
KTH Stockholm
Department of Numerical Analysis

# How can we generate private heartbeat data for arrhythmia detection?

Master Thesis Report

Sijun John Tu

Supervisors: Anders Szepessy (KTH) and Shahid Raza (RISE)

**Abstract**

# STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

January 17, 2024

Sijun John Tu

# Contents

# Notation

Mathematical conventions and notation used in this thesis:

| | |
|---|---|
| $\mathbb{R}$ | the real numbers |
| $\sqcup$ | disjoint set union |

Additionally, we introduce the following conventions to describe various elements from different mathematical objects to make the notations and their meaning as consistent as possible:

| | |
|---|---|
| $\mathcal{S}$ | set of heartbeat samples |
| $s_i \in \mathbb{R}^L$ | seqeunce of ECG measurements |

# List of Figures

# 1 Introduction

## 1.1 Motivation

Data-driven technology and especially machine learning have gained a lot of momentum the past years. Models like ChatGPT or BERT heavily depend on large datasets that are available publicly. At the same time machine learning models are now being considered in other data sensitive domains like health care [see 5, 6, 18, 22]. One exciting field within health care is arrhythmia detection for heartbeats, where machine learning methods can aid physicians to detect irregular heartbeat conditions. Recently, several methods have been proposed, ranging from SVMs to neural networks [see review 4].

When working with those sensitive data, privacy plays a major role in general acceptance of those models. Now governmental institutions like the European Union have established a right to privacy manifested in the General Data Protection Regulation laws [1]. Previous simple anonymisation attempts that simply removed some identifying attributes (e. g. name, birthday etc.) have been proven to be ineffective. For example, user profiles from the anonymised dataset used in the infamous Netflix prize have been reconstructed with the help of publicly available data from IMDB [16]. This is why technological advances in the area of privacy-preserving machine learning have increased in the past few years, with the development of various machine learning models that aim to preserve the privacy of individual data records. Protecting privacy becomes crucial for heartbeat data because it can be used to identify patients, thus heavily impacting the patient's privacy [see heartbeat biometrics 21, 11].

One promising solution [see 13] is to replace the original, possibly sensitive data set with a synthetic data set that resembles the original raw data in some statistical properties. Much research has been done to generate tabular or image data ???REF, whereas dedicated time series data generation is still a *"burgeoning"* area of research according to a recent benchmark [3]. Regardless of the data type, data generators with no formal privacy guarantees have been shown to still be susceptible to privacy leaks [20].
To improve privacy, this thesis aims to analyse the combination of synthetic data with tools from so-called differential privacy. Differential privacy has been developed by Dwork et al [7] and is widely considered as the mathematical framework to rigorously provide privacy guarantees to privacy-preserving algorithms, relying on applied probability theory and statistics. This thesis will study existing architectures based on private generative AI models, as well as explore the possibility of new solution. Experiments were conducted to assess the performance of these models using the MITBIH dataset on heartbeat arrhythmia [15]. Unfortunately, there is no free lunch and privacy always comes with a decrease in utility [20]. A careful balance between privacy and utility needs to be established.

[1] see https://gdpr-info.eu/

## 1.2 Problem Definition

This thesis aims to examine how to generate private time series data for heartbeat arrhythmia detection. Let $\mathcal{S} = \{s_i\}_{i=1}^{N}$ denote a set of heartbeat samples, where $s_i = (s_i^0, ..., s_i^L)$ is a sequence of one-dimensional ECG measurements of fixed length $L$ corresponding to one heartbeat. Each heartbeat sequence is associated with a corresponding label denoting whether it is a normal or anomalous heartbeat according to ???. Therefore we separate the set into normal heartbeats $\mathcal{N}$ and $\mathcal{A}$ (i. e. $\mathcal{S} = \mathcal{N} \sqcup \mathcal{A}$)

