
Degree Programme in Computer Engineering

First Cycle 15 Credits

# Privacy preserving data access mechanism for health data

**NAJIIB ABDI DAHIR AND IKRAN DAHIR ALI**

STOCKHOLM, SVERIGE 2023

# Privacy preserving data access mechanism for health data

# Sekretessbevarande dataåtkomstmekanism för hälsodata

Najiib Abdi Dahir and Ikran Dahir Ali

Examensarbete inom datateknik
Grundnivå, 15 hp
Handledare på KTH: Reine Bergström
Examinator: Ibrahim Orhan
TRITA-CBH-GRU-2023:089

KTH
Skolan för kemi, bioteknologi och hälsa
141 52 Huddinge, Sverige

## Sammanfattning

Hälso- och sjukvårdsbranschen har länge varit en sektor som hanterar stora mängder känsliga patientdata och personuppgifter. Integriteten och säkerheten hos patientdata har blivit allt viktigare som en följd av ökad datavolym och digitalisering. Detta examensarbete fokuserade på att utforma och implementera en säker datadelning infrastruktur för att skydda integritet och sekretess för patientdata. Syntetisk data användes för att möjliggöra tillgång för forskare och studenter i reglerade miljöer utan att riskera patienters privatliv. Projektet lyckades genom att utvärdera olika integritetsbevarande mekanismer och skapa en maskininlärningsbaserad applikation för att visa den säkra datadelningsinfrastrukturens funktionalitet. Trots vissa utmaningar visade de valda algoritmerna lovande resultat i fråga om integritetsbevarande och statistisk likhet. Slutligen kan användningen av syntetiska data främja rättvisa beslutsprocesser och bidra till säkra datadelningspraxis inom hälso- och sjukvårdsbranschen.



### Nyckelord

Säker datadelning, syntetiska data, integritetsbevarande, hälso- och sjukvård, maskininlärning.

## Abstract

Due to the rise of digitalization and the growing amount of data, ensuring the integrity and security of patient data has become increasingly vital within the healthcare industry, which has traditionally managed substantial quantities of sensitive patient and personal information. This bachelor's thesis focused on designing and implementing a secure data sharing infrastructure to protect the integrity and confidentiality of patient data. Synthetic data was used to enable access for researchers and students in regulated environments without compromising patient privacy. The project successfully achieved its goals by evaluating different privacy-preserving mechanisms and developing a machine learning-based application to demonstrate the functionality of the secure data sharing infrastructure. Despite some challenges, the chosen algorithms showed promising results in terms of privacy preservation and statistical similarity. Ultimately, the use of synthetic data can promote fair decision-making processes and contribute to secure data sharing practices in the healthcare industry.

### Keywords

Secure data sharing, synthetic data, privacy preservation, healthcare, machine learning.

## Acknowledgements

We would like to express our deepest gratitude to Reine Bergström and Harsha Krishna for their invaluable assistance and support throughout this thesis. Reine's expertise and guidance, along with Harsha's technical contributions, have been instrumental in the success of our research work. Without the help and support of Reine and Harsha, this thesis would not have been possible.

# Innehållsförteckning

# 1 Introduction

## 1.1 Problem description

The healthcare industry has long been a sector that handles large amounts of sensitive patient data and personal data. Integrity and security of patient data have become more crucial as a result of increased data volumes and digitalization. Synthetic data have emerged as possible solutions to prevent data breaches and safeguard patient information. However, direct access to individual health data is limited due to privacy concerns and national regulations.

Furthermore, it is crucial to maintain the confidentiality and integrity of patient data. In order to solve this problem, synthetic data has emerged as a possible solution. With synthetic data, organizations can create realistic, but completely artificial, datasets that resemble real patient data, without sharing any sensitive information. This gives organizations the opportunity to develop software and algorithms without the risk of revealing patient data, thus protecting patient privacy and confidentiality.

Individual health data are regarded as a crucial source of information that can support precise and individualized care in the modern world. Due to privacy concerns, direct access to health data is not possible, and Sweden implements strict national regulation and the GDPR to protect the privacy of individuals. In order to give access to more students and researchers, even in regulated environments, new strategies are needed.

In this bachelor's thesis, the issue is to design and implement a secure data sharing infrastructure based on a known dataset and to show that it works with a sample machine learning application.

## 1.2 Goals

The project aim can be divided into three objectives:

- Identify and evaluate different solutions for meeting privacy requirements for data sharing.
- Develop a sample machine learning-based application that can use the secure data sharing infrastructure to demonstrate its functionality.
- Test and validate the secure data sharing infrastructure by proving that it meets integrity requirements.

## 1.3 Delimitations

In order to achieve the goals of this thesis, the study implements one option, therefore it doesn't aim to compare several implementations.

- This study implements one privacy preserving mechanism. Therefore, it doesn't aim to compare several implementations.
- The dataset used in this study was relatively small, which may limit its ability to fully represent the entire population or provide a comprehensive understanding.

- Only two algorithms were chosen for implementation and evaluation, limiting the exploration of alternative algorithms.

## 1.4 Methods

Our research will begin with a study that gives basic understanding on how synthetic data works and how to implement it so it can preserve privacy and integrity.

# 2 Theory and background

A brief introduction will be given to the definition of personal and sensitive data, followed by a detailed discussion of the topic of synthetic data.

## 2.1 Personal data and sensitive data

The paper [1] examines the historical progression of safeguarding personal data and privacy in healthcare, dating back to the 1970s when technological advancements and the increased utilization of personal information began. It delves into John Stuart Mill's notion of privacy as a means for individuals to realize themselves and shield against government authority. Additionally, the paper acknowledges the UN's International Covenant on Civil and Political Rights, which upholds privacy and acknowledges the existence and rights of individuals. However, it also recognizes that individual actions can encroach upon the privacy and freedom of others, implying that complete autonomy has limitations. Overall, the paper underscores the significance of privacy in healthcare, emphasizing its role in preserving individual autonomy while balancing the rights of individuals and others.

Personal data is defined by the European Commission [2] as any information that relates to an identified or identifiable living individual. This may consist of different bits of information that, when put together, may enable the identification of a certain individual. If personal data may still be used to identify a specific person after being transformed by techniques like de-identification, encryption, or pseudonymization, it is still covered by the GDPR. Nevertheless, if personal information has been anonymized so that it cannot be used to identify a specific person, it is no longer regarded as personal information. True anonymization needs to be irreversible to take place. Examples of personal data include surname and name, email address, home address, location data and identification card number.

As defined by the Swedish Authority for Privacy Protection [3], sensitive data involves details regarding an individual’s health and sex life, race, political view, genetic data, philosophical or religious beliefs and biometric data uniquely for a person.

Based on Pipeda's [4] definition, personal information is any accurate or personal information about an individual, whether it is documented. Personal data includes name, ID numbers, ethnic origin, blood type or age.

## 2.2 Synthetic data

This subsection will give basic understanding of what synthetic data are and investigate the various ways of generating synthetic data.

### 2.2.1 What is Synthetic data?

Data that is generated using a specific mathematical model or algorithm and intended to meet one or more data science goals is referred to as synthetic data [5]. Real data, which is obtained from authentic real-world systems like financial transactions, satellite photos, medical testing, and other such examples rather than a model, is separated from synthetic data. The synthetic data generator, or model, can take many different forms, including agent-based and econometric models, a set of (stochastic) differential equations that simulate a physical or economic system, and deep learning structures like the frequently used Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs).

