TU WIEN Informatics

Privacy-Preserving Data Sharing

Identifying Records at Risk for Membership Inference Attacks Against Synthetic Data

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

im Rahmen des Studiums

Data Science

eingereicht von

Nina Niederhametner, BSc.
Matrikelnummer 11718286

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Ao.univ.Prof. Dr. Andreas Rauber
Mitwirkung: Univ.Lektor Mag.rer.soc.oec. Dipl.-Ing. Rudolf Mayer

Wien, 4. Dezember 2023

Nina Niederhametner	Andreas Rauber

TU WIEN Informatics

Privacy-Preserving Data Sharing

Identifying Records at Risk for Membership Inference Attacks Against Synthetic Data

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieurin

in

Data Science

by

Nina Niederhametner, BSc.
Registration Number 11718286

to the Faculty of Informatics

at the TU Wien

Advisor: Ao.univ.Prof. Dr. Andreas Rauber
Assistance: Univ.Lektor Mag.rer.soc.oec. Dipl.-Ing. Rudolf Mayer

Vienna, 4th December, 2023

Nina Niederhametner	Andreas Rauber

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Nina Niederhametner, BSc.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 4. Dezember 2023

Nina Niederhametner

Acknowledgements

I would like to express my deepest gratitude to all those who have contributed to the completion of this master's thesis. This journey would not have been possible without the support and encouragement of my family, friend and colleagues.

A special thanks goes out to my family for their unconditional love and encouragement. Their understanding, patience, and belief in my abilities have helped me throughout my entire academic journey.

I extend my sincere appreciation to my co-supervisor Rudolf Mayer, whose guidance and expertise have been invaluable throughout this research. His expertise and constant encouragement have played a key role in shaping the outcome of this work. Finally, I would like to thank my supervisor Professor Andreas Rauber for his time and valuable feedback, which have immensely contributed the quality of my thesis.

Kurzfassung

Mit der stetig wachsenden Menge an verfügbarer Daten nimmt die Nachfrage nach datenschutzerhaltenden Maßnahmen immer mehr zu. Die Verwendung synthetischer Daten als Maßnahme zur Wahrung des Datenschutzes von Mikrodaten gewinnt immer mehr an Popularität, insbesondere aufgrund ihrer Fähigkeit, die Qualität der Daten, und somit den Datennutzen zu erhalten. Gleichzeitig versucht man mit synthetischen Daten Datenschutzrisiken, die durch die Veröffentlichung entstehen, zu reduzieren. Synthetische Daten werden von einem Modell, welches mit realen Daten trainiert wurde, generiert. Das bedeutet, dass die Beobachtungen in den synthetischen Daten nicht direkt einem einzelnen Individuum im ursprünglichen Datensatz entsprechen. Dies sorgt dafür, dass synthetische Daten weniger anfällig für die Verknüpfung von Datensätzen oder die Re-identifikation sind. Trotz dieses Vorteils haben jüngste Studien potenzielle Risiken synthetischer Daten aufgedeckt. Diese Studien zeigen, dass synthetische Daten nicht immun gegen sogenannte *Membership Inference Attacks* (MIA) sind. Diese Attacken, oder auch Angriffe, versuchen zu ermitteln, ob ein bestimmtes Individuum zum Trainieren eines Modells verwendet wurde. Der Fokus dieser Arbeit liegt darin, die Angreifbarkeit von Modellen, die synthetische Daten generieren, zu evaluieren und besonders gefährdete Individuen zu identifizieren. Wir erweitern bereits veröffentlichte Arbeiten, indem wir das Risiko jedes Individuums quantifizieren und mithilfe statistischer Tests bewerten, ob Ausreißer im Vergleich zu Nicht-Ausreißern anfälliger für die Angriffe sind. Darüber hinaus schlagen wir vor, Individuen, die einem hohen Risiko für MIA ausgesetzt sind, aus dem Trainingsdatensatz zu entfernen, um sich gegen die Angriffe zu verteidigen. Wir analysieren die Effizienz dieser Angriffsverteidigung und evaluieren in wie fern diese Verteidigung jeweils Datennutzen und Datenschutz, der durch die Verteidigung eingeführt wird, beeinflusst.

Abstract

As the volume of available data continues to surge, the demand for privacy-preserving measures intensifies. The use of synthetic data as a privacy-preserving measure for micro-data is gaining increasing popularity, especially due to its ability to maintain data utility while aiming to reduce disclosure risks. Synthetic data is artificially generated by a model that has been trained on real data. This means that the observations in the synthetic data do not directly correspond to any individual in the original dataset, making it less susceptible to record linkage or re-identification.

Despite this advantage, recent studies have revealed potential risks related to membership disclosure, which can occur through membership inference attacks (MIA) that aim to determine if a specific record was used to train a model when publishing synthetic micro-level data. This thesis explores the potential of synthetic data as a solution to privacy-preserving data publishing. We extend prior work by quantifying the risk of each record's membership being correctly inferred, and, using statistical tests, assessing whether outliers are more vulnerable to the attack compared to inliers. Furthermore, we propose to remove records that are at high risk for membership inference attacks from the training set as a defense against the attacks and evaluate the defense performance and quantify the utility-privacy trade-off introduced by the defense.

Contents

CHAPTER 1

Introduction

In data publishing, preserving individuals' privacy is crucial. This chapter describes the motivation behind privacy-preserving data publishing and the problems that come with it. We highlight the threat of membership inference attacks and define the research questions that this thesis aims to answer.

1.1 Motivation

With an ever-growing amount of data available, the call for privacy-preserving data measures increases. Extensive research has focused on making all kinds of data, such as image, text, and tabular, private. This thesis addresses the challenges of preserving privacy in micro-level tabular data, where each data row (also called record) represents one individual. While researchers want to make the data they collected publicly available, individuals represented in the data rely on staying anonymous. Previously, researchers and data holders widely believed that it is sufficient to remove unique identifiers (attributes that uniquely identify an individual, e.g. social security number) or alter quasi identifiers (attributes that become a unique identifier when combined with other attributes, e.g. name plus address) from the data to preserve its privacy. As Sweeney et al. [1] in 1997 showed, this does not suffice, as public data that was once believed to be anonymous has been used to re-identify individuals with little to no background knowledge about the individuals in the data. Re-identification is the process of identifying an individual in a data set that was believed to be anonymous. This can be done by so-called record linkage, where data from multiple sources that refer to the same individual, are found. Empirical evidence on this issue has been published by [2] using mobility data, [3] with health care data, [4] with movie watching and reviews, and [5] with credit card data. Ideally, publicly available micro-data derived from individuals should be resistant to privacy attacks, while keeping the data utility high. The authors of [6] state that "privacy" relates to "the ability to learn about individuals", "utility" defines "the ability to learn aggregate

statistics about large groups of individuals". Typically, when trying to make micro-data private, e.g. sanitization, perturbation, or synthetization, we will face a privacy-utility trade-off. Brickell et al. [7] measure this trade-off by quantifying the privacy as how much an adversary can learn about individuals from the anonymized data, and the utility as accuracy for data-mining algorithms applied to the same anonymized data. While methods like aggregation (e.g. group means) make the data more private, the utility suffers substantially.

1.2 Problem Statement

One promising approach to privacy-preserving data publishing is synthetic data. Synthetic data is artificially generated by a model that has been trained on real data and tries to mimic the properties of the real data. This means that the observations in the synthetic data do not directly correspond to any individual in the original dataset, making synthetic data less prone to record linkage or re-identification. Because of this, data synthesizing models are widely used across various domains like healthcare, finance, and research, where sharing or publishing sensitive data is essential for analysis, research, or collaboration while protecting individual privacy. Although the risk of linking an individual to a data record does not remain with synthetic data, membership disclosure might still pose a threat when publishing synthetic micro-level data, as recent studies show [8, 9, 10]. An individual's membership can be disclosed by so-called membership inference attacks (MIA). These are methods used to determine if a target record was used to train a model, referred to as the target model, and make use of that machine learning (ML) models leaking information about the records contained in the model's training set [11].

Membership inference attacks on supervised machine learning models were first introduced by Shokri et al. [11] in 2017. Since then, most research on MIA has focused on supervised machine learning models. Since the first work on MIA on generative models has been published [12], most research on MIA on synthesis models has focused on synthetic image data, using generative adversarial networks (GAN) as a synthesizer [13, 14, 12]. Only recently, researchers started analyzing the attacks on synthetic tabular data [8, 9, 10, 15]. According to recent studies [8, 9], synthetic data might be more susceptible to membership inference attacks than initially assumed. However, it is assumed that not all records are equally vulnerable to the attacks: some researchers claim that outlier records are particularly at a high risk for membership inference attacks [8, 16].

The goal of this thesis is to go beyond the work of [8] and quantify the risk of each record being inferred correctly, in order to find out which records are at a higher risk for MIA. While the authors of [8] test the hypothesis that outliers are more at risk for membership disclosure for only a small number of outliers, this thesis will consider all records and assess their risk for MIA. Outliers will be detected using different methods. This way, we can conclude if there is a significant difference in disclosure risk for out- versus inliers. Furthermore, we build on the unsupervised approach published by [9], and

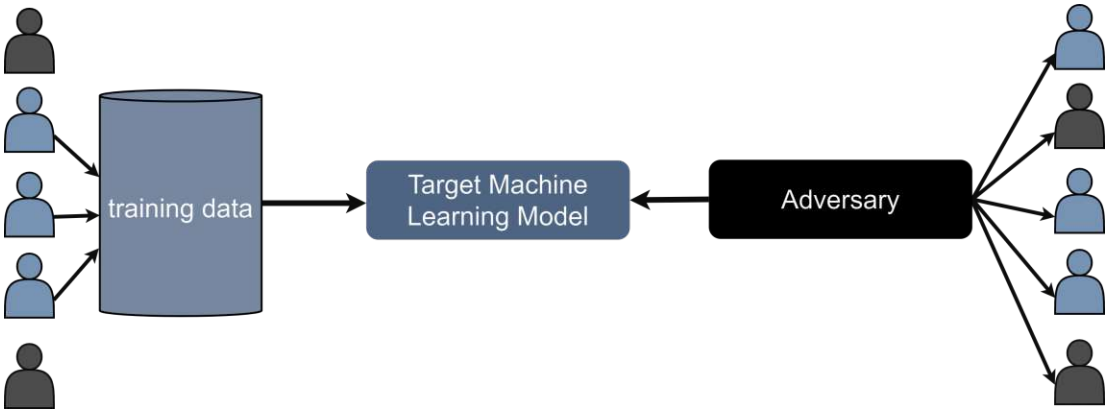

Figure 1.1: Basic Architecture of a Membership Inference Attack against a Machine Learning Model: an adversary creates an attack model that is specifically designed to find out whether individuals' data records were used to train the target model. Blue individuals represent members, while gray individuals represent non-members of the training data. When the adversary attacks the target machine learning model, they label the individuals according to whether they believe they were in- or outside the training set.

design a concrete attack approach that includes finding a suitable parameter value used to determine a record's membership, as this process was disregarded in their presented work.

As defense against membership inference attacks, we propose to identify and remove records at risk from the original training data to find out if this makes an attack less successful. For this, we also quantify and compare the data utility of synthetic data generated by a model that has been trained on the original data containing all records and a model that has had access to only records not at risk. This way, we can estimate the utility-privacy trade-off introduced by the defense. Although removing outliers from well-balanced data sets might not be a big issue, and could sometimes even be beneficial to the outcome as the model might generalize better, this is likely not the case for anomaly data sets, where the target classes are highly imbalanced. Outlier detection methods, specifically Local Outlier Factor (LOF) [17] and Isolation Forest (iForest) [18], are used to identify outliers and determine their risk for membership attacks.

1.3 Research Questions

We define three research questions, which this thesis will try to answer:

1. **To what extent can we identify records at risk for membership inference attacks by detecting outliers with algorithms like Local Outlier Factor and Isolation Forest?**

[8] uses a supervised approach, where classifiers label records as members or non-members. The prediction probability of these classifiers can be used to measure the attack risk. The distance-based approach by [9] calculates the distance from a target record to its closest synthetic record. Based on this distance, the risk for a record to be inferred correctly can be estimated. We then analyse if the outlier detection algorithms can predict the records most vulnerable against the attacks by using evaluation metrics recall and precision.

a) **By removing these records from the original training data as defense against the attack, to what extent can the threat of the MIA on the records at risk be decreased?**
For this, we will repeat the attack after removing the records at risk, and evaluate whether the MIA risk for the removed records decreases, by comparing the attack accuracy before and after the defense over all records at risk. Additionally, we compare the risk score distribution before and after the defense.

b) **To what extent are the remaining data records affected by removing the records at risk?**
As in research question 1a, we compare the overall attack accuracy of the remaining records and the risk score distribution before and after the defense.

2. **To what extent does the data utility suffer when synthesis models learn from the original data excluding the records at risk?**
To measure utility loss, the accuracy, precision, recall, and F1-measure will be evaluated for each data set according to their original machine learning task similarly to [19].

a) **By how much does the utility for synthetic data learned from imbalanced data sets decrease compared to synthetic data generated from a balanced data set?**
For highly imbalanced data, the evaluation metrics precision and recall are especially important, as highly imbalanced data sets are particularly affected by the precision-recall trade-off, and will therefore give a better understanding of the quality of a model, compared to more general performance metrics like accuracy.

b) **To what extent does the utility on the minority class of the imbalanced data suffer compared to the majority class?**
For this, we will compute evaluation scores per class and compare them. Here, we expect the utility of the minority class to decrease substantially, compared to the majority class.

