Sample a category k^* from D_{i^*} according to the logarithm of the observed frequency distribution of categories:

$$P(k^*) \propto \log(\text{frequency}(k^*))$$

4. Set the mask $m_{i^*}^{(k^*)} = 1$ and form the conditioning vector:

$$\text{cond} = m_1 \oplus m_2 \oplus \cdots \oplus m_{N_d}$$

5. Sample real rows with condition $D_{i^*} = k^*$ for discriminator training.

6. Generate synthetic rows by conditioning the generator on cond:

$$\hat{r} = G(z, \text{cond}), \quad z \sim \mathcal{N}(0, I).$$

In the CTGAN model, the conditional generator is able to generate synthetic rows based on specific values of the discrete columns of data. By using sample-based training, both the conditional vector and training data samples are drawn based on the logarithm frequency of each category; thus, the CTGAN-based model effectively explores all possible ranges of discrete values [7].

3.2 Thesis Adopted Model

The model adopted in this work is designed for the evaluation of generated synthetic tabular data along with three essential criteria: data fidelity, model variability, and privacy assessment across multiple independent training runs for both the CTGAN and DPCTGAN model variants. These models incorporate conditional generative adversarial networks along with additional techniques, such as mode-specific normalization and fully connected networks. Additionally, these models ensure more stability during training by leveraging the Wasserstein distance and gradient penalty.

The main idea of the adopted model is to evaluate how the model training across multiple runs with varying privacy budgets affects the run-to-run variability of the model, synthetic data fidelity, and the trade-off between fidelity and privacy associated with the generated dataset. The adopted model follows a structural experimental workflow that starts with the preprocessing of data that consists of a mix of continuous and categorical variables. Additionally, these models ensure more stability during training by using the Wasserstein distance and gradient penalty.

Safeguarding the privacy guarantee in the synthetic data generator process itself is an inherently complex process that usually requires a combination of multiple

additional techniques, such as k-anonymity, encryption, and differential privacy, to effectively protect against privacy risks [1]. To address the privacy challenges associated with synthetic data generation, this work adopts an extended version of the CTGAN model with the incorporation of a DP optimizer implemented with Opacus during model training, which adds noise and enforces DP guarantees on the trained model [41]. Integration of DP optimizers during model training ensures strong privacy guarantees. As a result, synthetic data generated using this model inherits the same privacy guarantee as the underlying trained model.

In the synthetic data generation process, samples were generated using models trained with two different synthesizers: one CTGAN without privacy constraints and another modified DPCTGAN with $\epsilon = 1$, 10, 100. For the CTGAN, the model trained with the synthesizer algorithm provided in [11]. For the DPCTGAN model, the trained model implements a synthesizer, documented in [12], with a key adjustment. For details on the synthesizer setup, see subsection (4.2.1). Additionally, the Opacus Privacy Engine's secure random number generator (RNG) parameter is enabled in the training pipeline, and the gradient penalty term is removed from the optimization process.

To support the reproducibility and comparability of results, the model has been trained multiple times. Generate synthetic tabular data by running a pretrained generative model multiple times for each model and save them separately to keep a robust and unbiased analysis. After successfully generating synthetic datasets, the thesis model calculated the Hellinger distance, Total Variation Distance analysis [42], the Kolmogorov-Smirnov (KS) statistic [43], and pairwise correlation to measure the statistical fidelity of the generated dataset. Additionally, a privacy risk evaluation was performed by adopting an attack-based Anonymeter framework defined in [13].

3.3 Data Overview

This thesis work adopted the Cardiovascular dataset that has been published by Svetlana Ulianova and the Kaggle community and is publicly available in [44]. Due to the open accessibility, it serves as a valuable resource for researchers working in the medical and data science communities. That supports enhanced reproducible analysis, benchmarking, and transparency in research on cardiovascular health using advanced machine learning techniques.

Feature	Variable	Format / Description
Age	age	integer (days)
Height	height	integer (cm)
Weight	weight	float (kg)
Gender	gender	categorical 1 = women, 2 = men
Systolic blood pressure	ap_hi	integer
Diastolic blood pressure	ap_lo	integer
Cholesterol	cholesterol	1=normal, 2=above normal, 3=well above normal
Glucose	gluc	1=normal, 2=above normal, 3=well above normal
Smoking	smoke	binary (yes/no)
Alcohol intake	alco	binary (yes/no)
Physical activity	active	binary (yes/no)
Cardiovascular disease	cardio	binary (yes/no)

Table 3.1: Summary of Dataset

As shown in table 3.1, the adopted dataset includes 70,000 records of patients with 11 distinctive features, including both demographic factors and life indicators, and one target variable. Data conversion was applied to age column values that were originally recorded in days to years.

4 Thesis Model Process and Evaluation

This chapter briefly explores data preprocessing and the experimental setup of the model, including model training, synthesization setup, model pipelining, synthetic data generation, and evaluation of the quality of the generated data. Additionally, assess the privacy guarantee of generated data using the Anonymeter framework.

4.1 Data Preprocessing

Data preprocessing is a fundamental process before model training, as the performance of the GAN-based models mainly depends on the cleaned and appropriately structured data to accurately learn the data distribution. This section consists of details discussing how original data has been preprocessed to feed and train the model.

Label Encoding

Label encoding transforms categorical features into different integers ranging from 1 to N, where N is defined as a cardinality of features [45]. In the adopted dataset, each categorical feature is available in an encoded format by assigning different unique integer value ranges, like gender (1: woman, 2: man). As a result, no additional categorical feature transformation or encoding is required during data preprocessing.

Data Sampling and Train-Test Split

Data sampling is a widely used process of selecting a subset of units from a large population dataset to represent information about the whole dataset [46]. This process allows research to make conclusions about the whole population without studying each and every dataset.

This study applied a stratified split based on the target class, 'Cardio,' to reduce the class imbalance and ensure a fair model assessment. After the sampling process, we collected 25,000 subset data for further analysis. Additionally, to reduce overfitting and enhance the generalization performance of models on unseen data, we divide the subset data into two distinct sets, such as a training and a test set, in an 80:20 proportion.

Data Binning

The differential privacy integrated generators cannot be directly applied to continuous data; they always require their input data in continuous or discrete values [13]. By following the equal-width discretization approach defined in [13], [47], we encoded each continuous column of our dataset into 50 equally spaced bins determined from the min-max range of the original dataset itself.

The differential privacy integrated model was trained using the original preencoded categorical variables and binned continuous variables. The synthetic data generated using the DPCTGAN model was in the form of bin labels. To preserve statistical fidelity for later analysis, synthetic bin labels have to be mapped back into corresponding real continuous variable formats. In this study, a synthetic bin label has been mapped back by drawing a random number uniformly from the interval defined by the respective bin edges of each synthetic bin label.

4.2 Experimental Setup of Model

After having knowledge and understanding of the dataset used for this work, the next step is to illustrate the practical implementation of our thesis work.

4.2.1 Model Training

The training process of our two distinct neural networks—a generator and a discriminator—uses an adversarial approach to generate better-quality, authentic, label-conditioned synthetic tabular data that effectively handles both continuous and categorical features. In this section, we briefly discuss the core elements of the thesis model process, including synthesization setup, initialization and hyperparameters, and DP optimizers in model training.

Synthesization Setup

The synthesization setup for generating synthetic medical tabular data has employed the CTGAN synthesizer and utilized default hyperparameters, as defined in [11]. While performing the search in prebuilt mode, we identified the default learning rate for both the generator and discriminator as 2×10^{-4} and set the log frequency for categorical levels to true.

Alternatively, for the DPCTGAN model, the study employed the synthesizer available in [12] with a certain modification: setting the learning rate of the generator and the discriminator as 10^{-5} and 10^{-3}, respectively, and updating the discriminator five times for each update of the generator to preserve the training balance. Change the set batch size to 64, the noise multipliers to 0.5, the clip norm to 10, and the remaining other hyperparameters to the same as default. Further, the total differential privacy budget has been split into two fractions, with 20% allocated for data preprocessing and the remaining 80% for the training of DPCGTAN synthesizers. The model has been trained with varying ϵ values to learn and generate

synthetic tabular datasets.

Differential Privacy in Model Training

In the thesis-adopted model, differential privacy has been integrated into the model training pipeline by using Opacus—a prebuilt library that is used to facilitate the training of PyTorch models with DP guarantee. The privacy engine is one of the major components of Opacus that preserves privacy by altering its gradients during the model training through the addition of calibrated noise and implementing gradient clipping. The privacy engineer included a noise multiplier parameter that quantifies the noise level by measuring the proportion of the Gaussian noise standard deviation and the model's L2-sensitivity computations [41]. In this model, we established a targeted value as $\delta = 1 \times 10^{-5}$, which stands for the maximum allowable probability of the tolerance of privacy leakage.

Furthermore, the privacy engineer parameter provides secure_model for cryptographically strong differential privacy guarantees, along with the secure random number generator for noise and shuffling, which helps prevent certain floating-point arithmetic-based attacks. In the adopted model, we set secure_model to establish a cryptographically strong differential privacy guarantee in our generated synthetic tabular dataset to reduce the privacy leakage and prevent arithmetic-based attacks.

Model Training Pipeline

After data sampling and preprocessing, we obtain 25,000 datasets that are further divided into training and testing sets in the 80:20 ratio: 20,000 datasets for training and 5,000 datasets for testing.