The use of computer-generated synthetic data for addressing certain problems has a long history, dating to the 1940s when Stanislaw Ulam and John von Neumann pioneered the Monte Carlo simulation techniques. Since it provides a *"ground truth*" that is helpful in developing and evaluating machine learning pipelines, this method has been widely used in research.
Synthetic data use has gained popularity as a method for tackling various problems in a variety of sectors. The release of sensitive information, eliminating bias and assuring fairness, and boosting data robustness through strengthening are three essential areas in the field of machine learning that have attracted a lot of attention [5].

### 2.2.2 What are synthetic data sets?

According to the paper "Privacy Enhancing Technologies and Synthetic Data" by Paul Wagner [6], synthetic data sets refer to the process of creating data with accurate statistical properties in order to enable businesses to function without disclosing actual user data. The paper explains that a synthetic data set consists of data that is specially generated to imitate the patterns and analytical capabilities of real data about actual people or events by mimicking their significant statistical aspects. A synthetic dataset can be classified into two types: fully synthetic and partially synthetic. Although there is a data set known as hybrid synthetic, due to its computational complexity, hybrid synthetic data does not exist in practice since it consists of both original and synthetic data. In the case of a fully synthetic data set, there is no original data contained in the data object. Therefore, there is little chance that any individual unit can be re-identified by reclassifying the synthetic data. Despite, all variables are still available. On the other hand, partially synthetic data sets contain some original data, but usually sensitive information is replaced.

### 2.2.3 Types of synthetic data

Synthetic data generation is a technique for safeguarding privacy in published data, offering an alternative to data masking. It involves creating randomized data with specific constraints to conceal sensitive private information, while maintaining key statistical information and relationships found in the original dataset. The resulting synthetic data can be classified into three main categories [7].

#### *2.2.3.1 Hybrid synthetic data*

Hybrid synthetic data refers to the process of merging synthetic data with original data. In this approach, each record from the original dataset is paired with the most similar record from the synthetic dataset, and the two records are combined. By incorporating both full and partial synthetic data, hybrid synthetic data offers several advantages, including enhanced privacy preservation and increased utility compared to fully synthetic and partially synthetic data. However, it should be noted that this method requires additional memory and processing time due to the combination of datasets.

#### *2.2.3.2 Partial synthetic data*

The method of generating partially synthetic data involves replacing only the values of a selected attribute with synthetic values, as opposed to fully synthetic data where all values are replaced [7]. This approach aims to protect privacy by substituting original values with synthetic ones to prevent re-identification. Multiple imputed values and model-based techniques are employed to avoid disclosure, and these techniques can also be used to substitute missing values in the original data. However, it should be noted that partially synthetic data carries a higher risk of disclosure compared to fully synthetic data, as it contains both original and imputed data.

#### *2.2.3.3 Fully synthetic data*

The methodology involves the use of fully synthetic data generators to create artificial data that does not include any original information [7]. The process begins by determining the density function of

attributes in the original data and estimating the associated parameters. Synthetic data is then generated by randomly selecting values from these estimated density functions, while preserving privacy. Only a select few attributes from the original data are replaced with synthetic data, and the protected series are aligned with the remaining attributes from the original data to maintain their relative ranking. Traditional techniques such as multiple imputation and bootstrapping are employed to generate the fully synthetic data. It is important to note that although this technique provides robust privacy protection by producing completely artificial data, the accuracy of the generated data cannot be guaranteed.

### 2.2.4 Challenges with synthetic data

Artificial data is being used as a more and more practical substitute for real data when training machine learning algorithms [8]. This is due to recent developments in algorithms and generative models. The use of synthetic data still presents several challenges that need to be addressed in order to achieve high performance. There are several limitations to machine learning, such as the absence of standardized tools for creating synthetic data, the distinction between artificial and actual data, and the extent to which imperfect synthetic data can be utilized by machine learning algorithms.

### 2.2.5 Synthetic data requirements

Four key conditions must be met by a high-quality synthetic data generator (SDG) [5], as mentioned below.

1. Syntactical accuracy: It must assure syntactic correctness by creating realistic data that retains the original data's structural qualities, such as avoiding the usage of future information in time-series data or retaining graph structure in financial transaction networks.
2. Privacy: It should emphasize privacy by calculating the amount of information revealed in synthetic data using approaches such as differential privacy. Yet, privacy assessment may differ based on the activity and data characteristics.
3. Statistical accuracy: It should offer statistical correctness by evaluating the similarity between synthetic and original data in terms of marginal distributions and variable connections, while providing control over these characteristics.
4. Efficiency: It should be efficient and capable of scaling well with the complexity of the data space, as the curse of dimensionality might impact distribution approximation.

### 2.2.6 Evaluating synthetic data

When evaluating synthetic data, it is usual practice to use generic metrics such as comparing variable distributions or correlations rather than accounting for the specific studies that will be done [9]. Another approach is to use a classification model to assess the distinguishability of the synthetic data. The usage of generic metrics is generally automatic and gives a decent indicator of the value of the data. If a dataset fails on broad metrics, it is unlikely to succeed on additional tests. The following metrics are common when evaluating synthetic data.

#### Kullback-Leibler divergence

The Kullback-Leibler (KL) divergence, also referred to as relative entropy, is a measurement used to quantify the distinction between two probability distributions [10]. It serves as an effective tool to assess the similarity or dissimilarity of different distributions. When applied to probability distribu-

tion P and Q, where P represents the true distribution of random variables and Q represents a theoretical or fitting distribution, KL(P||Q) is known as the forward KL divergence while KL(Q||P) is the backward KL divergence as shown in 2.1.

$$KL(P_1||P_2) = \int_{x \in X} P_1(x) log \frac{P_1(x)}{P_2(x)} \quad (2.1)$$

Kolmogorov-Smirnov test

In the paper [11], they consider two multi-sets of real numbers, denoted as $R$ and $T$. $R$ is a reference set with $n$ elements sampled from an unknown univariate probability distribution, and $T$ is a test set with $m$ elements that may or may not be sampled from the same distribution as $R$. The terms "set" and "multi-set" are used interchangeably, with multi-set being the default. The Kolmogorov-Smirnov (KS) test is used to determine if $T$ is sampled from the same distribution as $R$ by comparing their empirical cumulative functions as shown in 2.2. The null hypothesis is that $T$ and $R$ are sampled from the same distribution. The KS test involves three steps: computing the KS statistic, determining the target $p$-value for a user-specified significance level $\alpha$, and comparing $p$ and $D(R,T)$ to determine if the null hypothesis is rejected. If the null hypothesis is rejected, it means that the empirical cumulative functions of $R$ and $T$ are significantly different, and thus $T$ is unlikely to be sampled from the same distribution as $R$. Otherwise, if the null hypothesis is not rejected, we say that $T$ and $R$ pass the KS test. Sorting the elements in $R \cup T$ in ascending order is necessary to compute the KS statistic, and the time complexity of conducting the KS test is $O((n+m)\log(n+m))$.

$$D(R,T) = max_{x \in R \cup T} |F_R(x) - F_T(x)| \quad (2.2)$$

K-Anonymity

The k-anonymity definition for a table is as follows [12]: if T (AI,..., An) is a table with a quasi-identifier (QI) associated with it, T satisfies k-anonymity with respect to QI if every sequence of values in T[QI] appears at least k times in T[QI]. The k-anonymity model involves modifying the QID attributes using suppression and generalization operations to form groups of records that share the same QID values, called equivalence classes (EQ). This makes each record indistinguishable from a group of at least k-1 other records, thereby making the QIDs imprecise and less informative to prevent linking an individual to a record.