3. **Which data synthesizing models generate synthetic data that is more vulnerable to MIA?**
For assessing the success of the membership attacks, the overall accuracy will be computed and then compared for the different synthesising models.

The rest of this thesis structured as follows: In Chapter 2, we explain theoretical concepts for data anonymization, outlier detection and and data synthetisation methods, as well as membership inference attacks. In Chapter 3, state-of-the-art approaches to membership inference attacks on synthesizers are explained. Furthermore we describe our MIA experiment settings including attack methods and assumptions. Chapter 4 contains the analysis and evaluation of our experiment results. In Chapter 5, we summarize our findings by answering the research question, list our main contributions and discuss future work.

CHAPTER 2

Background

This chapter contains the background relevant to the thesis. First, an introduction to data anonymization followed by an overview of different anonymization techniques that were proposed over time is given. We then provide an introduction to synthetic data and its generation process. Furthermore, the concept of membership inference attacks and different approaches for it are given. Lastly, we discuss outliers and methods to detect them. For this literature review, the guidelines for systematic literature reviews proposed by [20] will be followed.

2.1 Privacy and Inference Attacks

Public micro-data can fall victim to privacy attacks, which can lead to disclosure, where an adversary infers hidden information about the subjects contained in the data. Disclosure risks have been broadly grouped into three types: identity disclosure, attribute disclosure, and inferential disclosure [21].

Identity disclosure happens when an adversary links a data record to an individual or entity. This can be performed by linking a record to some externally available information.

Attribute disclosure is the process of finding out a characteristic, i.e. attribute value, of an individual without having to rely on external data or record identification. This could for example happen when all individuals sharing a characteristic for one attribute, also share a characteristic for another attribute. Say all male patients aged 80 or older suffer from high blood pressure, an adversary can then infer that a male patient above 80 they know to be in the data has high blood pressure.

Inferential disclosure happens when the published data enables an adversary to determine an individual's attribute value more accurately with the published data than otherwise would have been possible. For example, the published data might show a high

correlation between age and some disease. An adversary could use this information to determine if someone suffers from this disease if their age is known to the adversary.

Closely related to the above privacy disclosures is **membership disclosure**, which describes the process of an adversary inferring whether an individual was used to train on a machine learning model, i.e. if the individual was present in the training data.

2.2 Data Anonymization

Many methods to mitigate privacy breaches by decreasing the disclosure risk have been introduced. In the context of data anonymization, the classification of attributes into unique and quasi identifiers is important. Unique identifiers are attributes that uniquely identify an individual, e.g. a social security number. Quasi identifiers are attributes that cannot themselves uniquely identify an individual but can become uniquely identifying when combined with other quasi identifiers. This could be attributes like name and address or name and date of birth. All remaining attributes that neither classify as unique nor quasi identifiers are called non-identifying attributes. These are attributes that cannot be used for re-identification. Besides this classification, attributes can also be categorized according to their sensitivity. While sensitive attributes contain confidential information that should be refrained from disclosing, e.g. income, medical records, or religion, non-sensitive attributes do not contain any confidential information, e.g. gender or race.

The first attempts at data anonymization included:

- **Generalization** [22]: Through generalization the level of detail in the data is reduced by replacing specific values with more general or less precise values. For example, the address can be replaced by the city an individual lives in.
- **Perturbation** [23]: Perturbation adds noise in different forms to the data. This can be done by adding randomly generated values to the original values.
- **Micro-Aggregation**: Individual data records are combined into groups or clusters. The aggregated data represents a summary or statistical information about the group rather than individual records. The aggregation can involve calculating summary statistics such as averages, counts, or percentages for a specific attribute across a group of individuals.
- **Suppression** [24]: This is the simplest form of anonymization, done by removing unique or quasi identifiers. If the original data contains attributes like name, address or phone number, these would be removed entirely from the data.

However, the above listed methods do not guarantee privacy. The concept of k-anonymity, first introduced by [25] and formally defined by [26], uses generalization, suppression and micro-aggregation to provide a higher degree of privacy by mitigating the risk of identity

disclosure. K-anonymity ensures that for any value combination of quasi identifiers, there are at least k records sharing the same values, making a subject anonymous in the specific group. Individuals sharing the same values for their quasi identifiers are often referred to as equivalence class or k-group. K-anonymity prevents an individual from being identified, by being, for example, the only record in an equivalence class.

Notably, k-anonymity does not guarantee to prevent attribute disclosure. Even if a data set is k-anonymous, a sensitive attribute's value can be inferred without re-identification: a sensitive value can be inferred correctly if all individuals belonging to the same equivalence class share the same value for this attribute. To counteract this problem, l-diversity [27] was introduced. It ensures that in each equivalence class, there are at least l different records for every sensitive attribute. Still, this does not necessarily ensure a high degree of privacy. For example, a 2-diverse data set, with an equivalence class consisting of 100 individuals where 99 people share the same value for the sensitive attribute will still pose a threat, as adversary can infer their target records sensitive value with a high likelihood. Another problem can occur if the sensitive attributes in the equivalence class are different, but semantically similar. This could be an attribute containing individuals' illnesses. Even though the adversary cannot infer which illness the target record has, they can still conclude that the target record suffers from some type of illness. This shows that diversity within an equivalence class is not enough, the sensitive attribute's distribution within a class needs to approximately represent the attribute's overall distribution. This is addressed with t-closeness [28]. For data to achieve t-closeness, the distance between the distribution of a sensitive attribute within an equivalence class and the entire data cannot be larger than t. This way, an attacker cannot infer a sensitive attribute more easily through the equivalence classes, than with the entire data set.

However, researchers across the board have shown that the above-mentioned simple anonymization techniques are not as successful as initially hoped, as utility suffers and the achieved privacy might not be sufficient. The authors of [29] shed light on the ineffectiveness of simple techniques like suppression, generalization, and perturbation without guaranteed k-anonymity, l-diversity, or t-closeness. They highlight how, with an abundance of information available, re-identification of supposedly anonymous data is often possible. Especially because there is no guarantee of what kind of data will be released in the future, the risk for identity disclosure remains with sanitized and perturbed data. The authors of [6] show that the privacy-utility trade-off for l-diversity and t-closeness behaves similarly. Another study empirically showed that the utility of 2-diversity can be worse than 1000-anonymity [7]. Additionally, the utility of t-closeness with $t = 0.4$ is about as bad as 2-diversity. Any data guaranteeing t-closeness of $t \leq 0.3$ obtained even worse utility results.

Another approach to data anonymization is differential privacy (DP) [30]. Differential privacy ensures that the presence or absence of an individual's data in a data set does neither significantly impact the results of the data's analysis, nor comprise the individual's privacy. A data set is said to be differentially private if it is not possible to determine whether an individual is present in the data set based on any analysis or query output.

This means that for any two data sets that only differ by one individual, the probability distribution of the output should be very similar. To achieve DP, randomness is introduced via algorithms that add controlled random noise or perturbations to the data. To quantify the level of privacy, the parameter ε is used. Small values for ε indicate a higher level of privacy, which, however, results in noisier data with lower utility.

Additionally to the privacy-preserving approaches explained above, synthetic data has been introduced as a privacy-preserving data publishing method; it has been shown to preserve data utility to a high degree [19, 31, 32]. Synthetic data is artificially created data that closely resembles real data while protecting the privacy of individuals in the original dataset. It is less susceptible to identity [33] and attribute disclosure, since fully synthetic records usually do not correspond to one specific individual of the original data set unless the synthesis model is overfit. In that case, the synthesizer could produce a synthetic record that matches a real individual perfectly, and hence pose a privacy threat for that specific individual. To protect individuals from inferential disclosure, [34] introduced a synthesizer that generates data lacking correlations between sensitive and non-sensitive attributes, making models that try to infer the data less reliable. These aspects make synthetic data a promising practice for privacy-preserving data publishing.

2.3 Synthetic data

Synthetic data as a measure for disclosure control was first introduced by [35] and [36]. Both ideas build on multiple imputation [37], a probabilistic approach originally used to impute missing values. Instead of generating values for missing data, they used multiple imputation to replace either all or only sensitive attributes in the data. While these initial concepts only focused on numeric data, [38] introduced the first method for synthetization of categorical attributes, using log-linear models. Other approaches using tree-based models for categorical attribute synthetization followed [39]. In order to keep utility even higher, approaches for generating partially synthetic data, where only some attributes (sensitive values or key identifiers) [36, 40, 39], or some records [41] are replaced with synthetic data, were introduced. Studies have shown that although partially synthetic data has a utility advantage compared to fully synthetic data [42], the risk for identity disclosure remains [43, 44, 45] since some of the attributes remain unchanged.

Generally, data synthesis models (also referred to as data synthesizers) learn from the real training data and aim to generate artificial data carrying similar overall statistical properties, e.g. mean, variances, or univariate distributions, as the original data. This general idea is visualized in Figure 2.1. Additionally, synthetic data preserves the relations between data attributes, such as correlation and joint distributions. Ideally, for a well-trained synthesizer, the synthetic data it generates does not yield any, or only minor utility loss [19, 31, 32, 46, 47].

There are various approaches to generating synthetic tabular data. It can be generated using generative models, e.g. Generative Adversarial Networks (GAN) [48], Variational Autoencoders (VAE) [49] or by using probabilistic approaches like Gaussian Copula

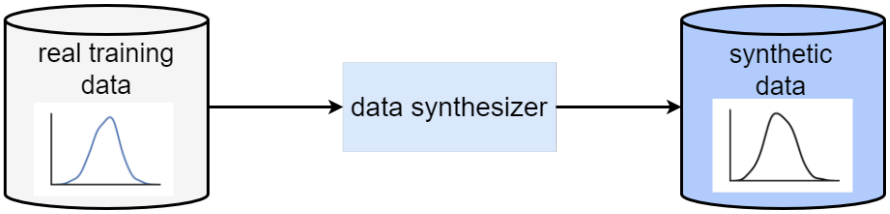

Figure 2.1: Basic Architecture of a data synthesizer: the synthesizer uses real data to train, and produces synthetic records with similar overall properties.

[50] and Bayesian Networks [51]. Another approach makes use of a tree-based method, generating data by using decision trees [52].

In the following, we describe the most important methods to generate synthetic data.

Bayesian Networks are graphical models that describe joint probability distributions. An example of such a network can be seen in Figure 2.2. A Bayesian network models conditional dependencies as a directed acyclic graph (DAGs). In a Bayesian Network, nodes represent attributes, while edges between nodes indicate conditional dependence. The Markov property is a fundamental concept used in Bayesian networks. It states that a node in a Bayesian Network is conditionally independent of its non-descendants, given its parents. In other words, once we know the values of a node's parents, the values of its non-descendant nodes provide no additional information about the node. This property is especially useful, as it allows simplifying computations in Bayesian Networks. Instead of considering all possible combinations of variables, only the parents of a node need to be considered to determine its conditional probability distribution. By exploiting this conditional independence, complex dataset structures can be efficiently modeled. Because of this, Bayesian Networks are often used in machine learning, e.g. for classification tasks. A prediction is made based on the most probable class, given the values of the target attributes values.

Data can then be generated from a Bayesian Network by first sampling values for the root node that does not have any parents, followed by each remaining node in the network based on its conditional probability distribution given the values of its parents [53]. This process is often referred to as importance or forward sampling.

The **Gaussian Copula** approach connects marginal distributions of individual random variables to their joint distribution. For every attribute, the univariate marginal distribution is modeled. A Gaussian Copula is then used to model the dependence structure between the attributes independently of the marginal distributions. The marginal distributions are transformed to be represented by a standard normal distribution, while the copula is represented by a $n \times n$ correlation matrix, where n is the number of attributes in the data set. This correlation matrix describes the pairwise correlations between attributes. Once the marginals and the correlation matrix are defined, the copula combines them using the multivariate Gaussian distribution to generate a joint distribution that captures the dependence structure. According to [54], samples can

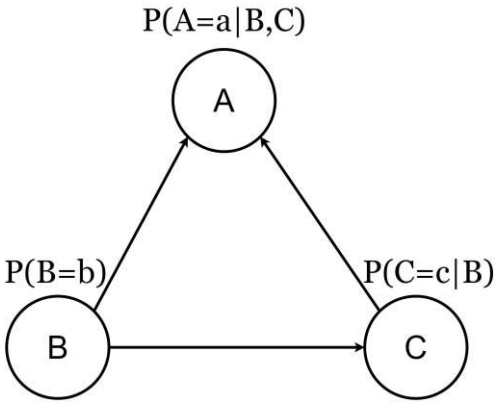

Figure 2.2: Example of a Bayesian Network: each attribute, represented by nodes, is conditionally independent of its non-decedents given its parents.

then be generated from this joint distribution by sampling from the standard normal marginals, multiplying them with the square root of the correlation matrix, obtained via a Cholesky decomposition. The Cholesky decomposition states that every positive definite matrix $A \in R^{nxn}$ can be factored as

$$A = LL^T, \tag{2.1}$$

where L is the lower triangular matrix with positive diagonal entries. The samples are then transformed using the inverse transformation, which maps the values back to their original distribution.