Firstly, we want to design a time series generator (TSG) that can model the true probability distribution $p(\mathcal{N})$ of the normal heartbeats. Here, we only consider normal heartbeats since for the subsequent task of arrhythmia detection we will follow an anomaly detection approach explained next. The aim of the TSG is to generate a synthetic data set $\widehat{\mathcal{N}}$ with distribution $p(\widehat{\mathcal{N}})$ that is "close" to the original data $p(\mathcal{N})$.
Secondly, the utility of the generated data is assessed in the downstream task of detecting anomalous heartbeats (heartbeat arrhythmia detection). We treat this task as an anomaly detection task based on reconstruction error ??REF, i. e. we want to train a model only on normal heartbeats that can reconstruct those samples with low error, but give high reconstruction error when inputting an anomalous sample. Alternatively, one could treat this as a binary classification task, that classifies a given heartbeat sample as either normal or anomalous. Since the ratio of those two classes are heavily imbalanced due to the nature of arrhythmias, we will favor the first approach ??REF.
Lastly, we will embed the generation procedure in a differential privacy setting. This will provide a theoretical framework to assess privacy.

## 1.3 Related Works and State of the Art

TBA

- sota privacy in ml
- why DP
- private data generation
- heartbeat data generation

### Privacy in machine learning

A lot of past efforts have been put into improving the performance of machine learning methods, where the privacy aspect has been neglected. Due to the increased awareness about private individual data and policies like EU's GDPR laws, big tech companies like Apple, Google and even the US Census have been implementing privacy measurements in the their data collection [see 8, 2]. One of the first groundbreaking works on actually quantifying the privacy leakage in machine learning models has been studied in [19], where Shokri et. al. have designed a framework to perform membership inference attacks (MIA) on basic classification tasks. MIA

on machine learning models try to infer whether a certain record has been used when training the respective model. This becomes a privacy issue when e. g. an adversary can infer whether a certain patient's data was used to train a model associated with a disease. Then the adversary can conclude that this particular patient likely has this disease [cf. 19, p. 5]. Hence, their results indicate a strong vulnerability in terms of privacy for data-based models.
Several notions of privacy have been proposed in the last decade, among which Differential Privacy (DP) has emerged as the *"de-facto standard in data privacy"* [14]. Reasons for its popularity according to a recent survey [9] are among others:

1. DP is future-proof and requires no extra knowledge about the adversary.
2. DP provides rigorous privacy guarantees.
3. DP provides a notion of privacy budget, which can be adapted to the specific use case to balance privacy and utility.

We will revisit the definition and most important results in Chapter 2 of this thesis. The basic idea is to add calibrated, random noise either the data or during model training.
Recently, a lot of popular neural network architectures have been "privatised" by adding DP noise, most notably there is a differential private version of stochastic gradient descent (SGD) called DP-SGD developed by Abadi et al [1] in 2016. Broadly speaking, differential private noise can be injected in three different stages of the modelling pipeline: input, hidden or output layer [cf. 23].
Applying some DP mechanism at the input stage can be seen as preprocessing step to either hide sensitive attributes in the data or generating new synthetic dataset, which will be the focus of this thesis. Hence, it will not be discussed in this chapter.
- hidden layer
- output layer

**Data generation and Privacy**

# 2 Theoretical background on Differential Privacy

In this chapter we briefly describe and derive the most important results from Cynthia Dwork's work on differential privacy that was first introduced in ??. This summary heavily relies on her writings in her as well as lecture notes from ????.

## 2.1 Defining differential privacy

Differential privacy (DP) should be understood as an agreement between the data holder and the data subject: the latter should not be "affected, adversely or otherwise, by allowing [her] data to be used in any study or analysis, no matter what other studies, data sets or information sources are available". This addresses the paradox of learning something useful about a population while learning nothing about the individuals

**Example 2.1.1** (Randomised response)**.** citation needed ??? proposes the following random answering procedure: In a study where participants are asked to answer with "Yes" or "No" whether they have engaged in an illegal or embarrassing activity $A$, they should:

1. Flip a coin
2. If the coin shows tails, then the participant should respond truthfully.
3. If the coin shows head, then the participant should flip the coin a second time and answer "Yes" if the second coin shows head and "no" otherwise.