L-diversity

The concept of l-diversity was introduced as an enhancement to k-anonymity in order to overcome its limitations [13]. It represents a novel approach that addresses the challenge of preserving data privacy without requiring knowledge of an adversary's background information to prevent attribute disclosure. This technique revolves around the idea of ensuring that sensitive attributes within each group are adequately represented. Essentially, l-diversity modifies the principle of k-anonymity by incorporating the concept of k-anonymity.

## 2.3 Synthetic data generation

Within this subsection, different approaches to generating synthetic data are explored, including the Variational Autoencoder (VAE), Generative Adversarial Network (GAN), and Bayesian Network.

### 2.3.1 Variational autoencoder

VAE is a type of neural network that utilizes a continuous latent variable to represent the data distribution [8]. The encoder and the decoder are the two neural network elements that make up the VAE architecture. The encoder maps the input data into a continuous latent space, and the decoder maps the latent variable back to the input space to recreate the original data.

A distinguishing feature of autoencoding is that it is a data-specific process, which implies that it can only compress data that is similar to the data on which it has been trained [14]. This contrasts with previous compression techniques that relied on data generalizations. Comparatively, earlier compression techniques based on generalizations of the input data could not achieve this result. Furthermore, autoencoders are "lossy," which implies that the decompressed outputs are inferior to their original form.

VAEs learn a probability distribution that models the input data, which enables them to capture a wider range of features compared to traditional autoencoders [14]. To train the VAE model, two loss functions are used: the reconstruction loss and the KL divergence between the learned latent distribution and the prior distribution. The three essential components required to build an autoencoder are an encoding function, a decoding function, and a loss function that measures the information loss between the compressed and decompressed representations.

### 2.3.2 Generative adversarial network

GAN is a type of neural network that solves the generation problem by transforming it into a learning problem [8]. There are two neural network components that make up this system: the discriminator and the generator. Utilizing random noise as input, the generator creates synthetic data that is similar to the actual data. In contrast, the discriminator makes use of a combination of real and synthetic data to categorize the data as genuine or fraudulent.

A GAN does not require the data to be remodeled since it uses a generator that can extract direct samples from distributions [15]. The greatest benefit of this approach is that it allows GAN to adapt completely to the distribution of the actual data. GANs play an essential role in generative models because of their ability to produce data that can be interpreted naturally. Due to the absence of a complex variational lower limit in GAN, training becomes simpler and more effective. Using GANs, points can be generated only on thin surfaces that are close to the data without having to resort to inefficient Markov chain techniques or approximate inference.

Although GANs have been shown to be successful at modeling continuous distributions, dealing with discrete data, such as text, presents difficulties [16]. Back-propagation during the learning phase is difficult because the loss function for discrete random variables is not differentiable, which is the problem. Adjusting the loss function and approximating the discrete distribution with a continuous distribution are two potential fixes. The generator may also be modeled as a stochastic policy in reinforcement learning, which solves the differentiability problem by changing the optimization strategy.

### 2.3.3 Bayesian Network

A graphical model called a Bayesian network is used to capture the joint probability distribution of a group of variables [17]. It consists of two essential parts: a graphical structure that illustrates the connections between variables, and a set of conditional probability distributions that explain the probabilistic relationships among the variables. This modeling approach, also referred to as a belief network, is commonly utilized for performing probabilistic inference, enabling inference about one variable given the values of other variables in the network. To generate data based on a Bayesian network, multiple synthetic datasets need to be created. These synthetic datasets can then be shared for broader usage while ensuring the confidentiality of the original data.

## 2.4 Algorithms

This subchapter will provide basic understanding regarding privacy focused algorithms and statistical focused algorithms.

### 2.4.1 Privacy focused algorithms

Differentially Private Generative Adversarial Network (DPGAN) is an algorithm used for privacy guarantees [18]. DPGAN ensures differential privacy by safeguarding the generator and discriminator parameters from revealing any information about the training data's privacy. Differential privacy is a privacy concept that assures that an algorithm's output remains unaffected even when an individual's data is added or removed from the dataset. The generator's parameters in DPGAN provide differential privacy guarantees concerning the training data.

The article [19] discusses GANs and their potential for generating high-quality data but highlights the obstacle of using them with sensitive data as it can lead to the disclosure of private information. In response, the authors suggest a solution in the form of Privacy-preserving Generative Adversarial Network (PPGAN) model that employs differential privacy and a Moments Accountant approach to manage privacy loss. The authors back their claim with a mathematical verification of the differential privacy discriminator and illustrate the effectiveness of PPGAN in creating high-quality synthetic data while keeping privacy risks to an acceptable level using case studies.

### 2.4.2 Statistical focused algorithms

The article [20] suggests that utilizing Least Squares Generative Adversarial Network (LSGAN) as a substitute for conventional GANs that rely on the sigmoid cross entropy loss function. It have been observed to generate higher quality images compared to regular GANs. LSGANs punish instances that are accurately classified but exhibit a considerable difference from the actual data, motivating the generator to produce samples that are closer to the decision boundary *(the boundary that separates the real and fake data)* and the genuine data manifold *(set of all possible variations of the real data)*. Furthermore, this approach results in more gradients, which results in a more consistent learning process and addresses the issue of vanishing gradients.

A technique called Conditional Generative Adversarial Network (CGAN) is employed for producing artificial data that corresponds to the class of the instances [21]. To assess the usefulness of the generated samples, two approaches are used: examining the correlation between the genuine and the artificial data and contrasting the classification accuracy of an algorithm when applied to each dataset. A GAN doesn't consider any conditions regarding the data. Typically, the generated synthetic data possesses a unique characteristic that needs to be utilized to achieve synthetic data that closely resembles real data. CGANs are an enhanced version of GANs that take a certain condition into account. This condition requires both the generator and discriminator to consider supplementary information,

referred to as "c". This additional information could be anything, including data from a different source or a class label.

Creating a realistic synthetic dataset and accurately modeling the probability distribution of rows in tabular data can be a challenging task [22], especially when dealing with data that includes both continuous and discrete columns. Continuous columns may have multiple modes, while discrete columns can be imbalances, posting difficulties for modeling. Standard statistical and deep neural network models have proven inadequate in handling this type of data. To address these challenges, CTGAN was designed, which uses a conditional generative adversarial network.

## 2.5 Tools

This subchapter describes the different tools used for creating synthetic data and how to use them.

### 2.5.1 Synthpop

In place of actual data, synthetic data is used to create publicly accessible datasets that may be utilized for inference [23]. This is true, but only if the model that was used to create the synthetic data properly reflects the real mechanism that produced the observed data. Users of confidential datasets can access test data that closely mimics the real data using the Synthpop tool for R. Using the artificial data, users can perform exploratory analyses and test models, but the code created on the artificial data is utilized to run the final analysis on the real data. This strategy acknowledges the limits of the synthetic data generated by these techniques.