Generative adversarial networks (GANs) [55] consist of two main components: the generator and the discriminator (see Figure 2.3), both of which are neural networks. The generator generates synthetic samples. Both synthetic and real samples are fed into the discriminator, which is a classifier that tries to label the samples as fake data sampled from the generator, or real data. At the beginning of a GAN's training process, the generator produces random data that is clearly recognized as fake by the discriminator. As training continues, the generator is able to produce fake samples that seem more real, so the discriminator is unable to distinguish between real and fake records. During training, the discriminator and generator parameters are updated according to the two model's losses, or errors. The discriminator loss serves as a penalty for misclassified samples. If the discriminator loss is large, its parameters need to be updated, while the generator parameters need little to no update. The generator loss penalizes the generator for sampling unrealistic instances. The generator parameters, i.e. weights, need to be updated for a large generator loss. Since training a GAN combines two neural networks, the training process will alternate between training the discriminator and the generator. Ideally, a GAN will stop training once the discriminator is unable to tell the difference between fake and real samples, i.e. at an accuracy of 50%. Training should stop once this point is reached, since the responses of the discriminator, and hence the signals sent

to the generator, become meaningless. Once the generator is trained, synthetic data can be sampled from it [55].

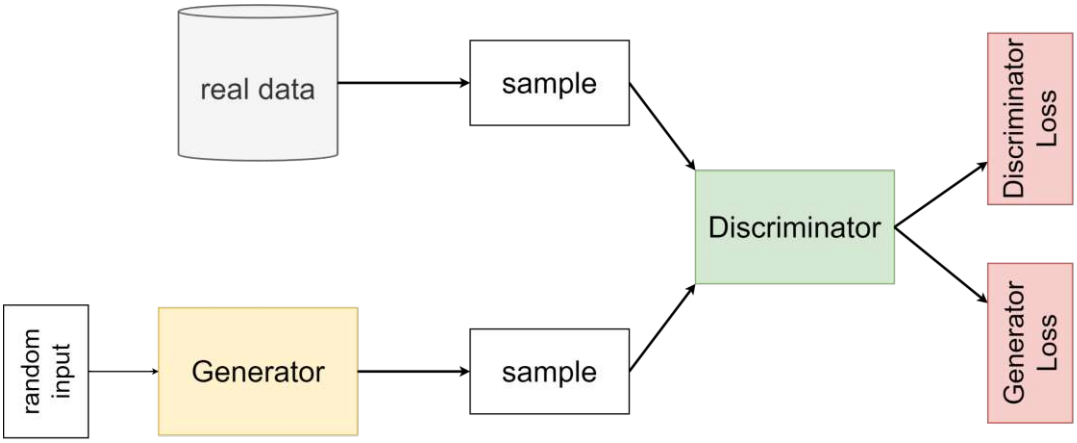

Figure 2.3: General Adversarial Network Architecture

As opposed to the original GAN architecture introduced by [55], conditional GANs are able to generate data records conditioned on a specific attribute of the data. The CTGAN by [56] uses this method with tabular data, to generate synthetic data with higher utility than synthetic data generated from an original GAN. One challenge for generating synthetic tabular data, especially for GANs, is dealing with mixed data types [56]. Tabular data will often contain discrete and continuous attributes. This problem does not exist for other data, like images or text. Furthermore, discrete attributes are often highly imbalanced, which makes training a synthesizer harder, and minority classes will not be able to be represented properly. To overcome these problems, [56] designed a conditional generator. In order to incorporate the conditional aspect into a GAN, the distribution is given by all samples conditioned on a discrete attribute having a specific class value. Formally, using the notation of [56]: for a synthetic sample $\hat{r}$, and k^* being a value from the discrete column D_{i^*}, where i^* describes the i^*-th discrete attribute in a data set, the distribution for $\hat{r}$ is given by: $\hat{r} \sim \mathbb{P}_G(row|D_{i^*} = k^*)$, with G being the synthesizer. This way, a CTGAN makes sure that all classes are represented fairly in the synthetic data it generates.

Autoencoders (AE) [57] comprise of two main components: an encoder and a decoder (see Figure 2.4). Input data x, with $x \in R^n$ is fed into the encoder, where it is encoded to a new feature representation in a lower dimension. This is called the encoded or latent space (R^m, where $m < n$). The decoder tries to reverse this process and bring the data into its initial space (R^n). The loss of this process is defined by the difference between the original input data x and the decoded data $\hat{x}$. This difference is often defined by the distance between an original and a decoded data point. For a given set of encoders and decoders, the goal is to find the encoder-decoder-pair with minimal loss, i.e. the pair that is able to retain the maximum information. Autoencoders are not regularized, meaning that there exist areas in the latent space that do not represent any of the input

data's points. These regions produce unreliable output samples $\hat{x}$. To account for this, Variational Autoencoders have been introduced.

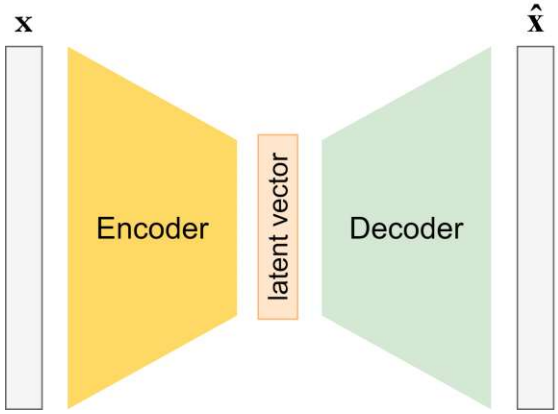

Figure 2.4: Autoencoder Architecture

Variational Autoencoders (VAE) [58] use regularization techniques during training to avoid overfitting and provide a regularized latent space. Compared to an AE, where *data samples* are encoded, the VAE encodes *distributions* (D). Hence, the VAE encoder outputs distribution parameters (θ) for each latent variable instead of a latent vector. With these parameters, the latent vector z is sampled (Figure 2.5). Oftentimes, the latent vector is sampled from a normal distribution with mean and standard deviation output from the encoder. The loss is then defined as the difference between the original input data x and the decoded data $\hat{x}$ plus the Kullback-Leibler divergence between the standard normal and the latent space distribution. Again, like for the AE, the optimal VAE consists of the encoder-decoder-pair with minimal loss. Synthetic data can then be generated by sampling data points from the learned distribution of the latent space [58].

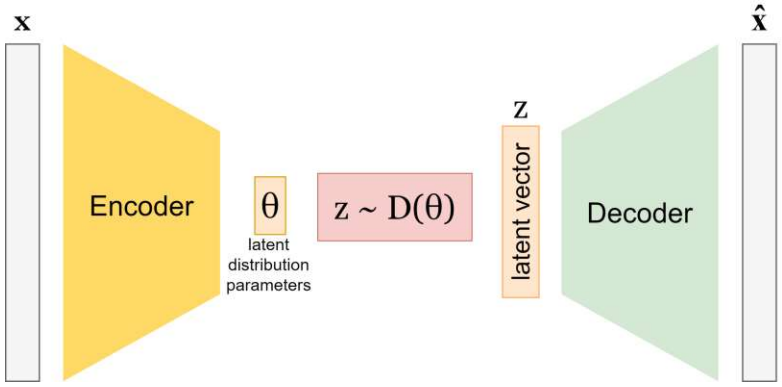

Figure 2.5: Variational Autoencoder Architecture

The Tabular Variational Autoencoder modifies the VAE for the use of tabular data by modifying the neural network to output a joint distribution of discrete and continuous

columns of a data set [56].

A non-generative model approach for synthesizing data is to use **decision trees**. Decision trees are supervised machine learning algorithms that can be used for both classification and regression tasks. They are tree-like structures composed of nodes and edges. Each internal node (also referred to as decision nodes) represents a decision based on an attribute, while each leaf node represents a class label or a predicted value. The decision nodes split the data into subsets based on different attribute values. This process is repeated recursively, creating child nodes until no further improvement in prediction accuracy is achieved or a stopping criterion is met. This criterion can be a maximum depth limit, a minimum number of samples required to split a node, or other conditions that prevent overfitting. To determine the most informative attribute for a decision node, splitting criteria like Gini impurity or information gain are used. These provide information about the impurity of each attribute. The next attribute to be split is the attribute obtaining the highest purity after a possible split. The most common class label (for classification) or the mean/median value (for regression) of the target values is assigned to each leaf node. [52] generate data by carrying out the following steps: given a data set with j attributes $x_0, ..., x_j$, starting with the first attribute x_0, n samples are drawn from its univariate distribution. The second attribute, x_1, is generated by building a decision tree using only the real values of x_0 as input ($x_1 \sim x_0$). For every value in the n samples drawn from x_0, this tree predicts the value for x_1. Each of the n samples now has two attributes: x_0 and x_1. These samples can now be used to generate the remaining attributes: for every attribute x_i, $i = 2, ..., j$, a decision tree is learned only from the real values of the preceding attributes $x_i \sim x_0, ..., x_{i-1}$. This tree is then used to predict the value for x_i from the preceding attributes. This step is repeated until the last attribute can be predicted from all of the preceding attributes from a tree that predicts $x_j \sim x_0, ..., x_{j-1}$.

2.4 Membership Inference Attacks

Membership inference attacks (MIA) [11] are attacks designed to perform membership disclosure. An adversary wants to learn if a target record was used to train a target model. For this, the adversary must have access to (at least some) information concerning the target record. With this information, the adversary can infer, with some certainty, whether the target individual was present in the training data. This can pose a threat to any individual's privacy included in the data set, especially if there is a certain characteristic that applies to every individual in the data set. Some examples would be medical data sets, containing only records of patients with a certain disease, a data set listing information about high-earning employees, or a store's data set containing all of its customers. An adversary could then, through a MIA, infer that a target subject suffers from a disease, has a high-paying job, or shops at some store. In general, membership inference attacks rely on the fact that a model overfits to its training data, and records used during the training will yield different response patterns than records that are new to the model [59, 60]. When evaluating this kind of attack, previous work on membership

inference attacks lacks a discussion of the semantic meaning of (mis-)classifying members and non-members. Naturally, a member that is labeled as a non-member will not face any negative consequences. An adversary will simply conclude that this person was not included in the training data even though they were a member of it. Arguably, a member being labeled as such is a higher privacy breach than a non-member being classified correctly. Again, being labeled as a non-member will not violate anyone's privacy. But what about non-members who are incorrectly labeled as members? This can lead to an adversary inferring incorrect and possibly harmful information about an individual. This problem, however, is inevitable. Even if an adversary is able to infer membership with high accuracy, they can most likely never be 100 percent certain their prediction is correct.

The majority of research has focused on membership inference attacks on classification (or generally supervised) models. There are several approaches on how to perform such an attack. An attack can either be supervised or unsupervised. Supervised attacks use a binary classifier as *attack model* that will then label input records as members or non-members. [11], for example, uses so-called shadow models that try to mimic the behavior of a target model. The data these shadow models are trained on, referred to as *shadow data*, and membership for each record are known to the attacker – unlike the inputs and associated labels that were used to train the actual target model. An attack model can classify members and non-members of the training data set based on learning the patterns from the prediction probability output by the shadow models. The idea is that the shadow models will achieve higher prediction scores for records they have been trained on, than for records that are new to the models – and that this learned knowledge transfers to the target model.

Unsupervised attack methods use the prediction metrics output from the target model. A simple unsupervised attack, based on prediction correctness [61, 62], labels a record as a member if the target model predicts it correctly. It builds on the assumption that the target model is fit to the training data and does not generalize to new data. Another approach uses prediction loss [59]: an adversary labels a record as a member if the prediction loss is smaller than some predefined threshold τ. This approach is based on the assumption that the target model is trained to minimize the prediction loss, and members of the training set should therefore have a smaller loss than non-members. The authors argue that the average training loss is often published with their respective architecture, and can be used as threshold τ. Similarly, an attack based on prediction confidence [59] assumes that this prediction confidence is higher for training than for test records. Records with a prediction confidence greater than a preset threshold will thus be labeled as members. The entropy-based attack [11, 63] assumes that the prediction entropy of the training samples is smaller than the one of the test samples. Again, a threshold τ is chosen, and only records having a prediction entropy lower than τ will be labeled as members. A modified entropy-based approach [63] combines a sample's entropy with its ground truth. The reasoning behind this is that a false positive prediction with a high confidence score will have a low entropy. According to the entropy-based attack, this

record would then be labeled as a member. However, it can be assumed that this record was not a member, due to it being falsely labeled. The modified entropy-based attack then labels a record as a member if its modified entropy is smaller than a threshold τ.

Membership Inference attacks on synthesizers are a relatively new concern compared to attacks on supervised machine learning models. The basic scheme of such attacks on synthesizers can be seen in Figure 2.6.

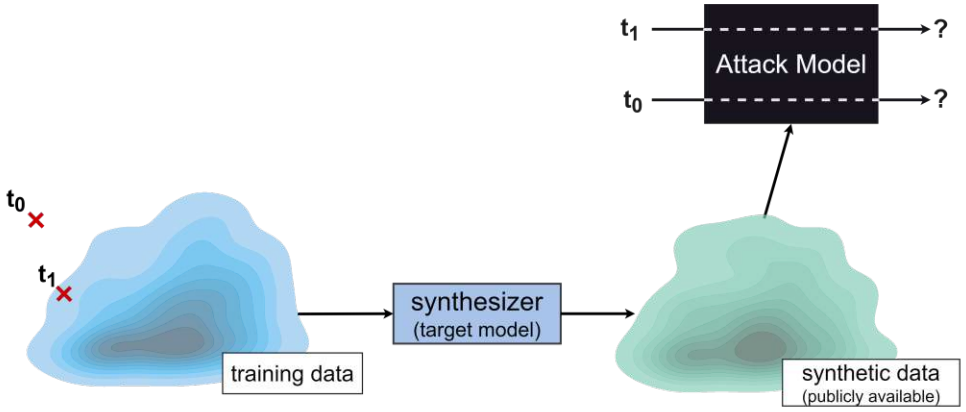

Figure 2.6: Membership Inference Attack on a Synthesizer: the data owner uses their data to train the synthesizer. The synthetic data generated from it is made publicly available. An adversary attacks the synthesizer to find out whether an individual was used to train the synthesizer.