This procedure ensures participants' privacy by "plausible deniability"; each participant's answer has non-zero probability of being truthful or not. By understanding the probabilities of the noise generation process, the data analyst can estimate the true number of "yes" and "no" answers. To this end, let $p$ b the true percentage of "yes" answers, $N$ the total number of participants, $n_{true}$ the true number of "yes" responses and $\hat{n}_{obs}$ the observed number of "yes" responses. We assume a fair coin with equal probability of showing heads or tails. Then the expected number of "yes" answers after applying the described procedure is:

$$\mathbb{E}(\text{"}Yes\text{"}) = \frac{1}{4}n_{true} + \frac{1}{4}(N - n_{true}) + \frac{1}{2}n_{true} = \frac{1}{4}N + \frac{n_{true}}{2} \tag{1}$$

We can estimate this using the $\hat{n}_{obs} \approx \mathbb{E}(\text{"}Yes\text{"}) = \frac{1}{4}N + \frac{n_{true}}{2}$ and finally solving for $n_{true}$ yields the estimate:

$$\hat{n_{true}} = 2\hat{n}_{obs} - \frac{1}{2}N \tag{2}$$

**Definition 2.1.2** (Probability Simplex)**.** Given a discrete set $B$, the probability simplex over $B$ is defined as the set

$$\Delta(B) = \left\{ x \in \mathbb{R}^{|B|}, x_i \geq 0 \text{ and } \sum_i x_i = 1 \right\} \tag{3}$$

**Definition 2.1.3** (Randomised Algorithm)**.** A randomized algorithm $\mathcal{M}$ with domain $A$ and discrete range $B$ is associated with a mapping $M : A \to \Delta(B)$. On input $a \in A$ algorithm $\mathcal{M}$ outputs $\mathcal{M}(a) = b$ with probability $(M(a))_b$

**Definition 2.1.4** (Histogram representation of a data base)**.** Given a set $\mathcal{X}$, the universe of all possible records, the histrogram representation of a database $x$ is the vector

$$x \in \mathbb{N}^{|\mathcal{X}|} \tag{4}$$

in which each entry $x_i$ represents the number of elements in database $x$ of type $i \in \mathcal{X}$.

**Definition 2.1.5** ($l_1$-norm of a database in histogram representation)**.** The $l_1$-norm of a database is a measure of the size of the database and defined as:

$$||x||_1 = \sum_{i=1}^{|\mathcal{X}|} |x_i| \tag{5}$$

This immediately gives rise to a notion of distance between two databases $x$ and $y$, namely:

$$||x - y||_1 = \sum_{i=1}^{|\mathcal{X}|} |x_i - y_i| \tag{6}$$

which basically counts the number of different entries.
Now we are ready to give the general definition of differential privacy:

**Definition 2.1.6** ($(\epsilon, \delta)$-DP)**.** A randomised algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ is $(\epsilon, \delta)$-differentially private if for all outcomes $S \subset ran\mathcal{M}$ and for all databases $x, y \in \mathbb{N}^{|\mathcal{X}|}$, such that $||x - y||_1$ (i. e. they only differ in one element) we have

$$\mathbb{P}(\mathcal{M}(x) \in S) \leq e^{\epsilon} \cdot \mathbb{P}(\mathcal{M}(y) \in S) + \delta \tag{7}$$

where the probability is taken over the randomness of $\mathcal{M}$. If $\delta = 0$, we say $\mathcal{M}$ is $\epsilon$-differentially private.