The article examines the creation and functionality of the R package Synthpop, which was developed as part of the SYLLS project [23]. The goal of the package is to produce synthetic data that imitates the observed data of the England and Wales Longitudinal Study, Scottish Longitudinal Study, and Northern Ireland Longitudinal Study, while preserving confidentiality. LS support staff can use the package to generate customized synthetic data for LS users, and the package includes routines to generate, summarize and compare the synthetic data and models to gold standard analyses. The article also notes that the synthpop package can be utilized for other confidential data where synthetic data would be beneficial. Additionally, the article provides a comparison of the synthpop package to other similar software such as simPop and IVEware [23].

### 2.5.2 DataSynthesizer

The DataSythesizer system is an end-to-end solution built on Python 3, which creates synthetic datasets from private datasets in CSV format [24]. The DataDescriber module analyzes the input dataset and deduces the domains and distribution estimates of its attributes, which are then saved in a dataset description file. The DataGenerator module uses this information to generate the synthetic dataset by drawing samples from the frequency distribution for categorical attributes, the equi-width histogram for non-categorical numerical and datetime attributes, and the length range for non-categorical string attributes. This methodology ensures that the statistical properties of the original dataset are preserved while maintaining privacy protection for sensitive information. Users can customize the data type and categorical status settings on an attribute level. The DataGenerator module has several modes to produce synthetic data and supports a unique random seed per user to prevent disclosure of private information through repeated data generation requests [24].

### 2.5.3 SynthCity

The SynthCity project is a collaborative effort aimed at developing an open-source software platform that can facilitate the innovative use of synthetic data in areas such as fairness, privacy, and augmentation across different data types [25]. The project has recently introduced the beta version of the synthcity library, which encompasses all the key applications of synthetic data generation, including fairness, privacy, and augmentation, while also providing evaluation metrics to measure the quality of the generated synthetic data. The library is equipped with several utility functions to automate and simplify workflows, such as conducting comparative evaluations of multiple data generators. The current iteration of the synthcity library concentrates on generating tabular data, which is widely used in a variety of industries, including regulated ones that have limited access to data [25]. The library can process different types of tabular data, such as static tabular data, time series data, and censored survival data. Future versions of the library will encompass additional data modalities and incorporate new generators. The project invites the community to participate in the development process by sharing feedback, reporting issues, and submitting pull requests on GitHub.

The Synthcity library presents a complete solution for producing and assessing synthetic data [25]. Its workflow encompasses loading data, training a data generator, generating synthetic data, and evaluating it using several metrics. Users can make use of the Plugin class to train and apply different data generators, as the library offers a uniform interface for loading diverse input data. Additionally, the Metrics class facilitates the assessment of synthetic data's fidelity, utility, and privacy. The Benchmark class is also at hand to compare and evaluate various data generators. In this project, SynthCity is the tool chosen to be used (see further section 3.4)

## 2.6 Related work

In [26], the authors investigate the current status of synthetic medical data generation and its increasing popularity as a means of protecting individual privacy while enabling medical research and innovation. The study discusses three techniques for creating synthetic data: knowledge-driven, data-driven, and hybrid approaches. Knowledge-driven approaches offer excellent privacy protection, but they require manually specifying the generative model. Developing a suitable generative model poses difficulties for data-driven approaches when dealing with complex and interrelated medical data. Hybrid approaches can overcome these difficulties by integrating domain expertise with data-driven methodologies.

Additionally, the study emphasizes the importance of accurate metrics to measure the realism of synthetic data. Evaluating the realism of data-driven generative processes throughout development could improve accuracy.

In their article [27], the authors discuss how machine learning-based services can assist the healthcare industry in shifting its focus towards prevention. However, safeguarding individuals' privacy when sharing sensitive personal information is crucial. To address this challenge, synthetic datasets generated with generative models like GANs offer a promising solution for privacy-preserving data sharing. Generating realistic synthetic data that maintains the statistical properties of the original dataset is particularly challenging for smart healthcare data, which involves various data types and distributions. To address this, the authors propose a GAN that utilizes differential privacy mechanisms to generate a synthetic and differentially private smart healthcare dataset. The evaluation of the proposed approach using a real-world Fitbit dataset demonstrates its ability to generate a realistic and differentially private dataset while preserving the original dataset's statistical properties.

# 3 Methods

In this chapter, the methodology used to achieve the results of the thesis will be presented. The chapter will detail the steps taken to create an implementation, including decisions made regarding the selection of techniques used to generate synthetic data. Additionally, a description of how knowledge and resources were gathered will be discussed in this section.
In the initial stage of the implementation, a tool mentioned in chapter 3.4 will be used to facilitate the creation of synthetic data. The implementation will consist of a given dataset which will be the “input real data”, as shown as step one in figure 3.1. Step two will be to training the generator using a plugin, for this implementation the generator used will be GAN, and two algorithms originating from GAN will be used, see chapter 3.3. Step three will be to generate the synthetic data and step four will be to evaluate the synthetic data using the metrics chosen in 3.5. The generated synthetic data will be analysed using the metrics mentioned in 3.5. Synthetic data serves as a secure data sharing infrastructure, safeguarding privacy, and integrity during information exchange. As mentioned in chapter 2.2.3, by generating randomized data that preserves statistical properties and relationships while concealing sensitive information, it allows for sharing without compromising individual privacy or disclosing personally identifiable data.

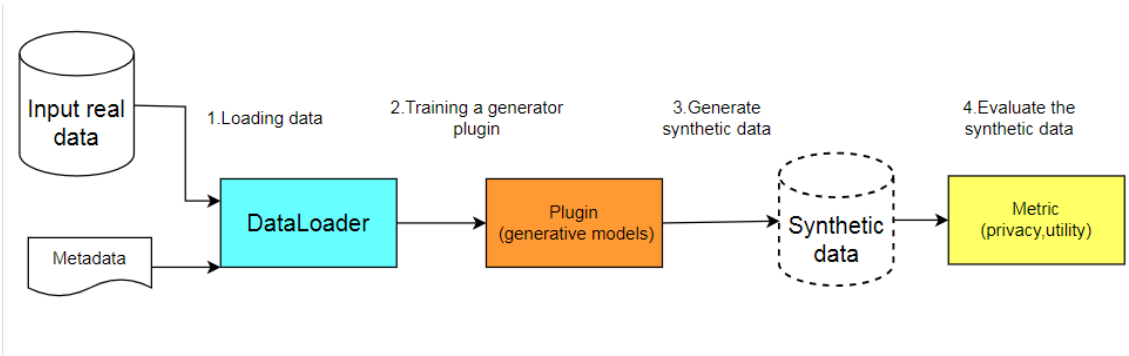


FIGURE 3.1: OVERVIEW ON HOW THE IMPLEMENTATION FUNCTIONS

## 3.1 Pre-study

A preliminary study was conducted over three weeks to examine the functions and techniques used to generate synthetic data. During the study, relevant information was acquired, and a general understanding of the topic was developed. Resources used to access and find information were obtained through scientific databases such as IEEE Xplore, Google Scholar and ACM Digital Library. Based on the knowledge gathered, a potential approach for creating synthetic data was presented.

## 3.2 Choice of synthetic data generator

Despite having the same objective, GANs, VAEs and Bayesian Networks discussed in chapter 2.3 differ in several ways. While VAE and Bayesian Network have their own merits, they may not be as well-suited for the task of generating synthetic data given this thesis specific dataset and objectives. VAEs are data-specific and "lossy," meaning the decompressed outputs are inferior to the original data. Bayesian Networks, on the other hand, require the creation of multiple synthetic datasets, which is not efficient for this thesis purposes.