Synthesizing models, unlike classification or regression models, lack prediction responses from the target model, like prediction confidence, correctness, etc., which are often used to train attack models against classifiers. However, some membership inference attacks on synthesis models can still be derived from the approaches designed for supervised ML tasks.

2.4.1 Supervised Attack Approaches

In the work of Stadler et al. [8], the authors rely on shadow modeling, a supervised approach introduced by [11], where a synthesizer is learned from a reference data set, coming from the same population as the original training data[1]. This reference set is used to train a synthesizer and generate several data sets from it. This process is done twice, once including and once excluding a target record in the reference data set. The generated data sets are then labeled zero or one, depending on whether the target record was used to train the synthesizer or not. Each data set is then flattened to be represented by a one-dimensional vector. This way, a classifier can learn and predict the membership

[1]Population in this context describes the entirety of individuals from which the samples in a data come from. All students from one school, for example, are the population of students from that group. Ten randomly chosen students of this population would then be a sample

of the target record. A visualization of the attack can be seen in Figure 3.1. To assess the privacy risk, the authors of [8] define the so-called privacy gain (PG), which measures the privacy gain when publishing synthetic instead of the original data. A record with $PG = 0$ can be inferred correctly by the adversary and does not gain any privacy through synthetic data publishing. A record having $PG = 1$ on the other hand, supposedly protects the records against the attack. The results published in the empirical study show PG values for **ten selected records** in total. The majority of these records show PG values of around 1. Only some outliers have a privacy gain of about 0.5. One outlier shows $PG = 1$.

The authors of [16] use the approach proposed by Stadler et al. [8] and show that there is no overall privacy risk for all data records by selecting ten random records. Their attack model was not able to distinguish between members and non-members on these randomly selected records. Combined with the results found by Stadler et al. [8], we hence conclude that the shadow model attack is not a serious privacy threat for the entirety of records contained in the original data set, but rather only for single, more vulnerable records. In [16], "vulnerable records" are defined as records with a larger distance to their neighbors, i.e. records that are not similar to other records. To identify these records, they define a vulnerability score V_k, that is, in comparison to our proposed risk scores as defined in Section 3.4.6, computed using the original data only. Using the notation from [16], for any dataset D, a record $x_i \in D$, and a distance d, the vulnerability score V_k is defined as:

$$V_k(x_i) = \frac{1}{k} \sum_{j=1}^{k} d(x_i, x_{i_j}) \tag{2.2}$$

where $x_{i_1}, ..., x_{i_{|D|-1}}$ are the records ordered according to their distance to x_i. The distance d is computed by first splitting the attributes into subsets of categorical (F_{cat}) and continuous (F_{cont}) attributes. Every categorical attribute $x_{i,f}$, $f \in F_{cat}$ in D is then converted into a one-hot-encoded vector $h(x_{i,f})$.

After getting one one-hot-encoded vector for each categorical attribute, they are all concatenated into one single vector.

The continuous attributes $x_{i,f}$, $f \in F_{cont}$ are scaled using min-max-scaling:

$$n(x_{i,f}) := \frac{x_{i,f} - \min_{j=1,...,|D|}(x_{j,f})}{\max_{j=1,...,|D|}(x_{j,f}) - \min_{j=1,...,|D|}(x_{j,f})} \tag{2.3}$$

Again, the attributes are concatenated into a single vector:

$$c(x_i) := (n(x_{i,f}))_{f \in F_{cont}} \tag{2.4}$$

As presented in [16], the distance between two records x_i and x_j is then given by:

$$d(x_i, x_j) := 1 - \frac{|F_{cat}|}{F} \frac{h(x_i) \cdot h(x_j)}{||h(x_i)||_2 * ||h(x_j)||_2} - \frac{|F_{cont}|}{F} \frac{c(x_i) \cdot c(x_j)}{||c(xi)||_2 * ||c(xj)||_2}. \tag{2.5}$$

Here, $*$ denotes scalar multiplications, $\cdot$ denotes the dot product between two vectors, and $||y||_2$ denotes the Euclidean norm of a vector y. The values for $d(x_i, x_j)$ can range between 0 and 1. Records that are less similar to other records will have larger values, and only identical records will have a distance of zero. The authors show that this vulnerability score effectively identifies records at risk. Their experimental set-up, however, only considers a subset of ten vulnerable records to show the MIA performance.

2.4.2 Unsupervised Attack Approaches

The idea of using a distance approach to infer membership was originally introduced by [13] for MIA on image data, and later used by [9] on synthetic tabular data. In this unsupervised approach, it is assumed that the distance from a target record to its closest synthetic data record is smaller for members of the training set than it is for non-members. This way, target records will be classified as members, if the distance to the closest synthetic record is smaller than some threshold ε. Different values for ε are used to compute the receiver operating characteristic - area under the curve (ROC-AUC) value, and to compare the attack performance. The authors found that if the adversary has access to the synthesizer and can generate synthetic data from it, this approach works well to infer membership on target records.

2.4.3 Attack Settings

In general, membership inference attacks (against discriminative and generative models) can be classified according to the adversary's knowledge: we distinguish white- and black-, and no-box attacks as defined in [64]. For the white-box attack, the adversary has access to the target model. This means that they can feed their own data into the model and collect its responses. These responses can be prediction properties when supervised models are attacked, or synthetic data, for attacks against synthesizers, similar to the work in [65]. Furthermore, information on the architecture of the model and its learned parameters, as well as how the target model is trained, can be accessed. The black-box-setting on the other hand assumes that an adversary only has access to the model in- and output, i.e. input data and model response. If the target model is a classifier, this means that it can be queried by feeding a record into the model and collecting its response, e.g. prediction. Sometimes the type of model used, e.g. which classifier, is also known. The no-box attack is the most restrictive attack setting. It assumes that an adversary does not have access to the model, but rather only to its output. For attacks against classifiers, this output can be a labeled data set, or a data set containing the prediction confidence for each label. For synthesizers, the adversary would only have access to the synthetic data set. Previous research suggests that white-box attacks are more successful than black-box attacks for generative models [9, 13]. For attacks against classification models, however, both scenarios seem to be equally effective [66].

2.5 Outlier Detection

Outliers are data points that significantly deviate from the majority of the data in a dataset. They are often unusual, rare, or extreme observations that don't conform to the typical patterns or distribution of the data. They are often referred to as anomalies or extreme values. [67] defines outliers as "an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism". Outliers can be observed in various types of data, including numerical, categorical, and multivariate data, and they can arise for different reasons:

- **Natural Variation:** Outliers occur naturally due to the inherent variability in a dataset. These outliers represent genuine, albeit rare, observations that are an essential part of the data distribution.
- **Measurement Errors:** Outliers can result from errors during data collection or measurement. These errors can include sensor malfunctions or human mistakes when collecting data. Such outliers are often considered as noise and may need to be corrected or removed.
- **Data Entry Errors:** Human errors in data entry can introduce outliers into a dataset.
- **Novel Events:** Outliers may indicate significant, unexpected, new events or changes in the data-generating process.

Outliers can occur in either one or more variables [68]. We distinguish the following:

- **Univariate Outliers:** Outliers in a single variable.
- **Multivariate Outliers:** Outliers that occur in multiple variables simultaneously.

Furthermore, outliers can be grouped with respect to their surrounding data points [69]:

- **Global Outliers:** Outliers that are unusual across the entire dataset.
- **Local Outliers:** Outliers that are unusual within a specific subset or cluster of data.

Another classification for outliers [70] (or anomalies) distinguishes between the following:

- **Point Anomaly:** A single data point behaves differently than the rest of the data.
- **Collective Anomaly:** A group of data points shows deviating patterns from the rest of the data.

- **Contextual Anomaly:** A data point that is anomalous only in a specific context. For example, while a high level of traffic might be normal at 5 p.m., high levels of traffic at 3 a.m. might be considered anomalous. Point or collective anomalies can simultaneously be contextual anomalies.

Outlier detection describes the process of detecting outlying data points within a population. Various statistical, graphical, and machine-learning techniques are used to detect outliers.

Outlier detection methods can be grouped into supervised, semi-supervised, and unsupervised mechanisms, depending on the availability of labels [71]. Supervised methods require pre-labeled data, classified as out- or inlier. This makes the outlier detection a supervised classification task. Unsupervised outlier detection lacks these labels. This is essentially analogous to an unsupervised clustering problem, where data is grouped into out- and inliers based on the unlabeled data. For semi-supervised detection, only class labels for the inlying records are available. A model trains only on these inlying records and aims to recognize data that differs from the inliers as outliers.

Some outlier detection methods only look at each attribute individually. One method defines outliers for continuous attributes to be records that are three inter-quartile ranges below the first quartile or above the third quantile [72], where the inter-quartile range is the distance from the first to the third quartile. A classical approach for normally distributed data uses standard deviation (SD). All points outside the $\bar{x} \pm k\ SD$, where $\bar{x}$ is the mean and k a preset value for the number of standard deviations, are considered outliers. Another widely used method for normally distributed data is the Z-score. It is defined as

$$Z_i = \frac{x_i - \bar{x}}{SD}, x_i \sim N(\mu, \sigma^2). \tag{2.6}$$

Now, Z follows a standard normal distribution and all values larger than three would be considered outliers. Alternatively, uni- and bivariate outliers can easily be detected using a graphical approach: Plotting each data point will allow detection of outlying points visually.

Other algorithms, like Local Outlier Factor [17] and Isolation Forest [18], consider all data attributes to divide the population into two groups: outliers and inliers.

[17] defines the "Local Outlier Factor" (LOF) to be the degree of a record being an outlier. The degree depends on how isolated a record is in regard to its neighborhood. The local density is detected by finding the mean distance from a data point to its k-nearest neighbours. Points that have a higher local density than their neighbours are assumed to be part of a cluster. If a point's local density is lower than its neighboring point's density, it is considered an outlier.

Isolation Forests [18] use tree-based models to isolate data points. The idea is that when building a binary tree and isolating every data point, i.e. building the tree until every leaf node contains a single data point, outliers will be closer to the root than regular

points. These trees are called isolation trees. Several isolation trees are used to form the isolation forest. Outliers are assumed to have a short average path from the root to the isolated point. Figure 2.7 visualizes this idea.

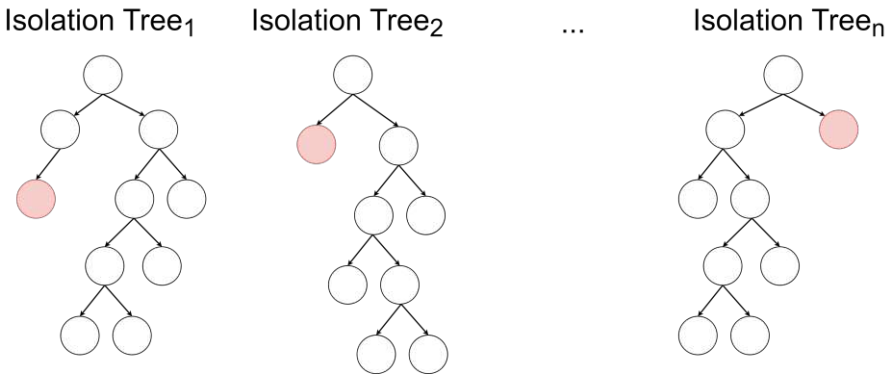

Figure 2.7: Isolation Forest: each isolation tree finds the distance from each point (leaf node) to the tree root. This distance is on average smaller for outliers (red nodes). The average distance is calculated over all n isolation trees that make up the isolation forest.

2.6 Summary

In this chapter, we summarized different attempts at data anonymization and described their (dis)advantages. Furthermore, we described the concept of synthetic data as a method for privacy-preserving data publishing and explained the synthesis models Bayesian Networks, Gaussian Copula, Generative Adversarial Networks, Autoencoders, Variational Autoencoders, and decision trees in detail. We gave an introduction to membership inference attacks and presented attack approaches on classification models. We then highlighted the differences for MIA on synthesis models as opposed to classifiers and explained two recently proposed approaches. Lastly, we gave an overview of outliers, and how they can be classified and detected. Additionally, we illustrated the two main outlier detection algorithms used in this thesis: Local Outlier Factor and Isolation Forest.

CHAPTER 3

Experiment Design

In this chapter, we present our approach and experiment design. We list and explain the steps necessary to be able to answer our research questions.

We choose to evaluate the MIA and risk identification empirically on publicly available data. For conducting the experiments and gathering insights, we will follow the Cross Industry Standard Process for Data Mining (CRISP-DM) [73]. The following sections list the six phases of the CRISP-DM methodology and describe how they are applied in the thesis.

3.1 Business Understanding

The importance of data privacy, its current shortcomings, and measures against it are discussed in Chapter 1. The goal of this thesis is to identify the most vulnerable records for MIA, and, as defense against the attack, remove them from the training data. The defense can be considered a success if i) the MIA performs worse overall, on records at risk and on records not at risk, and ii) the data utility does not suffer from the defense.