why $e^{\epsilon}$

**Example 2.1.7** (Randomised response revisited)**.**

## 2.2 Important results for Differential Privacy

**Theorem 2.2.1** (DP requires randomisation)**.** *Any non-trivial DP-mechanism requires randomisation.*

*Proof.* TBA □

**Theorem 2.2.2** (Post-processing)**.** *Let $\mathcal{M} : \mathbb{N}^{|\mathcal{X}|} \to R$ be a randomised algorithm that is $(\epsilon, \delta)$- DP. Further let $f : R \to R'$ an arbitrary function. Then $f \circ \mathcal{M}$ is also $(\epsilon, \delta)$ -DP.*

*Proof.* First fix data sets $x, y \in \mathbb{N}^{|\mathcal{X}|}$, s. t. $||x - y||_1 \leq 1$ and outcome $S' \subseteq R'$. Define a set $S = \{r \in R : f(r) \in S'\}$. Then we have:

$$\begin{aligned} \mathbb{P}(f(\mathcal{M}(x)) \in S') &= \mathbb{P}(\mathcal{M}(x) \in S) \\ &\leq e^{\epsilon} \cdot \mathbb{P}(\mathcal{M}(y) \in S) + \delta \\ &= e^{\epsilon} \cdot \mathbb{P}(f(\mathcal{M}(y)) \in S') + \delta \end{aligned} \tag{8}$$

where the inequality follows from the $(\epsilon, \delta) - DP$ of $\mathcal{M}$. □

**Theorem 2.2.3** (Group privacy). *Let $\mathcal{M} : \mathbb{N}^{|\mathcal{X}|} \to R$ be a randomised algorithm that is $(\epsilon, \delta)$- DP, then $\mathcal{M}$ is $(k\epsilon, ke^{k\epsilon}\delta)$- DP for groups of size $k$, i. e. it holds for databases $x, y \in \mathbb{N}^{|\mathcal{X}|}$ such that $||x - y||_1 \leq k$ and for all $S \subseteq R$:*

$$\mathbb{P}(\mathcal{M}(x) \in S) \leq e^{k\epsilon} \cdot \mathbb{P}(\mathcal{M}(y) \in S) + k\delta \tag{9}$$

*Proof.* First fix data sets $x, y \in \mathbb{N}^{|\mathcal{X}|}$, s. t. $||x - y||_1 \leq k$ and outcome $S \subseteq R$. Now there exists a series of databases $z_0, ..., z_k$, such that $x = z_0$ and $y = z_k$ and $||z_{i+1} - z_i||_1 \leq 1$, i. e. we can find a series of databases that transforms $x$ into $y$ by removing or adding one record at a time. Then we have:

$$\begin{aligned}
\mathbb{P}(\mathcal{M}(x) \in S) &= \mathbb{P}(\mathcal{M}(z_0) \in S) \\
&\leq e^{\epsilon} \cdot \mathbb{P}(\mathcal{M}(z_1) \in S) + \delta \\
&\leq e^{\epsilon} \left(e^{\epsilon} \cdot \mathbb{P}(\mathcal{M}(z_2) \in S) + \delta\right) + \delta \\
&\leq ... \\
&= ke^{\epsilon} \cdot \mathbb{P}(f(\mathcal{M}(y)) \in S') + ke^{k\epsilon}\delta
\end{aligned} \tag{10}$$

□

**Theorem 2.2.4** (Standard composition). *Let $\mathcal{M}_1 : \mathbb{N}^{|\mathcal{X}|} \to R_1$ and $\mathcal{M}_2 : \mathbb{N}^{|\mathcal{X}|} \to R_2$ be two randomised algorithms that are $(\epsilon_1, \delta_1)$- and $(\epsilon_2, \delta_2)$ DP, then their composition defined by $\mathcal{M}_{12} : \mathbb{N}^{|\mathcal{X}|} \to R_1 \times R_2$, $\mathcal{M}_{12}(x) = (\mathcal{M}_1(x), \mathcal{M}_2(x))$ is $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$ DP.*