GAN is a powerful choice for generating synthetic data from the given dataset, as it offers adaptability to the data distribution, natural interpretation and quality of generated data, simplicity and effectiveness of training, and the ability to generate data on thin surfaces close to the original data distribution. Considering these factors, GAN is the most suitable choice for this thesis topic.

## 3.3 Choice of algorithm

Both DPGAN and PPGAN, as mentioned in 2.4.1, can preserve privacy in the context of generating synthetic data. However, DPGAN is the better choice for this thesis for several reasons.

Firstly, DPGAN provides differential privacy guarantees for the training data by safeguarding the generator and discriminator parameters. This means that even if an individual's data is added or removed from the dataset, the algorithm's output will remain unaffected. In contrast, while PPGAN also utilizes differential privacy, it is difficult to draw a conclusion regarding why PPGAN should be preferred over DPGAN due to the limited information available.

Secondly, DPGAN has been specifically designed for privacy preservation in GAN. On the other hand, PPGAN is a model that has been proposed as a solution to the problem of sensitive data disclosure in GAN. The available information does not allow us to conclude that PPGAN is optimized for privacy to the same extent as DPGAN, which is specifically designed for privacy preservation. Although PPGAN has shown effectiveness in generating high-quality synthetic data with acceptable privacy risks, a comprehensive and direct comparison is challenging without additional detailed knowledge. Lastly, in the context of healthcare, DPGAN is more suitable as it provides better privacy guarantees.

Two GANs that could be used to generate synthetic data for statistical similarities are LSGAN and CTGAN, mentioned in 2.4.2. While LSGAN is useful for generating images, it may not be the best choice for generating synthetic healthcare data because it does not consider the conditions of the data [20].

In contrast, CTGAN is a conditional GAN that is designed specifically for generating synthetic tabular data. CTGAN considers the conditions of the data, making it more suitable choice for generating synthetic healthcare data with different features from the given dataset, including age, sex, episode number and hospital outcome. CTGAN can model the probability distribution of rows in tabular data accurately, even when dealing with data that includes both continuous and discrete columns.

In conclusion, based on the nature of the healthcare dataset and the goal of using it as a statistically focused algorithm, CTGAN is the most appropriate algorithm for this purpose.

## 3.4 Choice of tool

Synthpop, DataSynthesizer and SynthCity, as mentioned in chapter 2.5, are three different tools for generating synthetic data. Each of these tools has its own strengths and weaknesses. In comparison to Synthpop and DataSynthesizer, SynthCity presents a complete solution for producing and assessing synthetic data. It offers a uniform interface for loading diverse input data and can process different types of tabular data. The library is equipped with several utility functions to automate and simplify workflows, such as conducting comparative evaluations of multiple data generators.

SynthCity is particularly well-suited for the purpose of this bachelor thesis because it offers privacy protection and fairness in data generation, as well as evaluation metrics to measure the quality of the

generated synthetic data. Additionally, the SynthCity library can process different types of tabular data, including static tabular data, time series data, and censored survival data.

One potential disadvantage of SynthCity is that it is still in beta version, which means that it may have bugs or incomplete futures. However, the project invites the community to participate in the development process by sharing feedback, reporting issues, and submitting pull requests on GitHub, which can help improve the tool over time.

In summary, SynthCity is a better choice for this thesis than Synthpop and DataSynthesizer because it offers a complete solution for generating and assessing synthetic data, including privacy protection, fairness, and evaluation metrics. It can also process different types of tabular data and provides a uniform interface for loading diverse input.

## 3.5 Choice of metrics

The Kolmogorov-Smirnov test and the inverse of the Kullback-Leibler Divergence, as mentioned in 2.3.7, will be used as statistical similarity metrics, and the k-anonymity and l-diversity tests, as mentioned in 2.3.7, will be used as privacy loss metrics.

The KL divergence will be used because it is a widely used measure to assess the difference between probability distributions, making it suitable for comparing the synthetic data distribution with the true distribution of patient attributes. As the generated synthetic data is based on patient attributes, understanding the dissimilarity between the synthetic and real data distribution is essential, which can be accomplished with KL divergence.

The KS test will be used as a complement for KL divergence because it provides an additional perspective on the similarity or dissimilarity of the distributions, focusing on the empirical cumulative function rather than the explicit probability distribution.

K-Anonymity will be used because it modifies quasi-identifier attributes like age, sex, and episode number, in a dataset, similar to the one used in this thesis. K-Anonymity also creates equivalence classes to safeguard privacy by reducing the risk of linking individuals to their data.

L-diversity will be used because it is an advanced method of privacy protection that builds upon the concept of k-anonymity. It overcomes the limitations of the attribute disclosure and provides enhanced safeguards for sensitive information. Unlike k-anonymity, which focuses on making records indistinguishable within a group, l-diversity ensures that each group contains a variety of sensitive attribute values.

# 4 Results

This chapter presents the results of this thesis research, focusing on the evaluation of two generative adversarial network GAN architectures: DPGAN and CTGAN. The process is outlined and followed, including the development of the DPGAN and CTGAN pipelines, and present the evaluation methodologies employed to assess their performance.

## 4.1 Process

Both DPGAN and CTGAN were tested on a dataset consisting of 15,000 rows and 4 columns. The objective was to create synthetic data that maintains privacy and integrity.

Using DPGAN, the model was trained on the original dataset, resulting in the generation of 1,500 synthetic data points. Different epsilon values were evaluated to assess the level of privacy protection, with epsilon representing the degree of privacy preservation. The epsilon values chosen was (0.1,5,10). The evaluation specifically focused on protecting sensitive attributes like age and sex, employing differential privacy constraints.

Similarly, with CTGAN, the model generated 4,000 synthetic data points after completing the training process. The evaluation also emphasized looking into the statistical similarities while safeguarding age and gender attributes, which were classified as sensitive.

## 4.2 DPGAN pipeline

This subchapter presents an overview of the DPGAN pipeline and its evaluation. The step-by-step implementation of the DPGAN pipeline is presented, followed by a presentation of the evaluation of the obtained results.

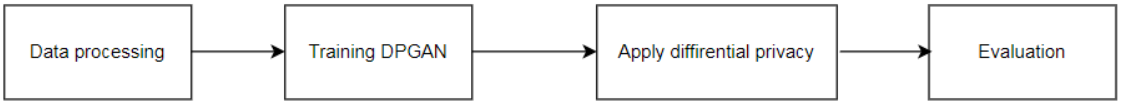


FIGURE 4.1: HOW THE PIPELINE FOR DPGAN IS OUTLINED

1. Data processing

   The first step is to preprocess the desired dataset into a csv file to be used for training the DPGAN model as shown in table 4.1.

TABLE 4.1: EXAMPLE OF ORIGINAL DATASET IN A CSV FILE

| age_years,sex_0male_1female,episode_number,hospital_outcome_1alive_0dead |
|---|
| 21,1,1,1 |
| 20,1,1,1 |
| 21,1,1,1 |
| 77,0,1,1 |
| 72,0,1,1 |
| 83,0,1,1 |
| 74,0,1,1 |
| 74,1,1,1 |

2. Training the DPGAN model

The next step is to train the DPGAN model using the preprocessed data. The generator generates synthetic data samples that are similar to the original data while enforcing differential privacy constraints. The discriminator distinguishes between real and synthetic data samples and provides feedback to the generator on how to improve its output and the realism of the generated samples. Shown in table 4.2 is the generated synthetic data after the training is complete.