3.2 Data Understanding

For the experiments, exclusively data sets with an associated classification task are chosen. We conduct our experiments on four different data sets, varying in size and target class distributions. The exact dimensions and distributions are listed in Table 3.1. All data sets are obtained from the UCI Machine Learning repository. The Caesarian[1] data set contains information about 80 women with variables related to medical attributes, like heart problems or blood pressure, concerning pregnancy. The binary target attribute denotes if a woman gave birth via a Caesarian section.

[1] https://archive.ics.uci.edu/ml/datasets/Caesarian+Section+Classification+Dataset

The heart disease[2] data set contains data of 303 patients and 14 health-related variables with the target being a binary attribute denoting whether a patient was diagnosed with a heart disease or not. The breast cancer Wisconsin[3] data set has 699 rows and describes patients with either benign or malignant breast tumors, which represents the moderately imbalanced target variable. The fourth data set used in our experiments, the thyroid[4] data set, also contains patient and health-related data. In contrast to the other data we use, this is a highly imbalanced data set. The target attribute consists of three groups, two of which express a thyroid dysfunction. These two classes make up five and two percent of the population respectively. We specifically chose the thyroid data to investigate the effects of removing outliers on the data utility, especially for the minority classes.

Table 3.1: Dataset characteristics

Dataset	# Instances	# Attributes	target variable split
Caesarian	80	6	58/42%
Heart	303	14	54/45%
Breast Cancer	699	10	65/35%
Thyroid	3428	22	93/5/2%

3.3 Data Preparation

Besides data cleaning, this step includes transforming the data according to the input data needed for the different attack approaches. Furthermore, outliers will be detected during this phase.

The only data set with missing values is the Breast Cancer data, with 16 rows containing missing values. As the number of rows containing missing values is relatively small, we delete these rows and end up with 683 instances.

We split each of the four data sets given in Table 3.1 into two equally sized data sets. We use one of these data sets as real training data and the other as a reference data set as done in [15]. The authors of [11] propose that the reference data could also be generated from the target model directly. For this, the assumption that the target model is available to the attacker has to hold. This approach has recently been studied on synthesizers [65]. Additionally, drawing the reference data from each attribute's marginal distribution, if they are known, was proposed [11]. Another approach samples the reference and real training data from the entire data set [8]. This will then allow for the reference and real data to overlap. The detailed description of the reference data set, and how it is used is given in the respective approach descriptions of the shadow modeling (Section 3.4.4) and distance-based approach (Section 3.4.5).

[2] http://archive.ics.uci.edu/dataset/45/heart+disease
[3] http://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original
[4] http://archive.ics.uci.edu/dataset/102/thyroid+disease

3.3.1 Outlier Detection

For our experiments, we test the relationship between records at risk and outliers. To detect outliers, we use the algorithms Isolation Forest and Local Outlier Factor, which we described in Section 2.5. For the LOF, we use $k = \{10, 15, 20\}$, where k is the number of nearest neighbors to calculate the mean distance from. For the iForest, we use $n\_estimators = 100$ for the number of trees that make up the forest. Furthermore, to compare our results with the ones found in [8], we further use the outlier definition proposed in their paper where outliers are defined as "records that either have rare categorical attribute values or numerical values outside the attribute's 95% quantile". As the authors of [8] do not further define how they identify "rare categorical attributes", we define them as those attribute values that occur in only five or less percent of records. In our analysis, we refer to this outlier detection approach as *Quantile approach*.

3.4 Modeling

During this phase, we will implement the membership inference attack models. Two different approaches for Membership Inference Attacks, motivated by [8] as well as [9], are implemented. To generate synthetic data, we use four different synthesis models.

3.4.1 Synthesis Models

The following synthesizers are used for our experiments:

- Bayesian Networks[5] (DataSynthesizer)
- Gaussian Copula[6] (SDV)
- CT-GAN[7] (SDV)
- TVAE[8] (SDV)

We specifically use these synthesizers as they are widely used, effective, and implementations are readily available. We decide to use these four synthesizers specifically, as they are commonly used in related work [9, 8]. By using different synthesis models, we can i) explore the overall MIA success across synthesizers and ii) compare the data utility between synthetic data generated by the synthesizers.

The python package "DataSynthesizer" [74] uses Bayesian Networks [75] in combination with a greedy algorithm introduced by [76] for building these networks. For our analysis, we try different values for the parameter p, i.e. the maximum number of parents for each node. We expect less utility loss, but less private data for networks with high values for p.

[5] https://github.com/DataResponsibly/DataSynthesizer
[6] https://docs.sdv.dev/sdv/single-table-data/modeling/synthesizers/gaussiancopulasynthesizer
[7] https://docs.sdv.dev/sdv/single-table-data/modeling/synthesizers/ctgansynthesizer
[8] https://docs.sdv.dev/sdv/single-table-data/modeling/synthesizers/tvaesynthesizer

The Synthetic Data Vault[9] (SDV) published several data synthesizers generating synthetic data, including Gaussian copulas [50], Conditional Tabular Generative Adversarial Networks (CTGAN) [56, 77], which are based on Generative Adversarial Networks (GAN) [55] and Tabular Variational Auto-Encoders (TVAE) [56] based on standard Variational Auto-Encoder(VAE) [58].

3.4.2 Implementation of Membership Inference Attacks (MIA)

For this thesis, two different MIA approaches on synthesis models will be implemented. One of them, the shadow model approach from [8], is a supervised attack, while the distance-based approach from [9] is an unsupervised attack. For each record in the training data set, we define a risk score, which gives information on the likelihood of the record's membership being inferred correctly. A record that is always labeled correctly receives a high risk score. Records that are rarely or never inferred correctly have a low risk score. Therefore, the risk of a record is determined by executing an attack. The mathematical definition of our risk score depends on the attack approach and is given in Section 3.4.6.

3.4.3 Threat Model

Whenever a MIA is simulated, certain assumptions about the adversary's and defender's motivation, goals, and knowledge need to be made. The threat model provides these assumptions, which can then be used to implement the attack and defense strategies. In our experimental setup, we define the adversary's and defender's profiles similar to the categorization of Biggio and Rolli [78], as follows:

- **Adversary's motivation:** An adversary wants to find out if a target record is a member of a training data set. With this knowledge, they can draw conclusions about the target individual.
- **Defender's motivation:** In the scenario of MIA, the defender plays the role of the data holder, who wants to publish the data they collected without neglecting the privacy of individuals present in the data. As collecting data is often costly and time-consuming, sharing it benefits others who might want to acquire the data for their own use.
- **Adversary's goal:** The adversary strives to build a model that is able to correctly classify members and non-members.
- **Defender's goal:** The defender aims to publish data with a high utility and privacy guarantee. Their goal is to successfully implement a strategy that both preserves privacy and maintains data utility.

[9] https://docs.sdv.dev/sdv/

- **Adversary's knowledge:** For our experiments, we imitate a "no-box" attack and evaluate the risk of synthetic data sharing, meaning an adversary only has access to the synthetic data generated. We do however assume that the adversary has some additional prior knowledge about the environment that the synthetic data was created in, similar to [8, 11]. First, we assume the adversary to have knowledge about the type of model used to generate the synthetic data and if applicable, its parameter settings. Second, the adversary is assumed to know the size of the original training and the synthetic data set. This knowledge might benefit the adversary, as they can then adapt the size of the reference data set – an attack model that was trained on 100 samples might not be able to capture the behavior of a target model that has been trained on 10.000 samples. Lastly, we assume that the adversary has access to a reference data set that stems from the same population as the original training data. This could for example be data containing information about the same individuals as the original data, collected in a different year. Or similar: data collected at the same time, but from people living in two different cities. Note, however, that the assumption that the reference data and the original training data do not overlap, holds.
- **Defender's knowledge:** The defender has access to the entire training data and is aware that if they publish the data, it might fall victim to ill-intentioned attacks.

3.4.4 Shadow Model Approach

Stadler et. al [8] use shadow modeling to infer memberships. The attack design can be seen in Figure 3.1. The adversary has access to the reference data set D_{ref}, coming from the same population P as the original training data set D_{raw}. In our setting, we randomly split each data set listed in Section 3.2 to use as original data (D_{raw}) and reference data (D_{ref}). Furthermore, we have a publicly available synthetic data set S, which was generated using D_{raw} as input. We assume that S and D_{raw} are the same size, which is known to the adversary. The size of D_{raw} is assumed to affect the data quality of S. If the size of S and D_{raw} is known to the adversary, they can use them for their generation process during the attack to better mimic the behavior of the target model.

To evaluate the influence of a single data record on the synthesizer and the resulting synthetic data, we carry out the following steps:

1. Select the target record $t^* \in P$.
2. Two synthesizers M are trained. One on D_{ref} including t^* (M^+) and one excluding t^* (M^-).
3. Multiple data sets that are the same size as D_{raw} are generated by M^+ and labeled 1, and by M^- and labeled 0, respectively.
4. All generated data sets are flattened (as detailed in section 3.4.4) to be represented by a one-dimensional array and merged into one data set, the attack training set.

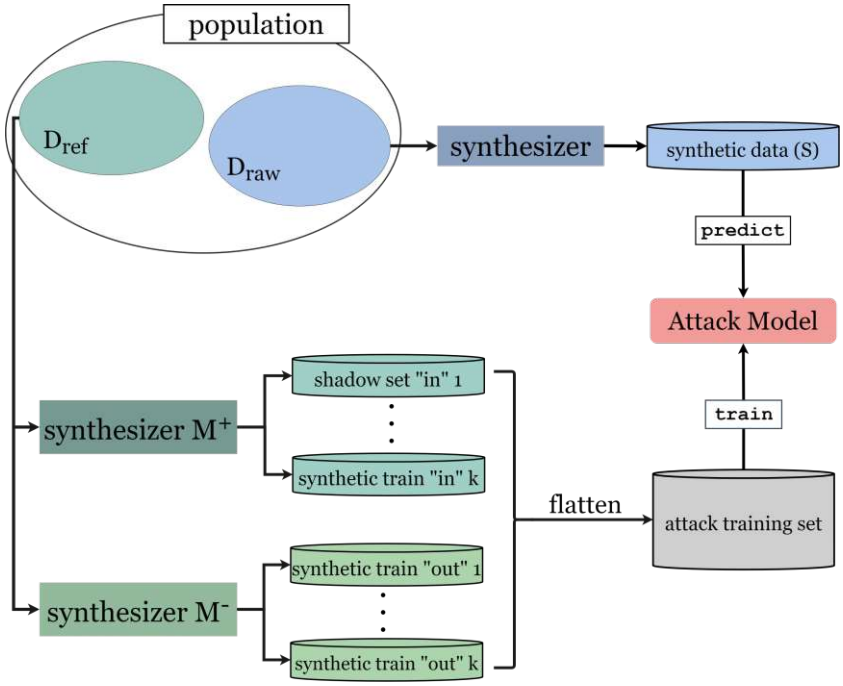

Figure 3.1: Shadow Model Approach

5. A binary classifier is used as attack model A, which has the task to label each synthetic data set in the attack training set according to the membership of t^*.

6. The public synthetic data S gets flattened in the same way as the shadow data sets.

7. The attack model A can now predict the membership of t^* in D_{raw}, and returns 1 if the model concludes that t^* was in D_{raw}, or 0 otherwise.

8. A risk score can be obtained by using the prediction confidence of the attack model.

The authors of [8] play out the attack on five outliers and five randomly chosen records only. For our analysis, each record is selected to be the target record once. This way records at risk can be identified. Additionally, we use five different classifiers acting as attack model A: Random Forest, Naive Bayes, Logistic Regression, Support Vector Machine, and k-Nearest-Neighbor classifier, where we use the scikit-learn implementations[10]. Earlier work used Random Forest, Logistic Regression and k-Nearest-Neighbor [8], as well as Neural Networks [11, 79]. The final membership prediction will be made according to the mean confidence score of all five classifiers: If the mean confidence is greater than 0.5, the membership prediction will be 1, and 0 otherwise.

[10] https://scikit-learn.org/stable/supervised_learning.html

Flattening Methods

As the attack model is a classifier that can only take a one-dimensional vector as input, the shadow data sets need to be transformed to be represented by such vectors. We use three different methods to do so:

- **Naive:** This method uses simple summary statistics: Numerical attributes are represented by their mean, median, and variance. For each categorical attribute we take the most and least frequent value, as well as the total number of classes per attribute, like it was done in [8].

- **Correlation:** For this, we dummy-encode categorical attributes as a pre-processing step. When dummy-encoding categorical attributes with j possible attribute values, the values are mapped to binary attributes. The j possible values are represented by $j - 1$ attributes. A record having value zero for all $j - 1$ dummy variables belongs to the category that is not represented by its own column. After dummy encoding, we compute the pairwise correlations.

- **Principal Components (PC):** This method computes the principle components and only uses the first PC to represent the data set. Principal components are a way of reducing data's dimensionality, by transforming a large set of attributes into a lower-dimensional space, all while preserving as much information as possible. The principal components are linear combinations of the original attributes in the data, and orthogonal to each other. The first principal component captures the most variance in the data, the second component explains the second most variance, etc. To find the principal components, the data has to be standardized, i.e. a mean of zero and a standard variance of one. Then the covariance matrix for all attributes in the data is computed. The eigenvalues and eigenvectors for the covariance matrix are calculated. The eigenvector corresponding to the highest eigenvalue is then the first principal component. This is the vector used to represent the entire data set in our analysis.