As shown in figure 4.1, model training has been exclusively performed on the training dataset using both CTGAN and DPCTGAN synthesizers with varying ϵ values. For each synthesizer, the model has been trained across 30 independent

iterations on the `train_df` and saved into separate folders for further use. To preserve dataset parity, the synthetic data generated is designed to be of the same number of samples as the training dataset. The generator model, denoted by G, has been responsible for producing synthetic data. To establish the basis of this work, training setups have been built by employing two different generators, such as a conditional tabular generative adversarial network and differential privacy mechanisms in conditional tabular GANs to enhance data privacy guarantees.

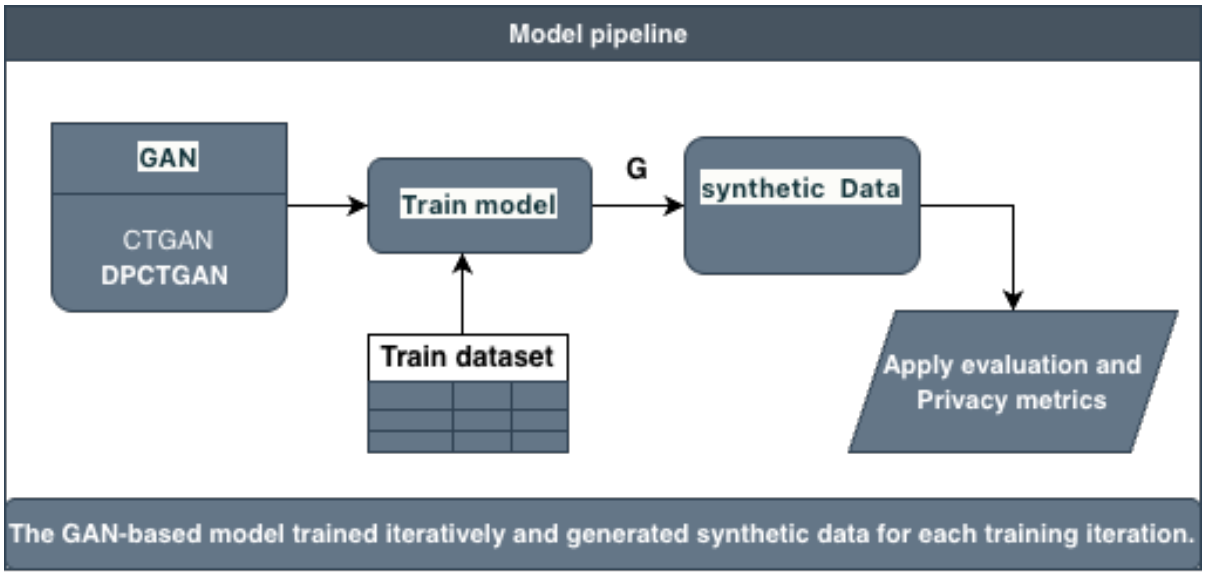


Figure 4.1: Thesis Model

4.2.2 Synthetic Data Generation

Once the model training had been completed, we utilized prebuilt run-to-run synthesizer models to fit and generate multiple synthetic datasets for each model and saved them in separate folders to maintain a model-robust and unbiased analysis. As shown in table 4.1, for each model, we generated 30 independent synthetic datasets consisting of 20,000 samples with 12 different attributes.

Synthesizer	**(ϵ)**	**Runs**	**Samples Per Dataset**	**Total Datasets**
CTGAN	–	30	20,000	30
DPCTGAN	1	30	20,000	30
DPCTGAN	10	30	20,000	30
DPCTGAN	100	30	20,000	30
Total		**120 Runs**		**120 Datasets**

Table 4.1: Summary of Generated Synthetic Datasets

Since the CTGAN model directly incorporates model-specific normalization into the prebuilt training pipeline to handle imbalanced class distribution [7], we don't have to further normalize data. In this work, label encoding has been performed on categorical attributes, defined as transforming categorical features into different integers ranging from 1 to N. For the CTGAN model, we used labeled encoded categorical variables and normal continuous variables without any normalization.

In contrast, for the DPCTGAN model variant, following the work of Giomi et al. [13], we preprocessed the continuous variables by splitting them into 50 equally spaced bins using the min-max value range of the real variables. Instead of directly extracting from real continuous data, models are trained on binned continuous and labeled encoded categorical datasets to subsequently learn, generate, and save synthetic datasets for each model. The synthetic datasets produced by using the DPCTGAN synthesizer are in bin label formats. To enable fidelity and privacy risk assessment, the labeled synthetic dataset bin has been further decoded by randomly picking values with the same bin edges (i.e., used to encode) to align with the real dataset format. The generative model incorporated a differential privacy mechanism to ensure the privacy guarantee; however, binning on continuous features potentially increases the risk of individuals' data disclosure because of the requirement of minimum and maximum ranges of original data information to establish the bin edge [32].

4.3 Evaluation Metrics

This section briefly discusses the different evaluation metrics performed to compute the data fidelity and the privacy risk of generated synthetic data.

4.3.1 Data Fidelity Assessment

The assessment of synthetic data fidelity is a complex task that can't be achieved by using a single metric. This work has adopted a set of quantitative evaluation metrics for both numerical and categorical relationships. To assess distributional dissimilarity between synthetic tabular data and real data across multiple runs, the study adopted the Hellinger distance metric. Further, the Total Variation Distance and the Kolmogorov-Smirnov (KS) statistic have been implemented to measure marginal distribution similarity. Additionally, for numerical variables, pairwise correlation with Frobenius norms has been computed from the difference between Spearman correlation metrics of real and generated data. To measure data fidelity assessment of categorical variables, the Cramer's V based on chi-square statistics has been performed.

Hellinger Distance

The Hellinger distance is employed to quantify how similar probability distributions of real and generated dataset's variables are, providing a bounded metric (0,1); a Hellinger distance of 0 means exactly the same probability distribution, and, in contrast, 1 indicates complete dissimilarity between real and synthetic data. In this work, this metric has been used to assess the data fidelity of the synthetic data generated by each synthesizer for both numerical and categorical datasets.

Mathematically, let us assume P and Q represent two discrete probability distributions of a given attribute in real and synthetic tabular datasets. The value of the Hellinger distance between P and Q, i.e., H(P, Q), is calculated using the formula

provided in [48]:

$$H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{k}(\sqrt{p_i} - \sqrt{q_i})^2}$$

Here, k stands for the count of discrete bins for the attributes, and p_i and q_i are probabilities in the i-th bin of distributions of both P and Q.

To evaluate the run-wise model stability and fidelity on synthetic datasets generated for each model variant, the Hellinger distance has been calculated for each variable between real and generated data. For each independent run and model, we compute variable-wise Hellinger distance in marginal 1D distribution metrics; as a result, we get 30 Hellinger distances for each variable, i.e., 30×12 metrics. Summarizing the statistics, the mean and standard deviation have been calculated among 30 Hellinger distance values for each variable. For cross-model comparison, the distribution of Hellinger distance scores for each variable across all models simultaneously has been visualized in a grouped box plot. The group boxplot illustrated the degree of similarity between the distributions of real and synthetic data across models per variable. In addition, assess the effect of privacy budget variation on distribution fidelity.

TVComplement and KSComplement Metrics

The TVComplement assesses fidelity through marginal distribution comparison [42]. It is utilized by the Synthetic Data Vault (SDV) ecosystem. This metric is derived from the Total Variation Distance (TVD) [49] that measures the difference between the probability distribution of real R and synthetic S data over a discrete domain [42]. Mathematically, the value of TVD was calculated using the following formula presented in [42].

$$\delta(R, S) = \frac{1}{2} \sum_{\omega \in \Omega} |R_\omega - S_\omega|$$

The normalized probabilities of observing categorical or binned values ω in real and synthetic datasets are indicated by R_ω and S_ω, respectively. Then TVComplement scores are computed as:

$$\text{TVC} = 1 - \delta(R, S)$$

This metric is applicable for both numerical and categorical variables. Where, for categorical variables, the probability distribution is calculated based on the relative frequency of each unique category. On the other hand, for numerical variables, this metric is employed with other metrics, such as Kolmogorov–Smirnov statistic.

In this work, we utilized two complementary metrics from the SDMetrics library, such as TVComplements and KSComplements, to evaluate the quality of generated data for both categorical and numerical variables. The TVComponents metrics have been computed to quantify similarity between the probability distribution of categorical features in real and synthetic data. A TVC score of 0.0 means the perfect dissimilarity, and a score of 1.0 indicates the perfect alignment between real and synthetic distributions. Next, for numerical variables, the KSComplement metric has been computed to measure the closeness between the empirical cumulative distribution of real and synthetic features, with higher values indicating strong similarity. As a result, we got feature-wise TVC scores and KSComplemts values across 30 independent synthetic datasets generated by each model and visualized them using box plots to illustrate the fidelity of generated synthetic data over multiple generations. To analyze the run-to-run model variability, we computed the attribute-wise average mean and standard deviation of TVComplemets and KScomplement values across multiple runs for each model.

Pairwise Correlation

The pairwise correlation analysis metric measures how well the generated synthetic data preserved the pairwise relationship between features compared to the real dataset. For continuous variables, the Spearman correlation matrix [50] on synthetic datasets was generated across 30 independent runs per model. Next, for categorical features, we computed Cramér's V, [51], [52], as a normalized chi-squared-based statistic to calculate the pairwise association between real and synthetic datasets defined in [53].

$$V = \sqrt{\frac{\chi^2/N}{\min(k_1 - 1, k_2 - 1)}}, \tag{4.1}$$

Where the chi-square statistic is denoted as χ^2, the total sample size as N, and the number of categories in each variable as k_1 and k_2.

The study computes mean absolute difference metrics by subtracting the synthetic data correlation metrics from the reference metrics and taking the absolute values. These metrics have been used to capture the local fidelity loss in replicating pairwise feature dependencies. To obtain the feature-wise fidelity, we computed the averaged means of the absolute difference metrics obtained across 30 independent synthetic dataset runs for each model.