TABLE 4.2: EXAMPLE OF GENERATED SYNTHETIC DATA IN A CSV FILE

| age_years,sex_0male_1female,episode_number,outcome_1alive_0dead |
|---|
| 46,1,1,0 |
| 51,0,1,0 |
| 34,0,1,1 |
| 56,0,2,1 |
| 64,0,2,0 |
| 27,1,2,0 |
| 40,0,2,0 |
| 48,0,2,1 |
| 45,0,1,0 |

3. Applying differential privacy

To ensure that the synthetic data generated by the DPGAN model is differentially private, noise is added to the gradient during the training. In this part an epsilon value needs to be chosen. Epsilon measures the privacy loss associated with algorithms or mechanisms. It serves as a quantitative indicator to gauge the maximum potential changes in the probability distribution of outcomes or outcome sets when including or excluding a single individual's data. The value of epsilon is critical in determining the strength of privacy guarantees provided by the algorithm or mechanism under consideration. A lower value of epsilon indicates a stronger level of privacy protection.

4. Evaluation of the data

The final step involves evaluating the synthetic data, which consists of several components that contribute to obtaining a score, as shown in Figure 4.1.

Firstly, in row number 2, the benchmarking process aims to assess the performance and utility of the synthetic data generated by the DPGAN model for different epsilon values.

Secondly, the loader serves as an interface between the original data and the synthetic data generation method. It facilitates access and manipulation of data, ensuring that the synthetic data generation process has sensitive attributes and the necessary input data in place. As a result, synthetic data that protects privacy can be created.

Thirdly, the synthetic size determines the quantity of synthetic data points generated for evaluation. By varying the size of the synthetic dataset, the evaluation provides valuable insights into the method's performance and effectiveness in handling different data quantities. This aids in assessing and optimizing synthetic data techniques.

Fourthly, the repeats parameter determines the number of iterations in which the evaluation process is performed for each combination of epsilon values. This repetition enables a more robust evaluation by reducing the impact of randomness and providing a more representative assessment of synthetic data utility.

Lastly, the score is presented, with the help from print() in the benchmark class, as shown in figure 4.2.

```
41
42  score = Benchmarks.evaluate(
43      [(f"test_eps_{eps}", "dpgan", {"epsilon": eps}) for eps in [0.1]],
44      loader,
45      synthetic_size=len(generated),
46      repeats=2,
47  )
48  Benchmarks.print(score)
49
```

FIGURE 4.2: CODE FOR APPLYING DIFFERENTIAL PRIVACY AND EVALUATING DPGAN.

### 4.2.1 Evaluation of results

The “+/-”-value suggests the range of uncertainty or variability associated with the measurement.

The stats.inv_kl_divergence.marginal is the mean inverse of the Kullback-Leibler Divergence, with a value of zero indicating that the datasets are from different distributions, while a value of one means that they are from the same distribution. As shown in table 4.3, the results indicate a high level of inverse KL divergence across various epsilon values.

The stats.ks_test.marginal, which is the Kolmogorov-Smirnov test, has a value of zero if the distributions are completely different, and one if the distributions are identical. The distributions of the synthetic data and the original data show a high degree of similarity, with the values shown in table 4.3.

When it comes to privacy.k-anonymization, ".gt", refers to the evaluation of the k-anonymization technique on the ground truth data and ".syn", represents the evaluation of the k-anonymization technique on the synthetic data generated.

The results suggests that, on average, each record in the ground truth dataset is indistinguishable from approximately 176 other records in different epsilon values.

The generated synthetic data does not achieve the same level of indistinguishability as the ground truth, with a minimum value of 51 when epsilon is set to 10. At epsilon 0.1, which corresponds to the highest privacy guarantee, the results exhibit a notable level of uncertainty as shown in table 4.3.

Similarly, for l_diversity, a higher value of "l" provides stronger privacy protection but may result in less utility in the data. On the other hand, a lower value of "l" may lead to more utility in the data, but weaker privacy protection. The result in table 4.3 shows that the l-diversity is zero for all the epsilon values.

TABLE 4.3: EVALUATION RESULTS USING DIFFERENTIAL PRIVACY WITH VARIOUS EPSILON VALUES

| Epsilon | 0.1 | 5 | 10 |
|---|---|---|---|
| privacy.k-anonymization.gt | 176.0 +/- 0.0 | 176.0 +/- 0.0 | 176.0 +/- 0.0 |
| privacy.k-anonymization.syn | 78.5 +/- 29.5 | 71.0 +/- 3.0 | 51.5 +/- 11.5 |
| privacy.distinct l-diversity.gt | 0.0 +/- 0.0 | 0.0 +/- 0.0 | 0.0 +/- 0.0 |
| privacy.distinct l-diversity.syn | 0.001 +/- 0.0 | 0.001 +/- 0.0 | 0.001 +/- 0.0 |
| stats.inv_kl_divergence.marginal | 0.914 +/- 0.017 | 0.959 +/- 0.01 | 0.866 +/- 0.003 |
| stats.ks_test.marginal | 0.879 +/- 0.037 | 0.873 +/- 0.034 | 0.801+/-0.001 |

.

## 4.3 CTGAN pipeline

This subchapter presents an overview of the CTGAN pipeline and its evaluation. The step-by-step implementation of the CTGAN pipeline is presented, as shown in figure 4.3, followed by a presentation of the evaluation of the obtained results.

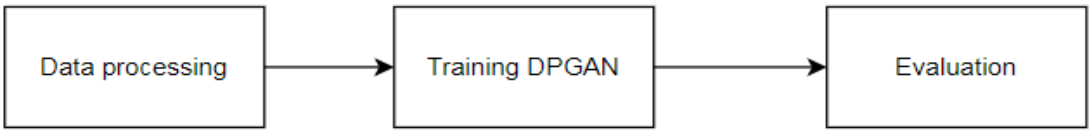


FIGURE 4.3: HOW THE PIPELINE FOR CTGAN IS OUTLINED

1. Data processing

CTGAN requires a tabular dataset as input as shown in table 4.1.

2. Training the CTGAN

   CTGAN's generator network takes random noise as input and generates synthetic samples. The generator network is trained to minimize the difference between the real and synthetic data distribution. CTGAN's discriminator network takes as input a batch of the real or synthetic data samples and predicts whether each sample is real or synthetic. The discriminator network is trained to maximize the difference between the real and synthetic data distributions. After completing the training, the synthetic data is generated as shown in table 4.2.

3. Evaluation of the data

   With the help of synthcity, the evaluation can be done with plugins that have several built-in metrics. That takes the original data and the synthetic data into consideration and performs various metrics that the plugin contains.

### 4.3.1 Evaluation of results

The stats.inv_kl_divergence.marginal is the mean inverse of the Kullback-Leibler Divergence, higher values indicate a closer match and stronger similarity between the datasets. As shown in table 4.4, the results indicate a high level of inverse KL divergence.