3.4.5 Distance-Based Approach

[13] originally introduced a distance-based approach for membership inference attacks on generative adversarial networks for synthetic image data. Later, [9] used this approach to conduct attacks on tabular data. While Hyeong et al. conduct the attacks on GANs and VAEs exclusively, we consider GANs and VAEs as well as Bayesian Networks and Copulas for the synthesizing process. By including additional synthesis models in our analysis, we can show differences in MIA performance that might occur when generating the data with different models. The basic idea is based on the assumption that the distance from a target record to its closest synthetic data record is smaller for members of the training set than for non-members (Figure 3.2). This way, target records will be labeled as members if the distance to the closest synthetic record is smaller than some

threshold ε. However, the method presented in [9] does not present a way to find a suitable value for ε, but rather provides the ROC-AUC value for varying ε values.

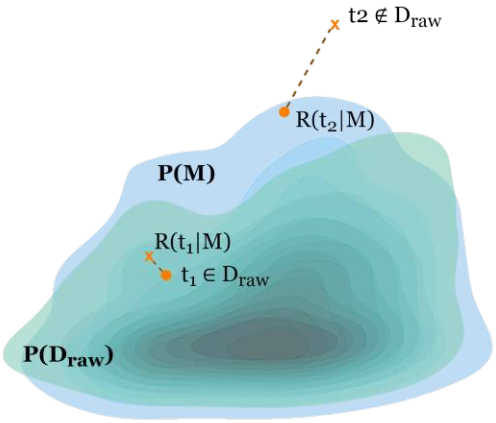

Figure 3.2: Distance-based Approach Idea: the approach builds on the idea that the distance from a record t_i, with $i = \{1, 2\}$ to their closest synthetic data record $R(t_i|M)$ generated by synthesizer M is smaller for records that were used to train M.

Adapting the notation used by [13] to our notation, an adversary wants to infer membership on a target record t^* by computing the probability $P(t* \in D_{raw}|t^*, M)$, where D_{raw} is the original training data and M is the target model, a synthesizer. The assumption is that this probability is proportional to the probability of a target model generating the target record $t*$. It is assumed that this assumption is valid since the synthesizer (M) is designed to approximate the distribution P of the training data D_{raw}, meaning $P_{D_{raw}} \approx P_M$. Formally we get

$$P(t^* \in D_{raw}|t^*, M) \propto P_M(t|M) \tag{3.1}$$

This probability can then be approximated using the Parzen window density estimation [80]. The Parzen Window estimates the density function of a continuous random variable. The probability density of any given point can be estimated by placing a window around that point and calculating the density within this window. We then get

$$P_M(t|M) \approx \frac{1}{k} \sum_{i=1}^{k} exp(-||t - M(z_i)||_2); \; z_i \sim P_z \tag{3.2}$$

with k being the number of synthetic records generated by M. The so-called reconstructed copy of record t is then given by

$$R(t|M) = \underset{\hat{t} \in \{M(.)_i\}_{i=1}^{k}}{\text{argmin}} \; ||t - \hat{t}||_2 \tag{3.3}$$

where $\{M(.)_i\}_{i=1}^{k}$ are all synthetic records collected from M. The adversary will then label the membership of the target record m_{t^*} as 1 (i.e. member) of D_{raw} if the distance

$L(t^*, R(t^*|M))$ between t^* and $R(t^*|M)$ is smaller than the threshold ε. Formally,

$$m_{(t^*)} = \mathbb{1}[L(t^*, R(t^*|M)) < \varepsilon] \tag{3.4}$$

The challenge for an adversary is finding a suitable value for ε, which can correctly label as many records as possible. Both [13] and [9] try different values for ε and evaluate the attack success. However, they disregard the process of an adversary finding this threshold and assume the adversary somehow knows a well-performing value for ε. The two empirical studies [13, 9] also lack any further testing on how the threshold ε generalizes to other data sets. We define a possible scenario in which an adversary can learn the threshold value beforehand, and use it to infer memberships later; this process is visualized in Figure 3.3. We assume an adversary to have access to a reference data set D_{ref}. We obtain this reference data set by doing the same process as for the Shadow Model approach described in Section 3.4.4. Both D_{raw} and D_{ref} are then split again into training and test sets. The training set of the original data ($D_{raw_{train}}$) will be used to train the synthesizer that generates a synthetic data set S that the data owner wishes to publish (membership=1). The testing data $D_{raw_{test}}$ with membership=0 is needed to evaluate the threshold's performance on D_{raw}. With $D_{ref_{train}}$ and $D_{ref_{test}}$ (membership = 1 and 0 resp.) an adversary can find a suitable threshold to distinguish between membership classes 1 and 0, which can be used to classify the memberships of records contained in D_{raw}. This is done by calculating the distances d for all data points $x_i \in D_{ref}$. These distances are then scaled to range between 0 and 1, using min-max scaling. The optimal distance threshold ε can then be found by looking at the true positive (TPR) and false positive rates (FPR) and choosing the threshold that maximizes $TPR - FPR$. Hence the optimal threshold ε^* can be defined as:

$$\varepsilon^* = \underset{\varepsilon}{\operatorname{argmax}} \ (TPR - FPR) \tag{3.5}$$

This threshold can then be used to classify the records in D_{raw}.

3.4.6 Risk Score

The method to identify records at risk is different, depending on which of the attack approaches that were described in Section 3.4.4 and Section 3.4.5 is used. For every record that was used in the training set, we will compute a risk score λ, which gives information on how much at risk for MIA each record is.

For the shadow model approach, the risk score for correctly labeled records is defined as the prediction confidence (PC) of the attack model A. The prediction confidence (or class probability) ranges between 0 and 1, and quantifies how confident the classifier's prediction is. Usually, a confidence of 0.5 or more for the positive class will be labeled positive by the classifier. If a prediction's confidence for the positive class is less than 0.5, it will be assigned a negative label by the classifier. We assume that records that can be predicted correctly with high confidence are somehow influential to the synthesis model and therefore an easy target for an adversary. For incorrectly labeled records, the

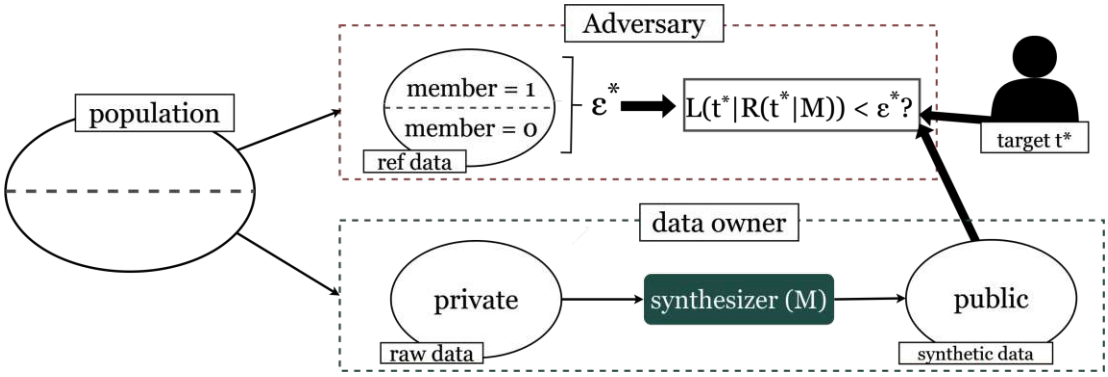

Figure 3.3: Distance-based Approach: the population, one of the data sets described in Section 3.2, gets split in half, one half represents the private data only available to the data owner, the other half is available to the adversary. The adversary splits their data into member and non-members and uses this to learn the threshold ε^* as described in Section 3.4.5. The target t^* and the public synthetic data published by the data owner are available to the adversary. With this, the adversary computes the distance from t^* to its closest synthetic record produced by the synthesizer M ($L(t^*|R(t^*|M))$). If this distance is smaller than the threshold ε^*, the adversary labels the target record as a member of the private raw data.

risk score is defined as $\lambda = 1 - PC$. As an example, if a record with membership 1 was labeled as such with a confidence score of 0.7, the record's risk score will be 0.7. However, if a record that was a member of the training data was classified as a non-member with confidence 0.7, the risk score will be $\lambda = 1 - 0.7 = 0.3$. This means that members who were misclassified are scored at a lower risk than members who were correctly labeled.

For the distance-based approach, we define the risk score λ as the following: for every record $t^* \in D_{train}$ we will use the scaled distance d from t^* to its nearest synthetic record and define the risk score $\lambda = 1 - d$. Records with λ close to 1 are considered to be more at risk than records with λ close to 0. A record with $L = 0.1$ for example is far from its closest synthetic record, and was therefore likely not used in the training process; it is hence not considered to be at risk. For a record with a very small distance, hence a large risk, e.g. $\lambda = 0.95$, we assume that the record was used during training and is very similar to an existing record in the original training data.

3.5 Evaluation

3.5.1 Attack Evaluation

To evaluate the MIA success we compare the predicted memberships to the actual memberships. We call a true positive (TP) prediction a sample with a positive label being predicted as such. True negative (TN) predictions occur when a negatively labeled sample is correctly predicted as such. False positives (FP) describe data instances belonging to

the negative class but are labeled as positives. False negatives (FN) are samples that are labeled as negative although they belong to the positive class. For our evaluation, we use the following evaluation metrics:

- **Accuracy** describes the percentage of correctly classified samples:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.6}$$

- **Precision** gives the proportion of correctly positively classified samples over all positively labeled instances, i.e.

$$Precision = \frac{TP}{TP + FP} \tag{3.7}$$

- **Recall**, or True Positive Rate (TPR) or sensitivity, gives the probability of a positive observation being labeled as such and is calculated by

$$PR = \frac{TP}{TP + FN} \tag{3.8}$$

- **False Positive Rate (FPR)** is the rate of negative samples classified as positive, over the total number of negative samples, i.e.

$$FPR = \frac{FP}{FP + TN} \tag{3.9}$$

- **F1-score:** gives the harmonic mean of recall and precision and is calculated by

$$2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3.10}$$

- **Area Under the Receiver Operating Characteristic Curve (ROC-AUC)** is calculated as the area under the ROC curve, which plots the TPR (recall) against the FPR while the discrimination threshold is varied. Models achieving an ROC-AUC score of 1 classify every sample correctly, while an ROC-AUC score of 0.5 corresponds to a model that is randomly guessing.

We use the accuracy, precision, recall, F1-score, and ROC-AUC to compare the overall MIA success on models trained on the entire data set and models trained on the subset that excludes records at risk. The ROC-AUC measures a model's ability to distinguish between classes for different thresholds. This is why it is especially important for the MIA evaluation, as it shows to what degree the attack model can distinguish between members and non-members.

3.5.2 Evaluating Trends for Records at Risk

Once the risk score is computed for each data record, we evaluate trends concerning these specific records. We investigate whether outliers, which are detected by the algorithms described in Section 3.3.1, are more at risk than inliers.

3.5.3 MIA Defense

As a defense against membership inference, we propose to remove the records at risk from the training data. Figure 3.4 shows the scenario of attack and defense. We use different cutoff values α for risk scores and remove records with risk scores higher than $\lambda = \{0.8, 0.85, 0.9, 0.95\}$ ($\alpha = \{0.2, 0.15, 0.1, 0.05\}$ resp.) for the distance approach. We chose these four values to evaluate the effect of the cutoff value on the defense performance and the utility. We assume that removing over 20%, i.e. $\alpha > 0.2$, results in substantial utility loss, and therefore we will only consider cutoff values of $\alpha \leq 0.2$. For the shadow model approach we chose $\lambda = 0.65$ ($\alpha = 0.35$ resp.). Due to the lengthy run-time of the shadow data approach (around three weeks for the largest data set), and limited computational resources, we restrict our analysis to only one cutoff value.

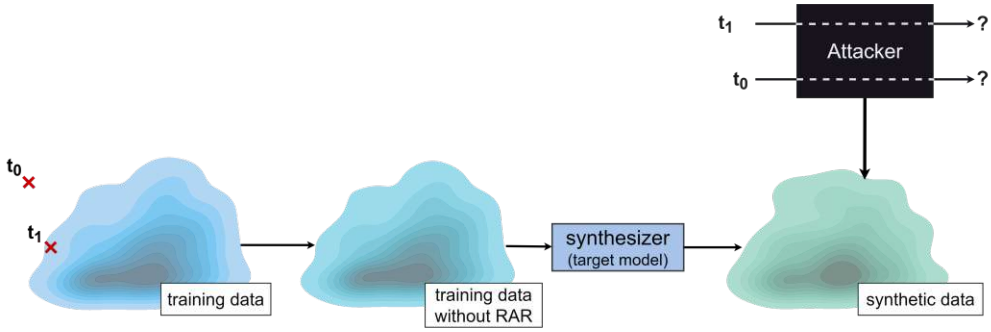

Figure 3.4: Defense: after identifying the records at risk from the training data, the data owner removes them, and thus publishes synthetic data that was generated by a synthesizer trained only on the records that are not at risk.

After generating new synthetic data from a model that was trained on data excluding records at risk, we perform the attack again. This way, we can investigate if i) the defense worked and the overall MIA success is smaller, and/or ii) there are new records at risk, and iii) if these new outliers coincide with the records that had the highest risk score of the records remaining after the removal. We will evaluate any changes in attack performance, with a focus on records previously labeled to be at risk. For this, the risk score can be computed and compared for both scenarios. Additionally, we want to reevaluate the attack risk for the remaining data records previously identified as records not at risk.