Additionally, we have been implementing annotated heatmaps with numerical annotations for the mean difference metrics for each model, which visualize the per-feature pairwise fidelity between real and synthetic data correlation metrics and highlight relationships that models mostly preserved. To get the per-feature discrepancy vector, a grouped bar chart has been used that eases a side-by-side comparison of each model's performance per feature with distinct colors. In summary, by aggregating over all 30 independent runs, the group bar chart provides both statistical fidelity and a fiddly pattern at all feature levels.

4.3.2 Privacy Risk Evaluation

In this section, we briefly look at the singling-out privacy evaluation that was performed to evaluate the privacy guarantee of synthetic data.

Singling-Out

Singling-out risk estimates the external adversary's ability to extract a uniquely individual data set by exploiting specific attribute combinations in synthetic data, even in the absence of explicit identifiers. This work adopted the attack-based privacy framework proposed in [13] by utilizing prebuilt, reproducible, and validated codes `SinglingOutEvaluator` hosted on GitHub [54] in our model evaluation pipeline. The main goal of an attack-based singling-out risk evaluator is to estimate the extent to which the generated data preserves unique individual information from the real dataset. Singling-out risk is measured by outputting a prediction from the generated dataset that is likely to single out rare records and successfully isolate individuals in the training data beyond what would be expected from population-wise rareness alone. The output from this is expressed as calibrated risk metrics that are denoted by R, and the result lies between 0 and 1. A value of R equal to zero indicates that there is no measurable singling out beyond baseline population uniqueness, and a value of R equal to 1 shows the maximum leakage in generated tabular data. In this privacy risk measurement, it is assumed that the attacker has full access to the generated synthetic data.

The single-out evaluation implements two different algorithms that show different adversarial scenarios: the univariate algorithm and the multivariate algorithm. First, the univariate algorithm evaluates the singling-out risk by examining each and every feature's value separately to estimate the probability that an adversary can single out rare values of individual attributes. For categorical, inspect the unique values, and for numerical, observe the outlier values, both at min and max. Second,

a multivariate algorithm constructs the complex predicate by using the logical conjunction of more than one attribute-value pair randomly from generated data. The algorithm estimates the risk using a combination of multiple attributes simultaneously, modeling a stronger adversary who exploits correlation among up to a specific number of attributes to uniquely identify attributes.

In this work, we run the `SinglingOutEvaluator` class instantiated to compare synthetic data generated by each model across multiple independent runs against the original training data and the test data. Here, the singling-out risk evaluator class first attempted a multivariate attack model and assessed singling-out risk based on the logical conjunction of three attributes. If multivariate evaluation has failed due to computational or data sparsity issues, then the evaluator goes to a univariate model and analyzes a singular attribute independently. For each run, singling out risk estimates sets a 95% confidence interval. Furthermore, for each query generation, we set the number of attacks to five hundred, represented by $n_{\text{attack=500}}$, balancing computational feasibility with sufficient statistical power to reliably estimate the risk metrics. In this privacy risk, an attack is considered successful if the adversary's ability to correctly single out individual information in the original train dataset i.e train_df based on the information inferred from the generated synthetic data significantly exceeds the baseline success rate achieved on an independent control dataset i.e test_df. In contrast, if the success rate on train_df is lower than or equal to the success rate on test_df or a similar baseline, then the attack is considered unsuccessful or a failure, which means it does not reveal meaningful individual privacy data.

To provide a comprehensive picture of the privacy risk distribution, the results obtained from multiple synthetic datasets for each generator have been aggregated and saved into a separate data frame. We computed the aggregate mean and standard deviation values, then used box plots to visualize the distribution of singling-out

risk for each method across multiple runs. The lower value of single-out risk shows a higher degree of privacy protection; in contrast, a higher value means greater vulnerability to a recorded single-out attack. In conclusion, this study of privacy risk evaluation provides an empirical assessment of the synthetic tabular dataset's privacy leakage. The results obtained from these metrics enable a direct comparison of privacy guarantees among synthetic data generation models and provide the guidelines for the model and privacy parameter section for practical implementation.

5 Results and Discussion

This chapter consists of a comprehensive discussion of the key findings obtained from various quantitative metrics that we implement in this work. These metrics have been adopted to evaluate the model variability, data fidelity, and privacy risk of synthetic data generated over multiple runs for both the CTGAN and DP-CTGAN model variants. For each model, we had 30 run-wise synthetic datasets, each consisting of 20,000 data with 12 features: 5 numerical, 6 categorical, and 1 target variable.

5.1 Data Fidelity Assessment Result

This section presents detailed results that we obtained from different quantitative metrics used for the univariate and bivariate analyses to assess the quality of synthetic tabular healthcare data generated across the multiple runs. In this work, we have performed univariate analysis using Hellinger distance metrics, total variation distance, and the Kolmogorov-Smirnov (KS) statistic. Additionally, we have computed bivariate analysis using pairwise correlation metrics.

5.1.1 Hellinger Distance Result

The Hellinger distance is computed to quantify the dissimilarity between probability distributions of real and synthetic dataset features, resulting in a bounded metric (0,1). A Hellinger distance score closer to 0 indicates better fidelity, and closer to

1 means high dissimilarity. This study computed the feature-wise average mean and standard deviation of the Hellinger distance between real and synthetic data distributions across each model for both continuous and categorical variables.

Variable	CTGAN		DPCTGAN $\epsilon = 1$		DPCTGAN $\epsilon = 10$		DPCTGAN $\epsilon = 100$	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
age	0.19	0.02	0.21	0.00	0.16	0.01	0.07	0.01
height	0.11	0.03	0.55	0.00	0.43	0.01	0.10	0.01
weight	0.10	0.03	0.52	0.01	0.38	0.01	0.08	0.01
ap_hi	0.23	0.02	0.57	0.01	0.46	0.01	0.24	0.00
ap_lo	0.13	0.03	0.65	0.00	0.53	0.01	0.11	0.01

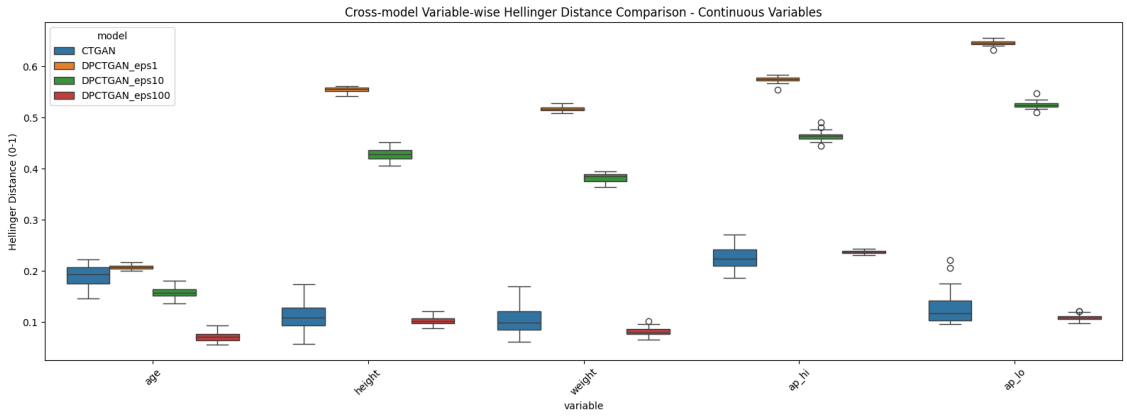
Table 5.1: Mean and Standard Deviation of the Hellinger distance for Continuous Variables

Variable	CTGAN		DPCTGAN $\epsilon = 1$		DPCTGAN $\epsilon = 10$		DPCTGAN $\epsilon = 100$	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
gender	0.04	0.02	0.71	0.01	0.73	0.01	0.71	0.00
cholesterol	0.08	0.04	0.75	0.00	0.74	0.00	0.75	0.00
gluc	0.09	0.04	0.82	0.00	0.83	0.00	0.82	0.00
smoke	0.09	0.04	0.18	0.01	0.02	0.01	0.04	0.01
alco	0.09	0.03	0.20	0.00	0.02	0.00	0.02	0.01
active	0.05	0.04	0.01	0.01	0.04	0.01	0.01	0.01

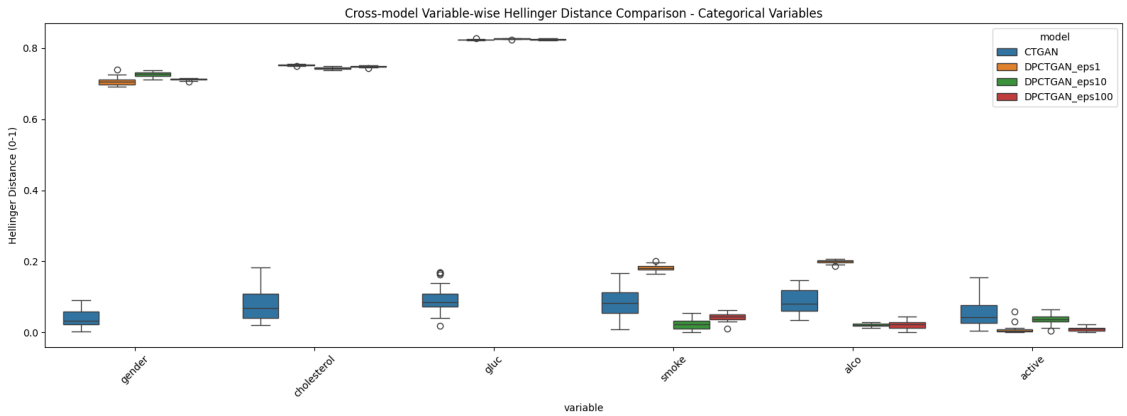
Table 5.2: Aggregate Mean and Standard deviation of the Hellinger distance for Categorical Variables

For continuous variables, as shown in the table 5.1, the CTGAN model demonstrated satisfactory performance with a mean Hellinger distance lower than 0.24. However, the DPCTGAN model at $\epsilon = 100$ has shown excellent fidelity with the lowest mean Hellinger distance across almost all variables. In contrast, the differential privacy integrated model at $\epsilon = 1$ has shown a higher mean Hellinger distance value for variables like height, weight, ap_hi, and ap_ho above 0.50 to 0.65, implying lower fidelity. As the value of ϵ increases, the mean Hellinger distance decreases, showing a direct relationship between privacy budget and data quality in the generated healcare tabular data.