The stats.ks_test.marginal, which is the Kolmogorov-Smirnov test, higher values indicate a closer match and stronger alignment. The result shows a relatively high min and max values, as well as a high mean value, as shown in table 4.4. The standard deviation suggests that there is some variability in the alignment.

When it comes to privacy.k-anonymization, ".gt", refers to the evaluation of the k-anonymization technique on the ground truth data and ".syn", represents the evaluation of the k-anonymization technique on the synthetic data generated. The result indicates that, on average, the synthetic data has groups with a minimum of 82 individuals sharing the same attribute values, as shown in table 4.4.

For l_diversity, a value of zero suggests a lack of diversity in the sensitive attribute values within the groups. As shown in table 4.4 the results indicate that the ground truth data has no diversity in terms of distinct values for sensitive attributes. A minimum value of 1e-08 and a maximum value of 1e-08, suggesting a minimal level of diversity in the synthetic data.

TABLE 4.4: DISPLAYS THE EVALUATION RESULTS OBTAINED THROUGH THE APPLICATION OF CTGAN.

| | min | max | mean | stdev |
|---|---|---|---|---|
| stats.inv_kl_divergence.marginal | 0.9577 | 0.9951 | 0.9753 | 0.0153 |
| stats.ks_test.marginal | 0.8788 | 0.9493 | 0.9173 | 0.0291 |
| privacy.k-anonymization.gt | 112.0000 | 112.0000 | 112.0000 | 0.0000 |
| privacy.k-anonymization.syn | 69.0000 | 95.0000 | 82.3333 | 10.6249 |
| privacy.distinct l-diversity.gt | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| privacy.distinct l-diversity.syn | 1e-08 | 1e-08 | 0.0010 | 0.0010 |

# 5 Analysis and discussion

During implementation, it became evident that training the GAN required significant computational resources, especially in terms of CPU usage. This posed challenges on the local computer, which struggled to handle the demands of the machine learning application, particularly when dealing with large datasets. The computationally intensive and time-consuming nature of GAN training had a noticeable impact on the achieved results. However, the problem was reduced by accessing a remote server that provided the necessary resources for more effective GAN training and better outcomes.

The evaluation of both the DPGAN and CTGAN algorithms used the built-in metrics of SynthCity, including k-anonymization, l-diversity, KS test, and KL divergence. The results for DPGAN indicated that the synthetic data fell short of achieving a high level of k-anonymization, exhibiting relatively high uncertainty (see Table 4.3). Better results could have been achieved if DPGAN was trained on a larger dataset. This will provide a greater attribute variation, simplifying the process of identifying suitable generalizations that satisfy the k-anonymity condition. Also, by adjusting the epsilon parameter in k-anonymity can increase privacy protection by enforcing stricter privacy constraints, resulting in higher levels of k-anonymity.

Additionally, both the ground truth and synthetic data obtained a score of zero for l-diversity, indicating a lack of diversity in both datasets. These outcomes suggest that DPGAN did not fully accomplish privacy preservation, which is critical, given the privacy constraints of age, gender, and GDPR requirements. Improving the dataset variation, for instance, including a broader range of values for attributes, could have potentially enhanced the results and enable a more secure data sharing infrastructure.

However, the KL divergence and KS test results demonstrated that the synthetic data generated by DPGAN exhibited high values with small uncertainty, indicating good statistical similarity with the original data (see Table 4.3).

For CTGAN, the KL divergence and KS test provided valuable insights into the quality and similarity of the synthetic data generated. Higher KL divergence values indicated a closer match between the synthetic and original data, with the minimum and maximum values suggesting a stronger alignment. The mean value reflected a higher overall similarity, and the low standard deviation indicated a consistent and stable distribution of synthetic data. Similarly, the KS test yielded a lower KS statistic, indicating a relatively good match between the synthetic and original data, with the minimum and maximum values supporting this alignment. Although the mean value represented a reasonably good overall fit, the standard deviation suggested some variability in the alignment.

The k-anonymization and l-diversity tests underscored the privacy properties of the synthetic data generated by CTGAN (see Table 4.4). The achieved k-anonymization level indicated a reasonable degree of privacy protection, making it challenging to identify individuals within the dataset. However, the results indicated a lack of l-diversity, which aligns with the findings of the previously mentioned algorithm. Both algorithms exhibited a similar absence of diversity in the synthetic data. To address this issue, similar improvements can be implemented, as mentioned earlier for DPGAN, to strive for better and more diverse results.

## 5.1 Social, economic, ethical and sustainability aspects

The use of synthetic data enhances privacy protection and reduces the risk of sensitive information being revealed. By utilizing data-driven technologies and participating in initiatives that share data, individuals feel more secure and confident sharing their information. While adhering to GDPR, synthetic data can foster a positive social environment that promotes data sharing for research, innovation, and public interest.

Furthermore, the use of synthetic data in research and decision-making processes can affect social fairness and equity. The use of synthetic data can reduce the biases associated with real datasets, leading to more objective and fair results. By minimizing the dependence on sensitive personal information, synthetic data can contribute to the development of fair algorithms and regulations.

By utilizing synthetic data, organizations can reduce costs associated with data collection, storage, and maintenance. Generating synthetic data decreases the need for extensive data collection efforts, particularly in the case of large or sensitive datasets. This can result in significant cost savings in the long term.

Synthetic data generation tackles ethical issues concerning data privacy and security by avoiding the use of real data. By employing synthetic data instead, the potential dangers associated with exposing sensitive personal information of individuals are eliminated. This method safeguards privacy and minimizes the possible harm that may result from mishandling or unauthorized access to genuine personal data. Upholding privacy rights and preserving data confidentiality are in accordance with ethical principles and encourage responsible utilization of data.

Unlike storing large amounts of real data, which necessitates dedicated server space and energy-intensive data centers, synthetic data can be generated and utilized in real-time without the requirement for long-term storage. Therefore, the demand for data storage infrastructure is reduced, resulting in a reduction in the environmental impact associated with data center operations and maintenance.

# 6 Conclusion

The healthcare industry handles sensitive patient data, and synthetic data has emerged as a solution to protect privacy and prevent breaches. However, privacy concerns and regulations limit direct access to health data. This bachelor's thesis presents a secure data sharing infrastructure, that was designed using a known dataset, demonstrating its effectiveness through a machine learning application.

## 6.1 Evaluating goals

The first two objectives stated in chapter 1.2 were accomplished and the third objective was accomplished to a certain degree, although, there are still work that must be done in the future to further implement a more secure data sharing mechanism which will be mentioned in section 6.2. The goals were to:

- Identify and evaluate different solutions for meeting privacy requirements for data sharing.
- Develop a sample machine learning-based application that can use the secure data sharing infrastructure to demonstrate its functionality.
- Test and validate the secure data sharing infrastructure by proving that it meets integrity requirements.

The goal of testing and validating the secure data sharing infrastructure by proving that it meets integrity requirements was met to a certain extent by evaluating two GAN algorithms, DPGAN and CTGAN, using various metrics that provided valuable insights into their privacy preservation capabilities. These findings contribute to the knowledge development in privacy-preserving data sharing, particularly in the context of healthcare. The study's outcomes inform researchers and practitioners about the challenges and possibilities associated with synthetic data generation and its impact on privacy preservation. Furthermore, the implementation of a secure data sharing infrastructure and the evaluation of GAN algorithms serve as a foundation for future research and the development of improved methodologies in privacy-preserving data sharing. Overall, this research expands the understanding of privacy requirements in data sharing and provides insights into the effectiveness of different solutions, paving the way for further advancements in the field.