3.5.4 Data Utility

Data utility refers to data's ability to obtain meaningful insights, conduct analyses, and make accurate predictions. We call utility loss of synthetic data the situation when synthetic data is not able to maintain this standard compared to the original data. We measure utility by using evaluation metrics on data labels derived from five different classifiers: Random Forest, Naive Bayes, Logistic Regression, Support Vector Machine, and k-Nearest-Neighbor classifier. Data with evaluation scores similar to the original one is considered to have high utility. Consequently, the utility loss is defined by how much the prediction accuracy decreases for a target attribute of classifiers trained on synthetic data instead of the original training data. Synthetic data with high utility will yield no or only minimal prediction loss for the data's initial classification task. To evaluate the data utility and assess the utility loss, we use accuracy, precision, recall, and F1-score. Although accuracy is probably the most intuitive evaluation metric, we need precision, recall, and F1-score to account for unbalanced target classes. For highly imbalanced data, a model that always predicts the majority class might achieve a high accuracy simply because most records belong to the majority class. Low precision values imply that the model tends to classify negative labels as positive. A model that tends to label the positive class as negative will lead to a low recall. High values for recall and precision are desired, hence a model should be trained to achieve a high F1-score, which is the harmonic mean of precision and recall.

In addition to computing the overall evaluation metrics, we also investigate and evaluate per-class predictions. This is especially relevant for imbalanced thyroid data set.

After removing the records at risk from the training data and generating synthetic data from it, we compare the utility not only to the original data, but also to the synthetic data that was generated by a model trained on the entire original data set. We repeat this five times using different seeds and present the average over all seeds to get more robust estimates. In our analysis, we also look at the differences in utility between the different synthesizers listed in Section 3.4.1 and evaluate their relation to attack successes.

3.6 Deployment

In this phase, we summarized and visualized all results. These are presented in Chapter 4.

3.7 Summary

In this chapter, we presented the experimental set-up, following the CRISP-DM [73] guidelines. We highlighted the importance of privacy-preserving data measures and data understanding. We further presented the data sets used for our experiments and analysis, as well as the data preparation process. We then introduced the outlier detection algorithms used in our analysis, and the parameter settings used for them. We listed the synthesizers used for our experiments and defined our threat model. Additionally, we

explained the two attack approaches used and defined their corresponding risk scores. We then described the attack evaluation process and proposed a defense for the attacks. Lastly, we described the process of measuring and quantifying data utility.

CHAPTER 4

Experimental Analysis and Results

In this section, we present, discuss, and analyze the results obtained from our experiments. We first present the results for the shadow model approach, and then for the distance-based approach. For each approach, we summarize the overall attack scores, i.e. how well each attack approach is able to infer membership for every data set-synthesizer combination. After that, we look at the relationship between outliers and records at risk and try to determine if outliers are more vulnerable to MIA. After applying the defense, we subsequently reassess the MIA performance. We first reassess this for all data points, and then separately for records at risk and records not at risk only. Finally, we analyze the data utility and compare it for the original data to the synthetic data sets before and after the defense. We highlight the per-class utility for the imbalanced Thyroid data set, to showcase possible effects on the utility caused by the defense. In the last section of this chapter, we will compare the two approaches and analyze their performance based on membership status. For our experiments and analysis, we use Python version 3.9.17. For the utility experiments, we use the scikit-learn default parameter settings, as we simply want to detect the differences in utility, and do not strive surpass the state-of-the-art models' performances. The evaluation metrics are also computed using their scikit-learn implementations[1].

4.1 Shadow Model Approach

For this supervised approach, we split the data into raw and reference data. Each record of the raw data set is going to be chosen as the target record, and for each, two synthesizers are trained: one including and one excluding the target record (thus, the synthesizers

[1] https://scikit-learn.org/stable/modules/model_evaluation.html

we train amount to twice the number of data samples). For all the experiments using Bayesian Networks as a synthesizer, we use a fixed number of maximum parents per node of three, meaning a node can either have one, two, or three parents. During our research, we found that this preserves the utility at a high level while keeping the run times reasonably low. The overall attack evaluation (Section 4.1.1) is computed by considering members and non-members. The risk identification (section 4.1.2), however, is only done for members of the training set.

4.1.1 Overall Attack evaluation

We analyze the overall membership inference attack performance by computing the accuracy, precision, recall, F1-score, and ROC-AUC for each synthesizer and data set. Figure 4.1 shows these results, where the data sets are ordered ascending according to the number of records. The accuracy gives the proportion of records that were correctly labeled according to their membership. With accuracies around 0.5, the attack is practically as reliable as random guessing. Related studies have already shown that MIAs are no threat to the entirety of records contained in a data set, but rather single records that are especially vulnerable to the attack. The privacy gain (PG), which we described in more detail in Section 2.4, defined by Stadler et al. [8] for five randomly selected records, is around one for all of these five records. Records with $PG = 1$ were not able to be correctly labeled by an adversary. This seems to be in line with our results, showing that overall, there is no serious privacy risk for the data. However, while our results show low accuracies, for some scenarios, e.g. Heart or Thyroid data generated by a Copula, the recall scores are above 0.9. This means that for these two cases, over 90% of members were identified as such by the attack. TVAE seems to be more vulnerable with the smaller data sets. In contrast, Bayesian Networks seem to be more vulnerable when used with larger data sets. We also observe that data generated with the CTGAN is the least vulnerable to the attack, with accuracy and ROC-AUC values ranging between 0.47 and 0.53. Overall, the ROC-AUC values range from 0.47 (CTGAN) to 0.91 (Copula). A related study measured the ROC-AUC for Bayesian Networks with 10 random records [16]. The resulting values range from around 0.47 to 0.8. Our results lie within this range, with the lowest ROC-AUC value for Bayesian Networks at 0.61 (Breast Cancer data) and the highest value being at 0.8 (Thyroid data). The reason this attack performs poorly overall is that most data records will not influence the training of a model to the point where the model produces noticeably different synthetic data when in- and excluding the target record. We observe that only a small amount of records are influential enough to change the characteristics of the generated data and enable an attacker to recognize these differences.

To show the difference in accuracies between members and non-members, the accuracy for each of the two groups is computed and the results are listed in Table 4.1. The highest value in each row is colored red. Note that the values for members are equivalent to the recall values in Figure 4.1. For 12 out of 16 data sets, the members were predicted more accurately than the non-members. For some settings, e.g. Thyroid and Heart data with

Copula, or Breast Cancer and Cesarean data with TVAE, members were predicted with much higher accuracy than non-members. Here, the Heart data and Copula show the highest difference, with members obtaining a 0.81 higher accuracy than non-members. For other cases, like Bayesian Networks with Caesarian or Heart data, the accuracies for members and non-members do not differ a lot and both stay around 0.5. Only for the Caesarian data and Copula, as well as Thyroid data with TVAE, the non-members obtain a considerably higher accuracy than members (+0.18 and 0.22 resp).

Table 4.1: Accuracy of correctly labeled records by membership (shadow model approach)

		member	non-member
Bayes	Caesarian	0.44	0.50
	Heart	0.45	0.57
	Breast Cancer	0.56	0.47
	Thyroid	0.68	0.33
CTGAN	Caesarian	0.69	0.25
	Heart	0.45	0.57
	Breast Cancer	0.57	0.41
	Thyroid	0.62	0.42
TVAE	Caesarian	0.75	0.38
	Heart	0.69	0.33
	Breast Cancer	0.78	0.21
	Thyroid	0.39	0.61
Copula	Caesarian	0.38	0.56
	Heart	0.92	0.11
	Breast Cancer	0.64	0.49
	Thyroid	0.94	0.27

4.1.2 Risk Identification

We compute the risk score for each record in the raw data set according to the definition described in Section 3.4.6. First, we want to test whether the prediction confidence, i.e. the risk score, is higher for outliers detected by the algorithms. In the plots and tables within this section, *LOF10*, *LOF15* and *LOF20* denote outliers detected using the Local Outlier Factor with $k = \{10, 15, 20\}$ respectively, where k describes the number of nearest neighbors of a data point used to calculate the mean distance from. *Quant* describes the outlier definition by Stadler et al. [8], where records with attribute values outside the 95% quantile are classified as outliers.

First, we analyze **whether the risk score is higher for outliers**. For this, we formulate a t-test to test if the outlier risk (r) is higher for outliers than inliers. The hypotheses

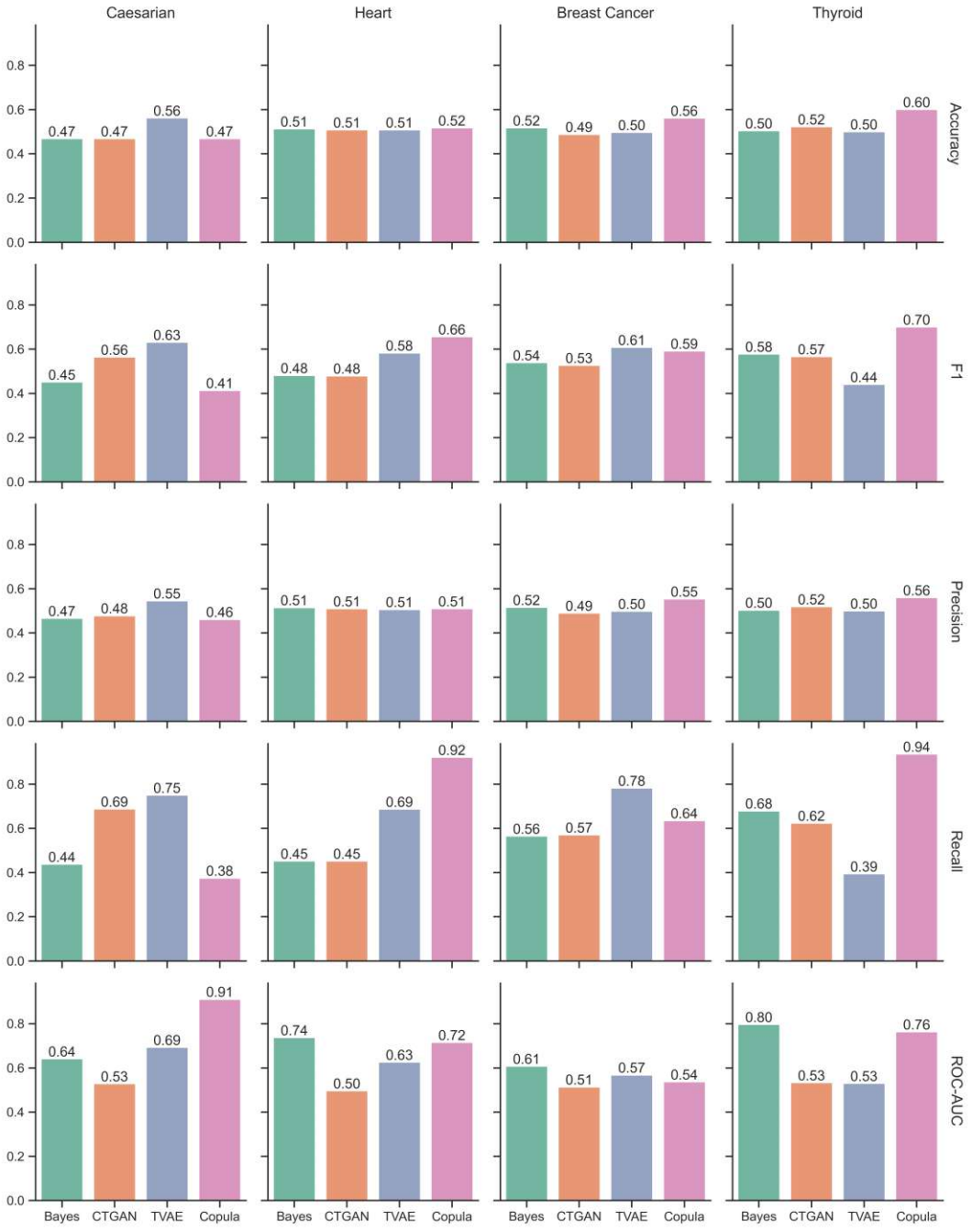

Figure 4.1: Overall MIA scores for the shadow model approach: accuracy, F1-score, precision, recall, and ROC-AUC for each data set and synthesizer combination.

are defined as the following:

$$H_{0_{S,1}} : \mu_{r_{out}} \leq \mu_{r_{in}} \tag{4.1}$$
$$H_{1_{S,1}} : \mu_{r_{out}} > \mu_{r_{in}}, \tag{4.2}$$

where $H_{0_{S,1}}$ denotes the Null-Hypothesis of the first test of the supervised attack $(S, 1)$, and $H_{1_{S,1}}$ denotes the respective alternative hypothesis. The remaining hypothesis tests will follow this syntax. The resulting p-values for this test can be seen in Table 4.2. Tests that are significant on a significance level of 0.05 are highlighted in red. In total, only four tests, all for the Thyroid data with CTGAN, returned a significant difference. For the reversed test:

$$H_{0_{S,2}} : \mu_{r_{out}} \geq \mu_{r_{in}} \tag{4.3}$$
$$H_{1_{S,2}} : \mu_{r_{out}} < \mu_{r_{in}}, \tag{4.4}$$

which tests if outliers have smaller risk scores than inliers, two tests (Heart with Bayesian Networks and TVAE) indicate significance. In Table 4.2 these are the p-values that are greater than 0.95. We, therefore, conclude that although outliers have, in some cases, higher risk scores than inliers, this difference is mostly not statistically significant.