On the other hand, for categorical variables, as shown in the table 5.2, the CTGAN model with the lowest mean Hellinger distance indicates better fidelity. While the model at $\epsilon = 1$ shows the higher mean Hellinger distance value lying between 0.7 and 0.8 for variables like gender, cholesterol, and glucose, it shows that synthetic data distributions are notably different in those categories than real data. The model at $\epsilon = 10$ enhanced the overall fidelity substantially with variables like active, smoke, and alcohol, producing values closer to 0.03. This pattern shows that model training with differential privacy parameters captures the variables' distributions, such as age, smoking, alcohol, and activity, more reliably even with strong privacy. Oppositely, variables like gender, glucose, and cholesterol with high mean Hellinger distance under lower ϵ show the poor synthetic fidelity for these variables.



(a) Continuous Variables



(b) Categorical Variables

Figure 5.1: Grouped Box plot Cross-Model Variable-Wise Hellinger Distance

In addition to the statistical summary of the Hellinger distance of each variable, a combined cross-model group chart for continuous and categorical variables has been generated. Figure 5.1 has shown the distribution of the Hellinger distance score across multiple runs for each variable over all models with different colored box plots. The CTGAN model has a lower Hellinger distance for all variables across multiple runs, showing the highest fidelity among the models. The DP-CTGAN models show a higher Hellinger distance, especially for certain variables like gender, cholesterol, and glucose, showing the lower fidelity under stricter privacy. However, as ϵ increases, fidelity increases for the DPCTGAN model as its box plots move close to non-privacy models' values for the variables increased. In conclusion, the DPCTGAN variant synthetic data generation models with lower standard deviation in Hellinger distance and narrow box plots reflect high stability and low variability. That means the model consistently reproduces synthetic data for both continuous and categorical features over multiple runs.

5.1.2 TVComplement and KSComplement Metrics Result

These metrics are utilized to quantify the fidelity and similarity between the feature probability distributions of real and synthetic datasets. Below, table 5.3 provides the average mean and standard deviation result for both numerical and categorical features obtained from the TVComplemts and KSComplemts metrics of 30 independent synthetic datasets across each model: CTGAN and DPCTGAN with varying privacy budgets. Additionally, to complement the statistical summary, a box plot as shown in figure 5.2 illustrated the variable-wise TVComplement and KSComplement of both continuous and categorical variables for each model across multiple runs.

Feature	CTGAN		DPCTGAN $\epsilon = 1$		DPCTGAN $\epsilon = 10$		DPCTGAN $\epsilon = 100$	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
TVComplement								
gender	0.946	0.035	0.348	1.67e-16	0.348	1.56e-16	0.348	1.56e-16
cholesterol	0.908	0.061	0.244	7.57e-17	0.244	7.57e-17	0.244	0.001
gluc	0.901	0.048	0.144	6.48e-17	0.128	0.003	0.138	0.005
smoke	0.917	0.044	0.807	0.011	0.981	0.012	0.969	0.007
alco	0.930	0.031	0.810	0.006	0.987	0.002	0.988	0.006
active	0.936	0.049	0.991	0.013	0.958	0.015	0.989	0.007
KSComplement								
age	0.904	0.032	0.843	0.010	0.907	0.020	0.969	0.012
height	0.911	0.040	0.565	0.007	0.700	0.016	0.928	0.008
weight	0.897	0.032	0.577	0.009	0.703	0.013	0.947	0.009
ap_hi	0.931	0.023	0.577	0.008	0.655	0.015	0.742	0.010
ap_lo	0.950	0.028	0.489	0.009	0.557	0.015	0.654	0.006

Table 5.3: Variable-wise TVComplement and KSComplement Across Models

TVComplement

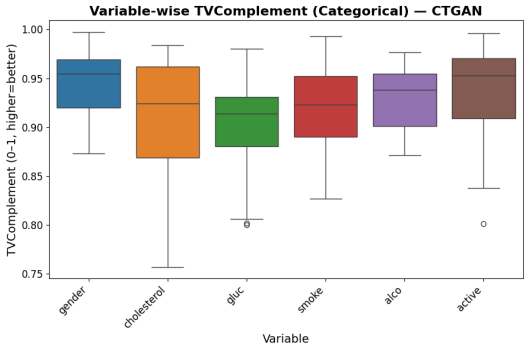
As shown in table 5.3, the CTGAN model has a consistently higher variable-wise TVComplements mean value close to 0.9 or higher, which indicates higher fidelity across multiple independent runs. Additionally, the boxplot shown in Figure 5.2 visualizations with narrow interquartile ranges and high medians for nearly all features also support the reproducibility of the high fidelity for CTGAN models. Similarly, the DPCTGAN models with different ϵ show the lower fidelity, especially with a strict privacy budget $\epsilon = 1$. As privacy budgets increase, a clear and distinct privacy–fidelity trade-off can be illustrated. At lower ϵ, it illustrates a wider boxplot and lower median, indicating reduced fidelity; with higher ϵ, the tightened boxes with shifts up visually show the improvement in synthetic data quality.

KSComplemets

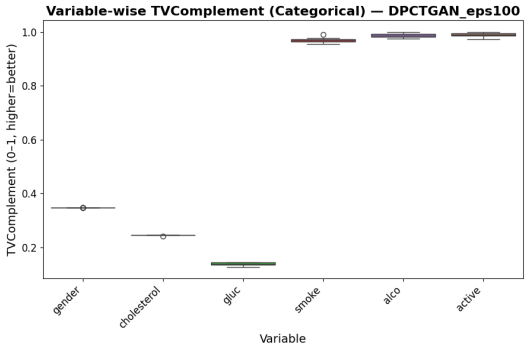
As shown in table 5.3, CTGAN models with a higher mean value closer to 0.9 and a lower standard deviation show better quality preservation of numerical variables as well as low run-to-run variability of models. In contrast, DPCTGAN variants show a clear privacy-fidelity tradeoff: at lower ϵ, the substantially lower mean values

indicate lower fidelity, and increasing the value of ϵ improves fidelity with the move from a moderate to a high KSComplements score. The model at $\epsilon = 100$ is shown to be even better than CTGAN values for some variables, like smoke, alco and active.

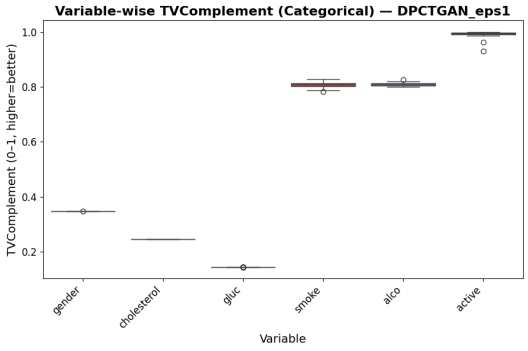
In conclusion, results from both the statistics summary and graphical visualizations show that variables like smoke, alcohol, activity, and age with higher TVC and KSC mean values correspond to better fidelity, whereas variables like height, weight, ap_hi, and ap_lo with comparatively lower values indicate moderate fidelity. In contrast, variables like gender, glucose, and cholesterol with lower values indicate poor synthetic fidelity. The model with a lower standard deviation indicates higher model consistency across multiple runs.



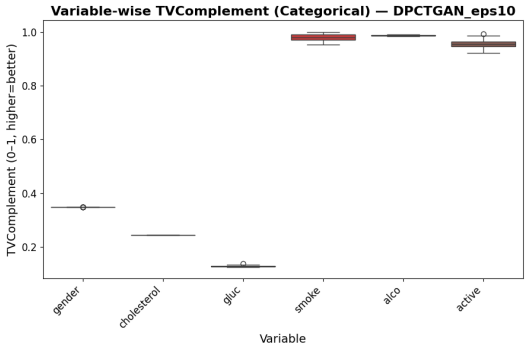
(a) CTGAN TVC



(b) DPCTGAN_ep1 TVC

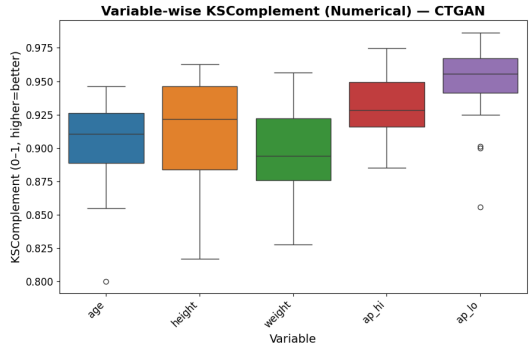


(c) DPCTGAN_ep10 TVC

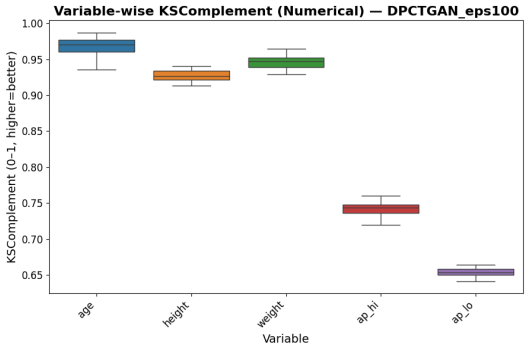


(d) DPCTGAN_ep100 TVC

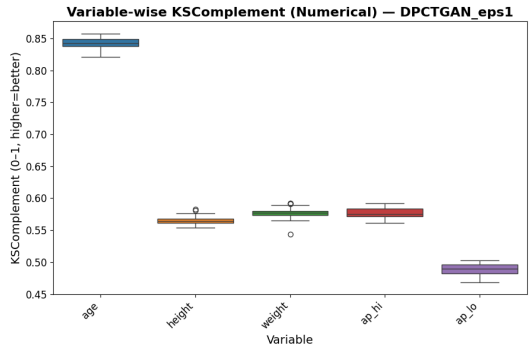
Figure 5.2: Box plots of Model Variants: Variable-wise TVComplement