## 6.2 Future works

Future work could involve expanding the project’s scope to larger datasets and more advanced machine learning algorithms, exploiting new problem areas such as integrating synthetic data with real-world dataset or improving the accuracy of synthetic data. Additionally, future works could consider the trade-off between privacy and utility, aiming to strike a balance that maximizes the usefulness of the data while safeguarding individuals' privacy. Overall, this study lays a solid foundation for future research and development in secure data sharing in the healthcare industry.

# Reference


1. Steindler D. Faculty of Law Protection of Personal Data in Healthcare and New Technologies [Internet]. [cited 2023 May 15]. Available from: https://dspace.cuni.cz/bitstream/handle/20.500.11956/176174/120407664.pdf?sequence=1&isAllowed=y

2. What is personal data? [Internet]. European Commission. [cited 2023 Mar 31]. Available from:https://commission.europa.eu/law/law-topic/data-protection/reform/what-personal-data_en

3. Sensitive personal data [Internet]. Imy. [cited 2023 Mar 31]. Available from: https://www.imy.se/en/individuals/data-protection/introduktion-till-gdpr/what-is-actually-meant-by-personal-data/what-is-meant-by-sensitive-personal-data/

4. Canada O of the PC of. PIPEDA in brief [Internet]. www.priv.gc.ca. 2018 [cited 2023 Mar 31]. Available from: https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/pipeda_brief/#_h2

5. Jordon J, Szpruch L, Houssiau F, Bottarelli M, Cherubin G, Maple C, et al. Synthetic Data - what, why and how? [Internet]. 2022. Available from: https://arxiv.org/pdf/2205.03257.pdf

6. Wagner P. Privacy Enhancing Technologies and Synthetic Data [Internet]. papers.ssrn.com. Rochester, NY; 2020. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3762686

7. A Review Of Synthetic Data Generation Methods For Privacy Preserving Data Publishing [Internet]. [cited 2023 Apr 3]. Available from: https://www.ijstr.org/final-print/mar2017/A-Review-Of-Synthetic-Data-Generation-Methods-For-Privacy-Preserving-Data-Publishing.pdf

8. Lu Y, Tech V, Wang U, Wei W, Wang H. Machine Learning for Synthetic Data Generation: a Review [Internet]. Available from: https://arxiv.org/pdf/2302.04062.pdf

9. El Emam K. Seven Ways to Evaluate the Utility of Synthetic Data. IEEE Security & Privacy. 2020 Jul;18(4):56–9.

10. Wu G, Zhang H, He Y, Bao X, Li L, Hu X. Learning Kullback-Leibler Divergence-Based Gaussian Model for Multivariate Time Series Classification. IEEE Access [Internet]. 2019 [cited 2023 May 16];7:139580–91. Available from: https://ieeexplore.ieee.org/abstract/document/8847405

11. Cong Z, Chu L, Yang Y, Pei J. Comprehensible counterfactual explanation on Kolmogorov-Smirnov test. Proceedings of the VLDB Endowment. 2021 May;14(9):1583–96.

12. Karle T, Vora D. PRIVACY preservation in big data using anonymization techniques [Internet]. IEEE Xplore. 2017 [cited 2023 May 16]. p. 340–3. Available from: https://ieeexplore.ieee.org/abstract/document/8073538

13. Rajendran, K., Jayabalan, M., & Rana, M. E. (2017). A study on k-anonymity, l-diversity, and t-closeness techniques. International Journal of Computer Science and Network Security, 17(12), 172.

14. Dai B, Wang Y, Aston J, Hua G, Wipf D. Hidden Talents of the Variational Autoencoder. arXiv:170605148 [cs] [Internet]. 2019 Oct 7 [cited 2023 May 15]; Available from: https://arxiv.org/abs/1706.05148

15. Gonog L, Zhou Y. A Review: Generative Adversarial Networks. 2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA). 2019 Jun;

16. Eigenschink P, Vamosi S, Vamosi R, Sun C, Reutterer T, Kalcher K. Deep Generative Models for Synthetic Data. Deep Generative Models for Synthetic Data [Internet]. 2021; Available from: https://research.wu.ac.at/en/publications/deep-generative-models-for-synthetic-data-5

17. Young J, Graham P, Penny R. Using Bayesian Networks to Create Synthetic Data [Internet]. [cited 2023 May 18]. Available from: https://www.scb.se/contentassets/ff271eeeca694f47ae99b942de61df83/using-bayesian-networks-to-create-synthetic-data.pdf

18. Xie L, Lin K, Wang S, Wang F, Zhou J. Differentially Private Generative Adversarial Network. arXiv:180206739 [cs, stat] [Internet]. 2018 Feb 19; Available from: https://arxiv.org/abs/1802.06739

19. Liu Y, Peng J, Yu J, Wu Y. PPGAN: PRIVACY-PRESERVING GENERATIVE ADVERSARIAL NETWORK A PREPRINT [Internet]. 2019 [cited 2023 May 16]. Available from: https://arxiv.org/pdf/1910.02007.pdf

20. Mao X, Li Q, Xie H, Lau RYK, Wang Z, Smolley SP. Least Squares Generative Adversarial Networks. arXiv:161104076 [cs] [Internet]. 2017 Apr 5 [cited 2023 May 16]; Available from: https://arxiv.org/abs/1611.04076

21. Belén Vega-Márquez, Rubio-Escudero C, Riquelme JC, Nepomuceno-Chamorro IA. Creation of Synthetic Data with Conditional Generative Adversarial Networks. 2019 May 13;231–40.

22. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling Tabular data using Conditional GAN [Internet]. Vol. 32, Neural Information Processing Systems. Curran Associates, Inc.; 2019. Available from: https://proceedings.neurips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html

23. Nowok B, Raab GM, Dibben C. synthpop: Bespoke Creation of Synthetic Data in R. Journal of Statistical Software [Internet]. 2016 Oct 28 [cited 2022 Oct 10];74:1–26. Available from: https://www.jstatsoft.org/article/view/v074i11

24. Ping H, Stoyanovich J, Howe B. DataSynthesizer: Privacy-Preserving Synthetic Datasets. In: Proceedings of the 29th International Conference on Scientific and Statistical Database Management (SSDBM '17). New York, NY, USA: Association for Computing Machinery; 2017. p. 1–5. Article 42. Available from: https://doi.org/10.1145/3085504.3091117

25. Qian Z, Cebere BC, van der Schaar M. Synthcity: facilitating innovative use cases of synthetic data in different data modalities. arXiv:230107573 [cs] [Internet]. 2023 Jan 18 [cited 2023 May 16]; Available from: https://arxiv.org/abs/2301.07573

26. Murtaza H, Ahmed M, Khan NF, Murtaza G, Zafar S, Bano A. Synthetic data generation: State of the art in health care domain. Computer Science Review. 2023 May;48:100546.

27. Imtiaz S, Arsalan M, Vlassov V, Sadre R. Synthetic and Private Smart Health Care Data Generation using GANs [Internet]. IEEE Xplore. 2021 [cited 2023 May 15]. p. 1–7. Available from: https://ieeexplore.ieee.org/document/9522203

TRITA – CBH-GRU-2023:089