*Proof.* TBA □

## 2.3 Example of DP-mechanism: Laplace mechanism

# 3 Time Series data generation

## 3.1 Overview

- Data generation in general
- what is special about time series
- what about Privacy
- choice of models

## 3.2 DP-MERF

DP-MERF [10] is an efficient all purpose data generation algorithm that is based on minimising the so-called Maximum Mean Discrepancy between the real and the synthetic data distributions. The authors mainly verfied their results using tabular data like ????, but also image data, notably the MNIST ???CITE data set. It has not been used for time series data, but we will consider this data generation for generating time series data in this thesis, because accorrding to a recent survey [12], DP-MERF delivers the best all purpose data generation performance.

### 3.2.1 Maximum Mean Discrepancy

There are different ways to measure the "distance" between two distributions $P$ and $Q$. On popular metric is the Maximum Mean Discrepancy (MMD) between $P$ and $Q$, where the random variables are projected into another feature space and the expected values are compared to each other in this space.

**Definition 3.2.1.1** (MMD)**.** Let $\phi : \mathcal{X} \to \mathcal{H}$, where $\mathcal{H}$ is a reproducing kernel hilbert space (RKHS) and $P$ and $Q$ some distributions over $\mathcal{X}$ and random variables $X \sim P$, $Y \sim Q$ given. Then the Maximum mean Discrepancy is defines as:

$$MMD(P, Q) = ||\mathbb{E}[\phi(X)] - \mathbb{E}[\phi(Y)]||_{\mathcal{H}} \tag{11}$$

Some "easy" features maps $\phi$ are for example:

**Example 3.2.1.2.** Let $P$ and $Q$ some distributions over $\mathcal{X}$ and random variables $X \sim P$, $Y \sim Q$ given.

- **Identity kernel**: $\mathcal{X} = \mathcal{H} = \mathbb{R}^d$ and $\phi(x) = x$, then we have:

$$\begin{aligned} MMD(P, Q) &= ||\mathbb{E}[\phi(X)] - \mathbb{E}[\phi(Y)]||_{\mathcal{H}} \\ &= ||\mathbb{E}[X] - \mathbb{E}[Y]||_{\mathbb{R}^d} \end{aligned} \tag{12}$$

  So we only compare the two distributions in terms of their means.

- **Quadratic kernel**: $\mathcal{X} = \mathbb{R}$ $\mathcal{H} = \mathbb{R}^2$ and $\phi(x) = (x, x^2)$, then we have:

$$\begin{aligned} MMD(P,Q) &= ||\mathbb{E}[\phi(X)] - \mathbb{E}[\phi(Y)]||_{\mathcal{H}} \\ &= ||\mathbb{E}[(X, X^2)] - \mathbb{E}[(Y, Y^2)]||_{\mathcal{H}} \\ &= || \begin{pmatrix} \mathbb{E}[X] \\ \mathbb{E}[X^2] \end{pmatrix} - \begin{pmatrix} \mathbb{E}[Y] \\ \mathbb{E}[Y^2] \end{pmatrix} ||_{\mathbb{R}^2} \\ &= \sqrt{(\mathbb{E}[X] - \mathbb{E}[Y])^2 - (\mathbb{E}[X^2] - \mathbb{E}[Y^2])^2} \end{aligned} \tag{13}$$

  So here we compare the two distributions in terms of their means and their variance (or first and second moments respectively).

- **Gaussian kernel** ????