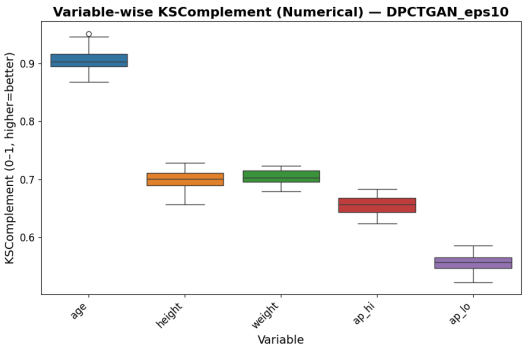
(a) CTGAN TVC



(b) DPCTGAN_ep1 TVC



(c) DPCTGAN_ep10 TVC



(d) DPCTGAN_ep100 TVC

Figure 5.3: Box plots of Model Variants: variable-wise KSComplement.

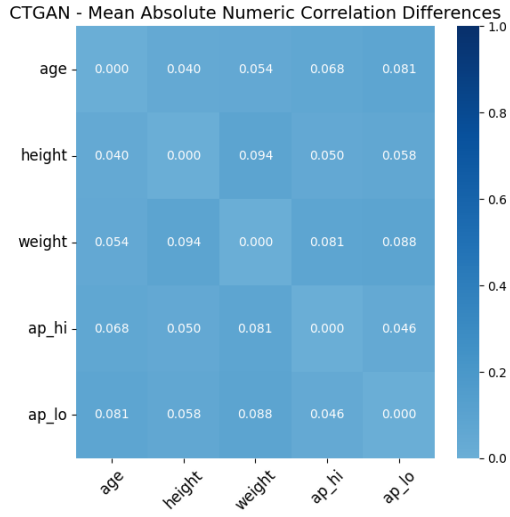
5.1.3 Pairwise Correlation Results

The pairwise correlation is computed to measure how well generated synthetic datasets preserve statistical relationships between features that exist in the original datasets. The study computed the feature-wise mean absolute correlation differences across each model for both continuous and numerical features: Figure 5.4 presents the mean absolute correlation differences for numerical variables, while figure 5.5 shows the mean absolute correlation differences for categorical variables.

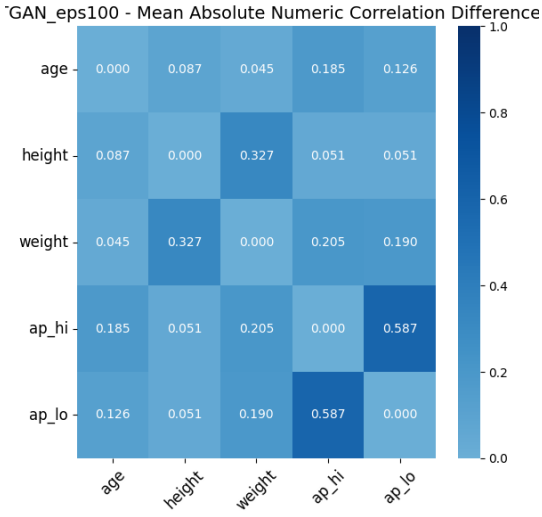
for continuous variables, as shown in figure 5.4, the CTGAN model with a low mean absolute difference range, i.e., 0.04 to 0.09, had better reproduced the joint structural relation among continuous variables. In contrast, the differential privacy

integrated model variants show a larger distortion for many pairs, like ap_hi and ap_lo (0.58 to 0.73) and weight and height (0.28 to 0.32). That means differential privacy-integrated models show the substantial distortion of medically important variables in comparison to original healthcare data. As the value of ϵ increases from 1 to 100, we can note a clear improvement in pairwise fidelity, with heatmap values decreasing towards those of CTGAN, but some numerical pairs, such as ap_hi and ap_lo, still show the high discrepancies.

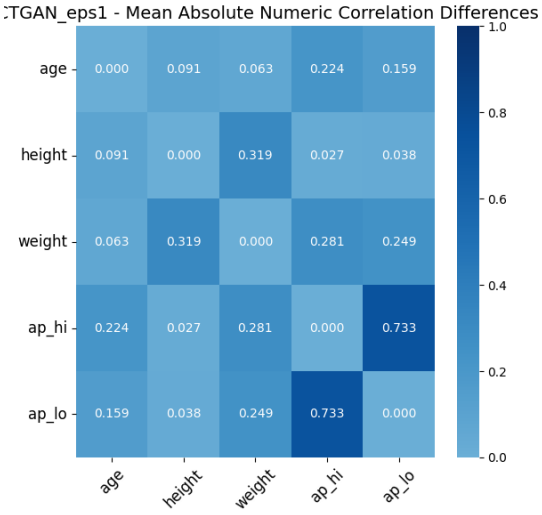
For categorical variables, as shown in 5.5, the CTGAN model has shown a relatively low mean absolute difference range. That means this model has been able to better preserve the structural relationship of categorical variables closer to real data. The DP integrated model at $\epsilon = 100$ better reproduced structural relationships close to real data, as compared to the other DP CTGAN variants. The model at $\epsilon = 1$ and 100 has demonstrated similar performance, with both models showing worse results across three pairs, i.e., gender and smoke, cholesterol and glucose, and smoke and alcohol.



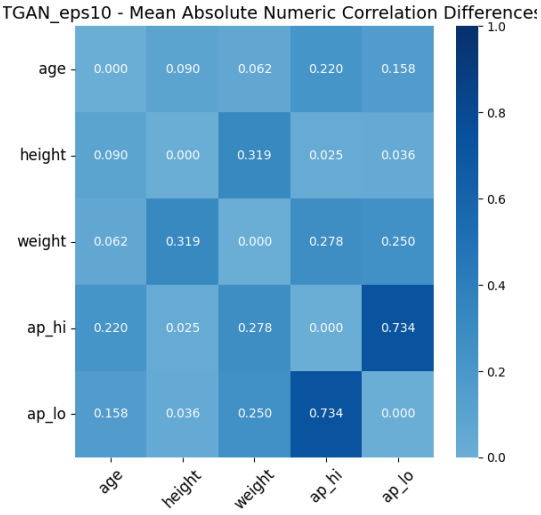
(a) CTGAN



(b) DPCTGAN_ep100

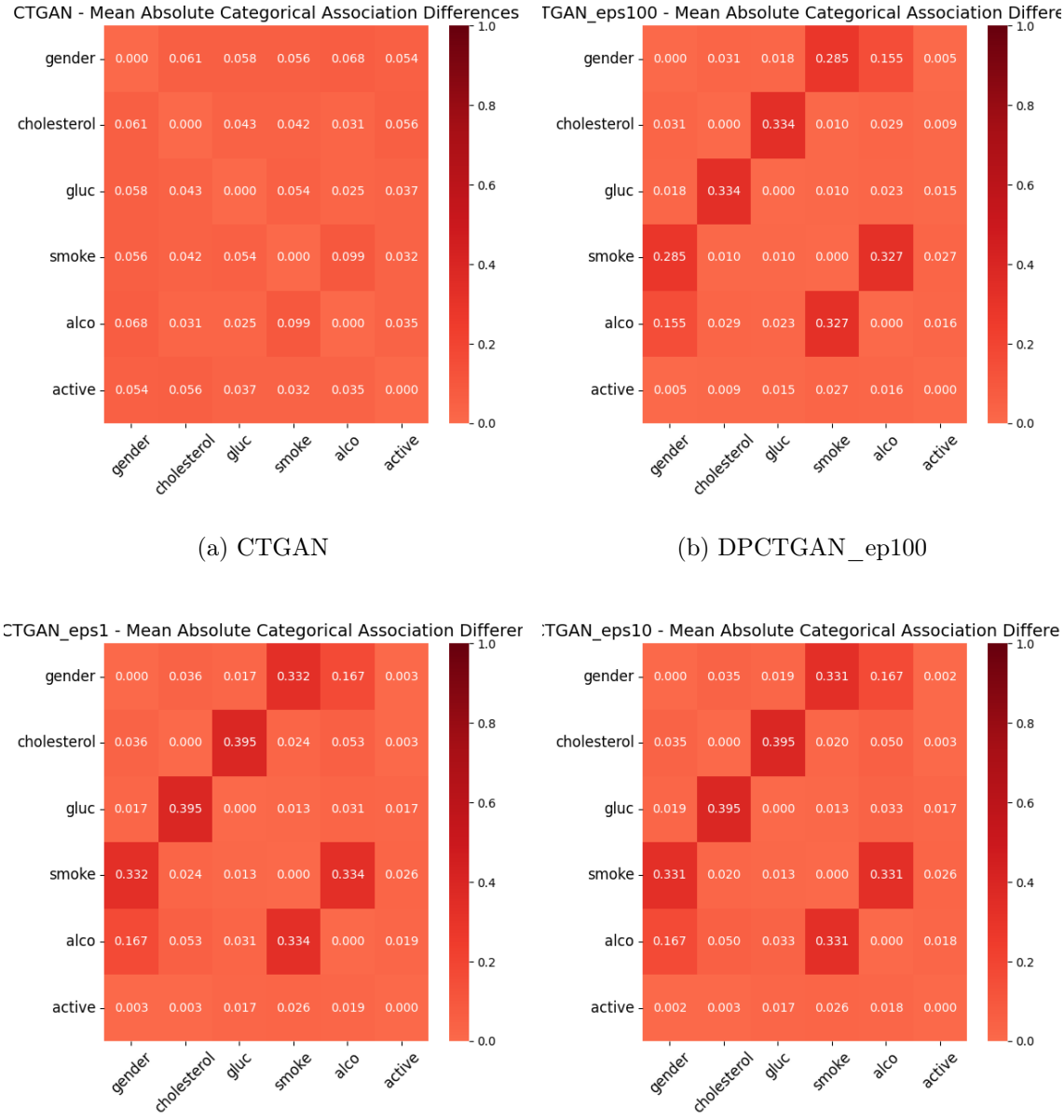


(c) DPCTGAN_ep1



(d) DPCTGAN_ep10

Figure 5.4: Numerical Feature-wise Mean Absolute Correlation Differences Across Models



(a) CTGAN

(b) DPCTGAN_ep100

(c) DPCTGAN_ep1

(d) DPCTGAN_ep10

Figure 5.5: Categorical Feature-wise Mean Absolute Correlation Differences Across Models

Additionally, the group chart has been used for demonstrating feature-wise comparison of average pairwise correlation discrepancies across each model. As shown in figure 5.6, the CTGAN model has shown superior average per-feature fidelity for both categorical and numerical features. Whereas DPCTGAN model variants

lag behind, with the largest error clustered around ap_hi and ap_lo, smoke, and cholesterol.

In conclusion, both the annotated heatmap and group chart plots provide a clear assessment of how well synthetic tabular datasets across multiple independent runs for each model preserve the underlying inter-feature relations that exist in the real datasets. Here, the CTGAN model outperforms all DPCTGAN model variants with the lowest feature-wise mean absolute correlation. The DPCTGAN models at lower epsilon show increased distortion in key pairwise associations. In contrast, the DPCTGAN model trained with a higher epsilon shows improvement in fidelity but is still not close to the non-privacy model. These results introduce a differential privacy trade-off that increases privacy protection and reduces the model's ability to reassemble complex pairwise dependencies.

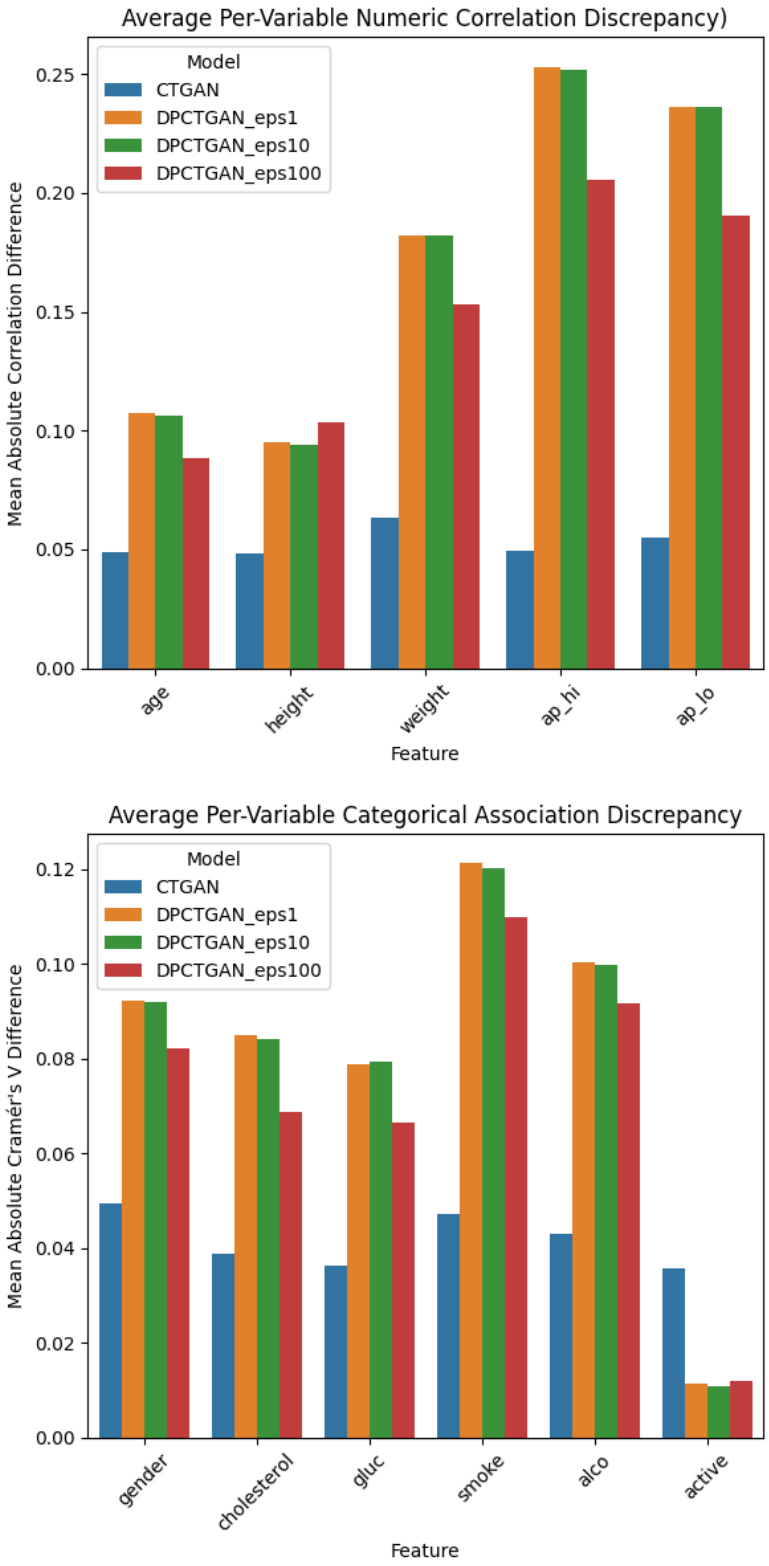


Figure 5.6: Grouped BarPlot: Numerical and Categorical Features Across Models

5.2 Privacy Risk Measurement Result

In this section, we show the performance of generators, CTGAN and DPCTGAN, with varying levels of differential privacy, using singling-out risk evaluation metrics under an attack-based privacy framework. Figure 5.7 shows the distribution of singling-out risk across 30 independent synthetic data points. Table 5.4 provides the mean and standard deviation values of each generator across multiple runs.

For distribution of singling-out risk across multiple runs, a boxplot visualized with different colors for each model was used. The distribution from the CTGAN model shifted upward with a tight interquartile range that indicates the persistent high privacy risk. On the other side, the DPCTGAN model at lower epsilon values shows the tightly clustered, lower-risk distribution. It has been shown that the model provides both strong and stable privacy protection. As privacy budget values increase, the distribution of singling out risk spread and median also increases, which indicates more privacy exposure and less predictability.

Model Type	mean	std	min	max	count
DP-CTGAN ϵ=1	0.038359	0.014054	0.010356	0.060801	30
DP-CTGAN ϵ=10	0.045024	0.015210	0.010239	0.068662	30
DP-CTGAN ϵ=100	0.100839	0.021027	0.050539	0.146106	30
CTGAN	0.131835	0.016744	0.102893	0.175945	30

Table 5.4: Aggregated Statistics Summary of Singling Out Risk

The CTGAN model, with the highest aggregate mean of singling out risk closer to 0.13, indicates the model is more prone to a successful re-identification attack compared to others or weak privacy. Further, the relatively low standard deviation indicates that vulnerability to singling out attacks is consistent across multiple runs. It makes this model both reliable and reproducible but still highly vulnerable in terms of accuracy.