Now instead of computing a possibly high or even infinite dimensional transformation $\phi$ one can use the well-known kernel trick ????REF. Let $k(x, y) =< \phi(x), \phi(y) >_{\mathcal{H}}$ be a kernel with corresponding reproducing kernel hilbert space $\mathcal{H}$, then the computation of the MMD simplifies to:

$$\begin{aligned} MMD^2(P,Q) &= ||\mathbb{E}[\phi(X)] - \mathbb{E}[\phi(Y)]||^2_{\mathcal{H}} \\ &=< \mathbb{E}[\phi(X)], \mathbb{E}[\phi(X')] > - < \mathbb{E}[\phi(X)], \mathbb{E}[\phi(Y)] > - < \mathbb{E}[\phi(Y)], \mathbb{E}[\phi(X)] > \\ &\qquad + < \mathbb{E}[\phi(Y)], \mathbb{E}[\phi(Y')] > \\ &= \mathbb{E}[< \phi(X), \phi(X') >] - 2\mathbb{E}[< \phi(X), \phi(Y) >] + \mathbb{E}[< \phi(Y), \phi(Y') >] \\ &= \mathbb{E}[k(X, X')] - 2\mathbb{E}[k(X, Y)] + \mathbb{E}[k(Y, Y')] \end{aligned} \tag{14}$$

Where we introduced independent random variables $X, X' \sim P$, $Y, Y' \sim Q$.
Now given a training data set $X_m = \{x_i\}_{i=1}^m \sim P$ and a synthetic data set $X'_m = \{x_i\}_{i=1}^m \sim Q$ we can estimate their $MMD^2$ by estimating the expected value with a mean estimate:

$$\widehat{MMD}^2(X_m, X'_m) = \frac{1}{m^2} \sum_{i,j=1}^{m} k(x_i, x_j) + \frac{1}{m^2} \sum_{i,j=1}^{m} k(x'_i, x'_j) - \frac{2}{m^2} \sum_{i,j=1}^{m} k(x_i, x'_j) \tag{15}$$

Unfortunately, this will require $\mathcal{O}(m^2)$ computations which grows quadratically in the number of samples. This will be too big for a large training data set. As a remedy, the authors of [10] propose to use Random Fourier Features based on a paper from 2007 [see 17], to approximate the kernel $k$ using its fourier transform and Monte-Carlo-Simulation. Thus,

$$k(x, y) \approx \hat{\Phi}(x)^T \hat{\Phi}(x') \tag{16}$$

where $\hat{\Phi}(x) \in \mathbb{R}^D$ and $\hat{\Phi}_j(x) = \sqrt{\frac{2}{D}} cos(\omega_j^T x)$.

## 3.3 GAN based

# 4 Models

describing the models used in this work and why

## 4.1 AE-DPMERF

## 4.2 RTSGAN

# 5 Experiment

## 5.1 Experiment setup

## 5.2 Results

# 6 Discussion

# 7 Outro

## 7.1 Future Works

## 7.2 Conclusion

# Appendix

# References


[1] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. "Deep Learning with Differential Privacy". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. CCS'16. ACM, Oct. 2016. URL: http://dx.doi.org/10.1145/2976749.2978318.

[2] Abowd, J., Ashmead, R., Simson, G., Kifer, D., Leclerc, P., Machanavajjhala, A., and Sexton, W. "Census topdown: Differentially private data, incremental schemas, and consistency with public knowledge". In: *US Census Bureau* (2019).

[3] Ang, Y., Huang, Q., Bao, Y., Tung, A. K. H., and Huang, Z. *TSGBench: Time Series Generation Benchmark*. 2023. arXiv: 2309.03755 [cs.LG].

[4] Apandi, Z. F. M., Ikeura, R., and Hayakawa, S. "Arrhythmia Detection Using MIT-BIH Dataset: A Review". In: *2018 International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA)*. 2018, pp. 1–5.

[5] Beam, A. L. and Kohane, I. S. "Big Data and Machine Learning in Health Care". In: *JAMA* 319.13 (Apr. 2018), pp. 1317–1318. eprint: https://jamanetwork.com/journals/jama/articlepdf/2675024/jama_beam_2018_vp_170174.pdf. URL: https://doi.org/10.1001/jama.2017.18391.

[6] Bhardwaj, R., Nambiar, A. R., and Dutta, D. "A Study of Machine Learning in Healthcare". In: *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*. Vol. 2. 2017, pp. 236–241.