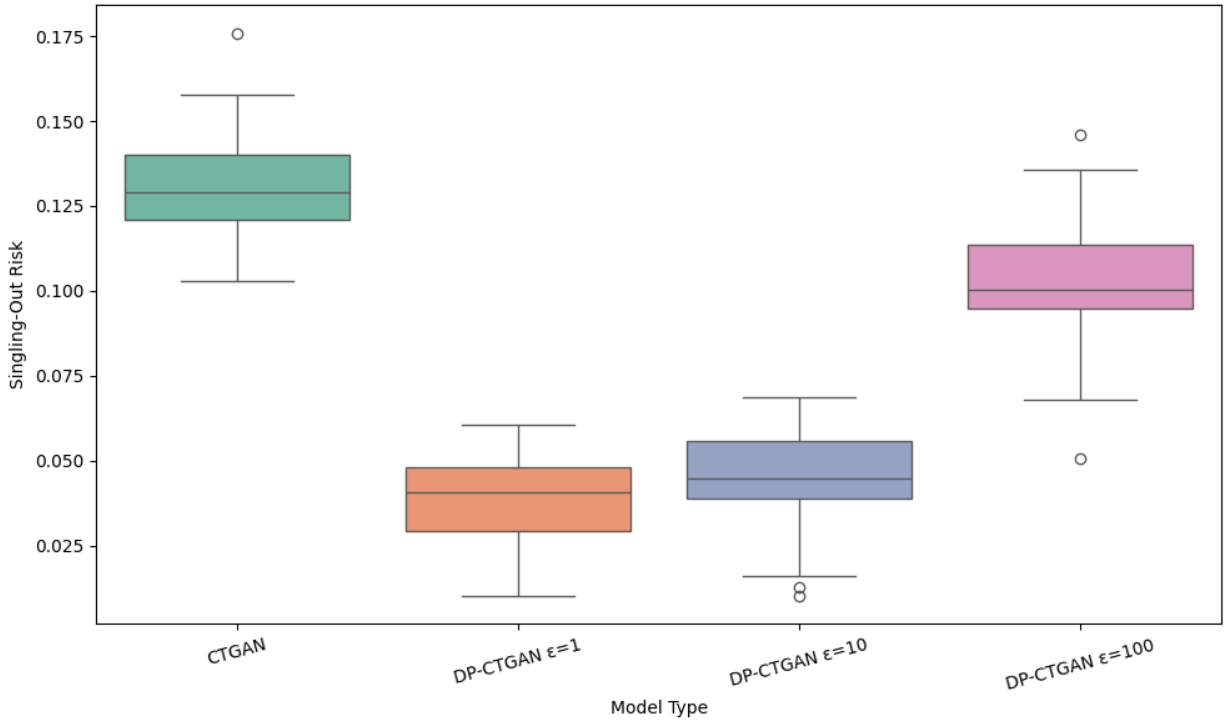


Figure 5.7: Boxplot: Distribution of Singling-Out Risk for Each Model

The model with the differential privacy mechanism has shown a higher degree of privacy with a consistent lower average, singling out the risk mean value. The DPCTGAN model at $\epsilon = 1$, with a lower mean value of 0.03 compared to other models, makes it significantly harder for the adversary to identify individual data. In addition to that, the smaller standard deviation shows the models low-risk stability across multiple runs. As the value of the privacy budget increases from $\epsilon = 1$ to $\epsilon = 100$, a clear privacy trade-off is visible: the mean risk increases and the distribution becomes more vulnerable. It shows that the weaker the privacy parameter, the lesser the model stability.

Summary of Privacy-Fidelity Metrics Across Models

This section consists of the summary results, i.e., the mean and standard deviation scores that have been obtained from the various privacy and fidelity metrics utilized

across each model, as demonstrated in table 5.5.

Model	SinglingOutRisk		TVC		KSC		Hellinger		MD.Spearman		MD.Cramér'sV	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
CTGAN	0.132	0.017	0.923	0.019	0.919	0.015	0.109	0.012	0.053	0.014	0.042	0.011
DPCTGAN $\epsilon = 1$	0.038	0.014	0.558	0.003	0.610	0.003	0.470	0.002	0.175	0.002	0.082	0.001
DPCTGAN $\epsilon = 10$	0.045	0.015	0.608	0.003	0.704	0.007	0.394	0.002	0.174	0.002	0.081	0.001
DPCTGAN $\epsilon = 100$	0.101	0.021	0.613	0.002	0.848	0.005	0.269	0.002	0.148	0.003	0.072	0.002

Table 5.5: Summary of Privacy-Fidelity Metrics Across Models

CTGAN The highest singling-out risk (i.e., weak privacy) compared to the other model shows that this model's privacy has not been protected by default. In contrast, the TVComponents and KSComponents metrics with height mean values closer to 1 indicate high fidelity, meaning that synthetic data closely resemble the distribution of real data. Additionally, the lowest mean absolute difference from both Spearman and Cramér's V shows the model's ability to capture pairwise relationships between features relative to the real dataset, indicating better model fidelity. The lowest standard deviation shows the consistency of models across multiple runs.

DPCTGAN at $\epsilon = 1$ The significantly lower singling-out risk compared to other models shows that the model has strong privacy protection. Oppositely, the TVComponents and KSComponents metrics with the lowest mean values and the Hellinger distance with the highest mean clearly show a privacy-fidelity trade-off, with reduced fidelity. The lowest standard deviation indicates the low run-to-run variability, that there is no such big fluctuation in generated data distribution across multiple runs.

DPCTGAN at $\epsilon = 10$ Almost the same singling out of risk results from the model at $\epsilon = 1$ indicates strong privacy protection. In contrast, the results of the TVComponents, KSComponents, and Hellinger distance are better than $\epsilon = 1$ but lower than $\epsilon = 100$, and the non-private method indicates intermediate fidelity. In addition, relatively higher average mean absolute deviation values for both Spearman and Cramer's V show that the model poorly preserves pairwise relationships

between features relative to the real dataset, indicating poor fidelity.

DPCTGAN at $\epsilon = 100$ the higher singling out risk than other DPCTGAN variant models, meaning weaker privacy. Whereas the results from the evaluation metrics are better compared to DPCTGAN variants, showing the usefulness of data quality with a weaker privacy guarantee.

In conclusion, the CTGAN model produced synthetic tabular datasets that better preserve marginal distribution and pairwise relationships between features relative to the real dataset. In contrast, the DPCTGAN models generated synthetic tabular datasets under varying privacy guarantees controlled by the value of ϵ. These models are able to better preserve the marginal distribution of features like age, height, smoking, weight, alcohol, and activity than ones like gender, cholesterol, and glucose, indicating poor synthetic fidelity for these variables. When strict privacy $\epsilon = 1$ 10 , the DPCTGAN variant shows a privacy-fidelity trade-off: a strong privacy guarantee with very low fidelity.

6 Conclusion

Accessing high-quality tabular healthcare data, including cardiovascular, heart disease UCI, and diabetes datasets, is an ongoing challenge in the medical and clinical research field due to data sensitiveness, strict privacy regulation, the risk of personal information disclosure, and the limited availability of real-world tabular datasets. Synthetic data generation techniques are considered a promising method for easing data sharing and reducing disclosure risk. These methods generate more realistic artificial datasets that preserve the key statistical properties of real datasets. However, the variability of these models across multiple independent runs is still not fully discovered. To address this gap, this work systematically evaluates multiple synthetic generative models with and without the integration of differential privacy, data fidelity, and run-to-run model variability with varying privacy budgets across multiple training runs. To carry this out, the thesis work has been split into two parts. The first part focused on generating tabular synthetic data using GANs-based models. The second part is using quantitative evaluation metrics and privacy risk assessment to evaluate fidelity, run-to-run model variability, and privacy characteristics of the generated synthetic tabular data.

The first part of this thesis work started with a literature review on GAN-based synthetic data generation methods with respect to their advantages and disadvantages and a review of the concept of differential privacy. Afterward, we select the GAN-based models, such as the standard CTGAN and the differential privacy inte-

grated version of the CTGAN with varying privacy control parameters, for synthetic tabular healthcare data generation due to their proven ability to handle tabular healthcare datasets. The result from this part provides information about how the synthetic data generation model behaves across multiple runs.

The second part of the thesis work quantifies the generated healthcare tabular synthetic dataset's fidelity, run-to-run model variability, and privacy risk across multiple runs. To quantify feature-wise quality and run-to-run model variability, different quantitative evaluation metrics have been performed. The result of these quantitative metrics provides information about feature-wise data quality and model variability across multiple runs. For privacy assessment, the attack-based singling-out risk has been employed. The result from this set of metrics provides information on privacy leakage and adversaries' ability to estimate uniquely found individual data records, even when direct identifiers are removed.

The primary focus of this thesis work is the analysis of variable-wise data quality, privacy risk assessment, and run-to-run model variability across multiple independent runs, which has been addressed. This work is useful for guidelines for generating synthetic tabular data and analyzing data quality, privacy, and model stability. However, there is still significant scope for future work to explore other aspects to enhance and validate this study. The adopted privacy framework in [13] defined multiple privacy aspects, including singling out, likability, and inference risk; this study only considered singling out. This suggests that extending privacy risk assessment to likability and inference risk is a significant avenue for future research. While quantifying the DP integration model's performance, it shows the variation in model performance across variables, especially categorical variables. For example, the models have shown better performance for some categorical variables like smoking, alcohol, and activity, while they have shown lower performance for other categorical variables like gender, cholesterol, and glucose. This suggests that the ex-

tended investigation of the distribution of these variants and the detailed statistical analysis to understand the factor behind this variation could be valuable for future work.

7 Disclosure of AI Assistance in the Thesis

While doing my thesis work, I utilized publicly available AI-based tools for the thesis outline, organization, and refinement processes. Firstly, I had used Perplexity AI (https://www.perplexity.ai) to support me in structuring my thesis outline and to resolve some issues that I faced in the model training code. Thereafter, QuillBot (https://quillbot.com) was used to test the correct typing and grammar. Additionally, in the literature review part, I had taken a little bit of assistance from ChatPDF (https://www.chatpdf.com) to document and summarize a lengthy research article. Moreover, ChatGPT (https://www.chatgpt.com) was used to help me when I became confused regarding the outline of thesis work. I have never used these tools to copy full content or paragraphs generated directly from them. But in some cases, I have used the QuillBot tool solely as a reference to maintain the flow of sentences and paraphrase.

References

[1] P. A. Osorio-Marulanda, G. Epelde, M. Hernandez, I. Isasa, N. M. Reyes, and A. B. Iraola, "Privacy mechanisms and evaluation metrics for synthetic data generation: A systematic review", *IEEE Access*, vol. 12, pp. 88 048–88 074, 2024. DOI: 10.1109/ACCESS.2024.3417608.

[2] R. B. Nelsen, *An Introduction to Copulas*. New York, NY, USA: Springer, 2006.