[7] Dwork, C. "A Firm Foundation for Private Data Analysis". In: *Commun. ACM* 54.1 (Jan. 2011), pp. 86–95. URL: https://doi.org/10.1145/1866739.1866758.

[8] Dwork, C., Kohli, N., and Mulligan, D. "Differential privacy in practice: Expose your epsilons!" In: *Journal of Privacy and Confidentiality* 9.2 (2019).

[9] Gong, M., Xie, Y., Pan, K., Feng, K., and Qin, A. "A Survey on Differentially Private Machine Learning [Review Article]". In: *IEEE Computational Intelligence Magazine* 15.2 (2020), pp. 49–64.

[10] Harder, F., Adamczewski, K., and Park, M. "Differentially Private Mean Embeddings with Random Features (DP-MERF) for Simple & Practical Synthetic Data Generation". In: *CoRR* abs/2002.11603 (2020). arXiv: 2002.11603. URL: https://arxiv.org/abs/2002.11603.

[11] Hegde, C., Prabhu, H. R., Sagar, D., Shenoy, P. D., Venugopal, K., and Patnaik, L. M. "Heartbeat biometrics for human authentication". In: *Signal, Image and Video Processing* 5 (2011), pp. 485–493.

[12] Hu, Y., Wu, F., Li, Q., Long, Y., Garrido, G. M., Ge, C., Ding, B., Forsyth, D., Li, B., and Song, D. *SoK: Privacy-Preserving Data Synthesis*. 2023. arXiv: 2307.02106 [cs.CR].

[13] Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., and Weller, A. *Synthetic Data – what, why and how?* 2022. arXiv: 2205.03257 [cs.LG].

[14] Kim, J. W., Edemacu, K., Kim, J. S., Chung, Y. D., and Jang, B. "A survey of differential privacy-based techniques and their applicability to location-based services". In: *Computers & Security* 111 (2021), p. 102464.

[15] Moody, G. B. and Mark, R. G. "The impact of the MIT-BIH arrhythmia database". In: *IEEE engineering in medicine and biology magazine* 20.3 (2001), pp. 45–50.

[16] Narayanan, A. and Shmatikov, V. "Robust De-anonymization of Large Sparse Datasets". In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. 2008, pp. 111–125.

[17] Rahimi, A. and Recht, B. "Random Features for Large-Scale Kernel Machines". In: *Advances in Neural Information Processing Systems*. Ed. by Platt, J., Koller, D., Singer, Y., and Roweis, S. Vol. 20. Curran Associates, Inc., 2007. URL: https://proceedings.neurips.cc/paper_files/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf.

[18] Shailaja, K., Seetharamulu, B., and Jabbar, M. "Machine learning in healthcare: A review". In: *2018 Second international conference on electronics, communication and aerospace technology (ICECA)*. IEEE. 2018, pp. 910–914.

[19] Shokri, R., Stronati, M., Song, C., and Shmatikov, V. *Membership Inference Attacks against Machine Learning Models*. 2017. arXiv: 1610.05820 [cs.CR].

[20] Stadler, T., Oprisanu, B., and Troncoso, C. *Synthetic Data – Anonymisation Groundhog Day*. 2022. arXiv: 2011.07018 [cs.LG].

[21] Wang, L., Huang, K., Sun, K., Wang, W., Tian, C., Xie, L., and Gu, Q. "Unlock with Your Heart: Heartbeat-Based Authentication on Commercial Mobile Phones". In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2.3 (Sept. 2018). URL: https://doi.org/10.1145/3264950.

[22] Wiens, J. and Shenoy, E. S. "Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology". In: *Clinical infectious diseases* 66.1 (2018), pp. 149–153.

[23] Zhao, J., Chen, Y., and Zhang, W. "Differential privacy preservation in deep learning: Challenges, opportunities and solutions". In: *IEEE Access* 7 (2019), pp. 48901–48911.