[3] IWCO, *Synthetic data generation: A comprehensive overview*, [Online]. Available: https://www.iwco.co/synthetic-data-generation-a-comprehensive-overview/, 2025.

[4] D. B. Rubin, "Statistical disclosure limitation", *Journal of Official Statistics*, vol. 9, no. 2, pp. 461–468, 1993.

[5] I. Goodfellow et al., "Generative adversarial nets", in *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[6] C. Doersch, *Tutorial on variational autoencoders*, arXiv:1606.05908, 2016.

[7] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan", in *Advances in Neural Information Processing Systems*, vol. 32, Curran Associates, Inc., 2019.

[8] M. Hernandez, G. Epelde, A. Alberdi, R. Cilla, and D. Rankin, "Synthetic tabular data evaluation in the health domain covering resemblance, utility,

and privacy dimensions", *Methods of Information in Medicine*, vol. 62, no. S01, e19–e38, 2023. DOI: 10.1055/s-0042-1760247.

[9] European Data Protection Supervisor, *Synthetic data*, [Online]. Available: https://www.edps.europa.eu/press-publications/publications/techsonar/synthetic-data_en, Accessed: Aug. 14, 2025, 2021.

[10] A. Alshantti, A. Rasheed, and F. Westad, "Privacy re-identification attacks on tabular gans", *Security and Privacy*, vol. 8, no. 1, e469, 2024. DOI: 10.1002/spy2.469.

[11] Synthetic Data Vault, *Ctgansynthesizer documentation*, [Online]. Available: https://docs.sdv.dev/sdv/single-table-data/modeling/synthesizers/ctgansynthesizer, Accessed: Sep. 14, 2025, 2025.

[12] OpenDP SmartNoise Team, *Differentially private conditional tabular gan*, [Online]. Available: https://docs.smartnoise.org/synth/synthesizers/dpctgan.html, Accessed: Sep. 14, 2025, 2022.

[13] M. Giomi, F. Boenisch, C. Wehmeyer, and B. Tasnadi, "A unified framework for quantifying privacy risk in synthetic data", in *Proceedings of Privacy Enhancing Technologies Symposium*, vol. 2023, 2023, pp. 312–328. DOI: 10.56553/popets-2023-0055.

[14] K. E. Emam, S. Rodgers, and B. Malin, "Anonymising and sharing individual patient data", *British Medical Journal*, vol. 350, h1139, 2015. DOI: 10.1136/bmj.h1139.

[15] L. Rodriguez and B. Howe, *Privacy-preserving synthetic datasets over weakly constrained domains*, arXiv:1808.07603, 2018.

[16] P. Nong, A. Williamson, D. Anthony, J. Platt, and S. Kardia, "Discrimination, trust, and withholding information from providers: Implications for missing

data and inequity", *SSM - Population Health*, vol. 18, p. 101 092, 2022. DOI: 10.1016/j.ssmph.2022.101092.

[17] U.S. Department of Health and Human Services, *Hipaa privacy rule and disclosures of information relating to reproductive health care*, [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/phi-reproductive-health/index.html, Accessed: May 1, 2025, 2024.

[18] European Parliament and Council of the European Union, *Regulation (eu) 2016/679 on the protection of natural persons with regard to the processing of personal data*, [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj, Accessed: Dec. 9, 2025, 2016.

[19] V. C. Pezoulas et al., "Synthetic data generation methods in healthcare: A review on open-source tools and methods", *Computational and Structural Biotechnology Journal*, vol. 23, pp. 2892–2910, 2024. DOI: 10.1016/j.csbj.2024.07.005.

[20] K. O'Shea and R. Nash, *An introduction to convolutional neural networks*, arXiv:1511.08458, 2015.

[21] R. M. Schmidt, *Recurrent neural networks (rnns): A gentle introduction and overview*, arXiv:1912.05911, 2019.

[22] P. Salehi, A. Chalechale, and M. Taghizadeh, *Generative adversarial networks (gans): An overview of theoretical model, evaluation metrics, and recent developments*, arXiv:2005.13178, 2020.

[23] M. Mirza and S. Osindero, *Conditional generative adversarial nets*, arXiv:1411.1784, 2014.

[24] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: K-anonymity and its enforcement through generalization and suppression", SRI International, Tech. Rep., 1998.

[25] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy", *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014. DOI: 10.1561/0400000042.

[26] M. I. Keskes, "Generative adversarial networks for synthetic data generation in deep learning applications", *Journal of Artificial Intelligence Research and Innovation*, 2025.

[27] J. M. Mendes, A. Barbar, and M. Refaie, "Synthetic data generation: A privacy-preserving approach to accelerate rare disease research", *Frontiers in Digital Health*, vol. 7, p. 1 563 991, 2025. DOI: 10.3389/fdgth.2025.1563991.

[28] I. M. Perez, P. Movahedi, V. Nieminen, A. Airola, and T. Pahikkala, "Does differentially private synthetic data lead to synthetic discoveries?", *Methods of Information in Medicine*, vol. 63, no. 1-2, pp. 35–51, 2024. DOI: 10.1055/a-2385-1355.

[29] J. Jordon et al., *Synthetic data: What, why and how?*, arXiv:2205.03257, 2022.

[30] J. Hayes, L. Melis, G. Danezis, and E. D. Cristofaro, "Logan: Membership inference attacks against generative models", *Proceedings on Privacy Enhancing Technologies*, vol. 2019, no. 1, pp. 133–152, 2019. DOI: 10.2478/popets-2019-0009.

[31] D. Susser et al., "Synthetic health data: Real ethical promise and peril", *Hastings Center Report*, vol. 54, no. 5, pp. 8–13, 2024. DOI: 10.1002/hast.4911.

[32] T. Stadler, B. Oprisanu, and C. Troncoso, "Synthetic data: Anonymisation groundhog day", in *Proceedings of the 31st USENIX Security Symposium*, Boston, MA, USA: USENIX Association, 2022, pp. 1451–1468.

[33] G. Astolfi and D. Köpfer, "Generating tabular data using generative adversarial networks with differential privacy", in *Conference of European Statisti-*

cians: Expert Meeting on Statistical Data Confidentiality (SDC 2021), Poznań, Poland: UNECE, 2021.

[34] C. Yan, Z. Zhang, S. Nyemba, and Z. Li, "Generating synthetic electronic health record data using generative adversarial networks: Tutorial", *JMIR AI*, vol. 3, no. 1, e52615, 2024. DOI: 10.2196/52615.

[35] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation", in *Advances in Cryptology – EUROCRYPT 2006*, ser. Lecture Notes in Computer Science, vol. 4004, Berlin, Heidelberg: Springer, 2006, pp. 486–503.

[36] Joseph Near and David Darais, *Differential privacy: Future work and open challenges*, [Online]. Available: https://www.nist.gov/blogs/cybersecurity-insights/differential-privacy-future-work-open-challenges, Accessed: Sep. 6, 2025, 2022.

[37] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax", in *International Conference on Learning Representations*, 2017.

[38] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", in *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37, PMLR, 2015, pp. 448–456.

[39] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting", *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[40] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY: Springer, 2006.

[41] A. Yousefpour et al., *Opacus: User-friendly differential privacy library in pytorch*, arXiv:2109.12298, 2021.

[42] SDMetrics, *Tvcomplement— quality metric in sdvmetrics*, [Online]. Available: https://docs.sdv.dev/sdmetrics/data-metrics/quality/tvcomplement, Accessed: Sep. 14, 2025, 2023.

[43] SDMetrics— Quality metric in SDVMetrics, *Kscomplement*, [Online]. Available: https://docs.sdv.dev/sdmetrics/metrics/quality-metrics/kscomplement, Accessed: Sep. 14, 2025, 2023.

[44] Kaggle Community, *Cardiovascular disease dataset*, [Online]. Available: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset, Accessed: Aug. 14, 2025, 2019.

[45] E. Poslavskaya and A. Korolev, *Encoding categorical data: Is there yet anything 'hotter' than one-hot encoding?*, arXiv:2312.16930, 2023.

[46] W. G. Cochran, *Sampling Techniques*, 3rd. New York: John Wiley Sons, 1977.

[47] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd. Waltham, MA: Morgan Kaufmann, 2012.

[48] H. 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*, vol. 7, p. 1 576 290, 2025. DOI: 10.3389/fdgth.2025.1576290.

[49] Wikipedia contributors, *Total variation distance of probability measures*, [Online]. Available: https://en.wikipedia.org/wiki/Total_variation_distance_of_probability_measures, Accessed: Dec. 9, 2025, 2025.

[50] C. Spearman, "The proof and measurement of association between two things", in *Studies in Individual Differences: The Search for Intelligence*, J. J. Jenkins and D. G. Paterson, Eds., Appleton-Century-Crofts, 1961, pp. 45–58. DOI: 10.1037/11491-005.

[51] H. Cramér, *Mathematical Methods of Statistics* (Princeton Mathematical Series). Princeton, NJ: Princeton University Press, 1999, vol. 9.

[52] Ruben Geert van den Berg, *Cramér's v - what and why?*, [Online]. Available: https://www.spss-tutorials.com/cramers-v-what-and-why/, Accessed: Dec. 1, 2025, 2019.

[53] Wikipedia contributors, *Cramér's v*, [Online]. Available: https://en.wikipedia.org/wiki/CramÃľr's_V, Accessed: Dec. 1, 2025, 2025.

[54] M. Giomi, F. Boenisch, C. Wehmeyer, and B. Tasnádi, *Anonymeter: A unified framework for quantifying privacy risk in synthetic data*, [Online]. Available: https://github.com/statice/anonymeter, Accessed: Dec. 9, 2025, 2023.