Table 4.2: P-values for the test $H_{1_{S,1}} : \mu_{r_{out}} > \mu_{r_{in}}$ (shadow model approach)

		LOF10	LOF15	LOF20	iForest	q_outs
Caesarian	Bayes	0.424	0.424	0.424	0.769	0.136
	CTGAN	0.871	0.871	0.871	0.871	0.870
	TVAE	0.594	0.242	0.242	0.204	0.254
	Copula	0.899	0.899	0.899	0.899	0.925
Heart	Bayes	0.676	0.655	0.652	0.958	0.425
	CTGAN	0.528	0.338	0.338	0.602	0.871
	TVAE	0.909	0.731	0.711	0.845	0.972
	Copula	0.293	0.245	0.133	0.145	0.321
Breast Cancer	Bayes	0.297	0.063	0.056	0.703	0.693
	CTGAN	0.641	0.613	0.533	0.525	0.402
	TVAE	0.744	0.599	0.880	0.423	0.353
	Copula	0.165	0.474	0.474	0.898	0.129
Thyroid	Bayes	0.402	0.425	0.275	0.461	0.290
	CTGAN	0.022	0.024	0.004	0.334	0.042
	TVAE	0.278	0.520	0.690	0.338	0.100
	Copula	0.884	0.949	0.942	0.888	0.710

We now want to analyze **whether the attack accuracy is higher for outliers than for inliers**. To do this, we compute the accuracy for all outlying and inlying records found by the detection algorithms separately and compare them. Figure 4.2 shows the obtained results averaged over all four data sets. The plot shows that on average outliers

do have higher accuracy than inliers for data generated with Bayesian Networks, Copula, and CTGAN. For TVAEs the opposite seems to be true: inliers have higher accuracy values than outliers. To further look into this relation, we define a t-test where we test if the mean outlier accuracy ($\mu_{c_{out}}$) is significantly larger than the inliers' ($\mu_{c_{in}}$) with the following hypothesis:

$$H_{0_{S,3}} : \mu_{c_{out}} \leq \mu_{c_{in}} \tag{4.5}$$

$$H_{1_{S,3}} : \mu_{c_{out}} > \mu_{c_{in}} \tag{4.6}$$

We conduct this test for every combination of data set, synthesizer, and outlier detection algorithm and present the resulting p-values in Table 4.3. The significant p-values for a significance level of 0.05 are colored in red. For these cases, we conclude that outliers are significantly more at risk for MIA than inliers.

Table 4.3: P-values for the test $H_{1_{S,3}} : \mu_{c_{out}} > \mu_{c_{in}}$ (shadow model approach)

		LOF10	**LOF15**	**LOF20**	**iForest**	**Quant**
Caesarian	Bayes	0.338	0.338	0.338	0.252	0.578
	CTGAN	0.303	0.303	0.303	0.466	0.158
	TVAE	0.791	0.791	0.791	0.890	0.890
	Copula	0.209	0.209	0.209	0.366	0.718
Heart	Bayes	0.314	0.155	0.109	0.060	0.062
	CTGAN	0.786	0.451	0.368	0.095	0.300
	TVAE	0.589	0.790	0.790	0.740	0.203
	Copula	0.000	0.000	0.000	0.000	0.294
Breast Cancer	Bayes	0.209	0.610	0.842	0.341	0.928
	CTGAN	0.898	0.657	0.826	0.507	0.139
	TVAE	0.853	0.739	0.786	0.433	0.751
	Copula	0.930	0.930	0.955	0.477	0.894
Thyroid	Bayes	0.199	0.039	0.004	0.245	0.096
	CTGAN	0.658	0.686	0.347	0.191	0.280
	TVAE	0.262	0.054	0.083	0.234	0.930
	Copula	0.140	0.268	0.325	0.325	0.492

Only 6 out of 80 tests are significant: The heart data generated with a Copula and outlier detection algorithm LOF10, LOF15, LOF20, and iForest, as well as Thyroid data generated with a Bayesian Network using LOF15 and LOF20 as outlier detection. Note that the p-values for the Caesarian data and LOF10, LOF15, and LOF20 are the same for each synthesizer. This is because these three parameter settings all identify the same outliers. Generally, small p-values suggest that outliers could be predicted more accurately than inliers.

The significant tests for the reversed hypothesis:

$$H_{0_{S,2}} : \mu_{c_{out}} \geq \mu_{c_{in}} \tag{4.7}$$

$$H_{1_{S,2}} : \mu_{c_{out}} < \mu_{c_{in}}, \tag{4.8}$$

are then the settings with p-values greater than 0.95. As none of the p-values in Table 4.3 is greater than 0.95, we conclude that the accuracy of the outliers is never significantly smaller than the accuracy of inliers.

The aggregated accuracies over all data sets by in- and outliers can be seen in Figure 4.2. We observe that the outliers can be predicted with a slightly higher accuracy for all settings except Copula with outliers detected using the quantile method (-0.07), and all settings using TVAE, with an average difference in accuracy of 0.16. For the Bayesian Networks, the outliers detected with LOF could be predicted with 0.24 higher accuracy than the inliers. On the attacks on data generated with the CTGAN, outliers obtain, on average, a 0.12 higher accuracy than inliers. The accuracies for the Copula differ only slightly for in- and outliers. For all LOF outliers, the accuracy for the two groups is the same. The outliers detected using iForest obtain an accuracy that is 0.05 higher than the inliers' accuracy. For the quantile method, the outliers' accuracy is 0.06 lower than the inliers'. With these results, we conclude that whether outliers can be predicted more accurately highly depends on the synthesis model. Here, we find that the TVAE is the only synthesizer where outliers were predicted less accurately than inliers.

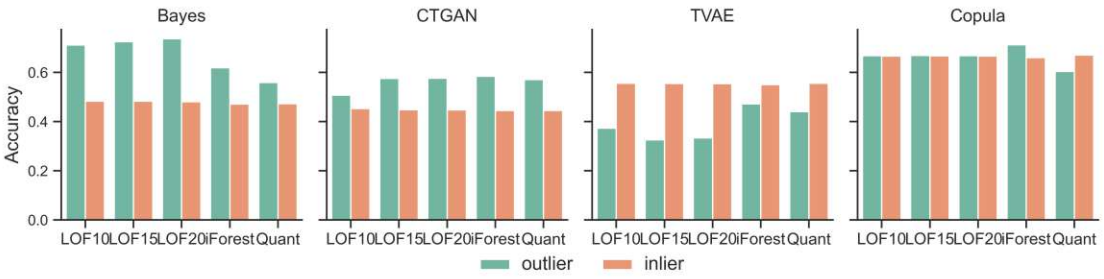

Figure 4.2: Accuracy for outliers vs. inliers for shadow model approach for members. Accuracy for every synthesizer and algorithm combination. Turquoise bars show the accuracies for outliers, and orange bars those of inlying records.

Additionally, we compare the accuracies for in- and outliers for non-members only. The results can be seen in Figure 4.3. Like for the members only, as seen in Figure 4.2, we observe that on average outliers are predicted with slightly higher accuracy for all synthesizers except the TVAE. On average, the outliers' accuracies are 0.02 higher than the inliers. Concerning the outlier detection algorithms, outliers identified using iForest show the highest difference in accuracy to inliers with 0.04 higher accuracy for outliers on average. The lowest difference of 0.012 is obtained for outliers detected using LOF20.

We now investigate if **outlier detection algorithms can predict the records at risk**. For this, we treat and evaluate the problem as a classification task where we set

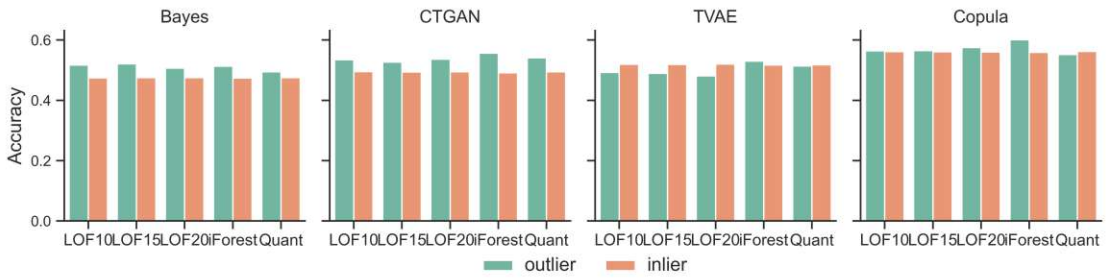

Figure 4.3: Accuracy for outliers vs. inliers for shadow model approach for non-members only. Accuracy for every synthesizer and algorithm combination. Turquoise bars show the accuracies for outliers, and orange bars those of inlying records.

the ground truth to be the variable *at_risk*. A record that was identified to be at risk according to its risk score will be labeled 1 for the variable *at_risk*, 0 otherwise. The outlier detection algorithms' labels are set to be the predictions. The assumption here is that the records that are at high risk for MIA are outliers, so we test if the outlier detection algorithms are suitable to predict records at risk. We compute the precision, i.e. the proportion of detected outliers that are at risk, and the recall, i.e. the proportion of all records at risk that were also labeled as outliers (see Figure 4.4). The results can

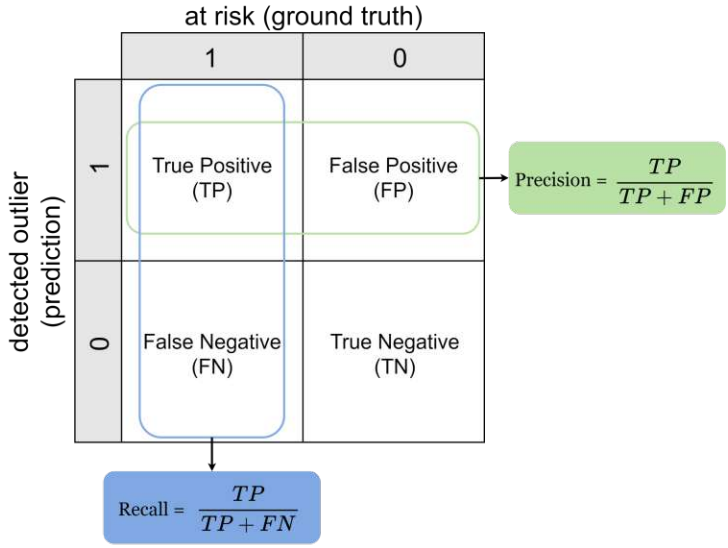

Figure 4.4: Confusion Matrix for outlier detection algorithms predicting records at risk.

be seen in Figure 4.5. With values at almost 0.4 for CTGAN, Bayesian Networks, and Copula, the precision scores show that up to around 40% of detected outliers were labeled at risk, according to the risk score. Recall scores on the other side are on average lower than the precision scores. For the Copula, however, they are between 30 to 40%. This means that up to 40% of records at risk were outliers.

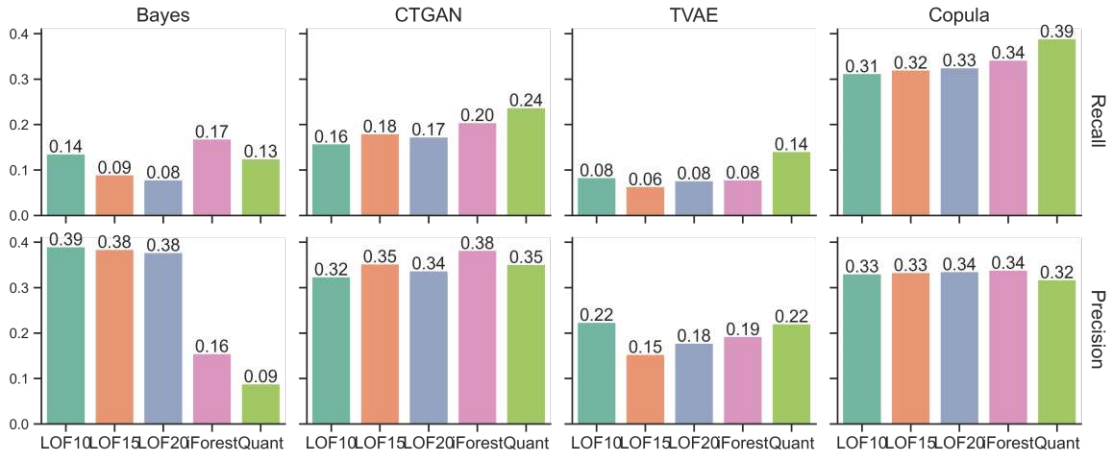

Figure 4.5: Relation between outliers and records at risk for shadow model approach. The plot displays how well outlier detection algorithms can predict records at risk with precision and recall.

Summarizing our analysis of the correlation between outliers and records at risk, we conclude that although outliers were predicted more accurately with the exception of synthetic data generated by TVAE (Figure 4.2), the increased risk for outliers is not substantial. The rather small values in Figure 4.5 and the p-values presented in Table 4.3 also suggest that outliers detected with LOF, iForest, or the Quantile method are not significantly more at risk for MIA than inliers.

4.1.3 Defense Evaluation

Once the records at risk are identified, we remove them from the original training data as part of the defense process. To evaluate this approach, after removal, the membership inference attack is run again, and the attack accuracy and risk scores are recomputed by running the attack for the remaining data samples. The overall attack accuracies are shown in Figure 4.6, where the dashed grey lines show the attack accuracy before and the bars after the defense. Note that these values differ from the values presented in Figure 4.1, as we only compute the accuracies for members of the training data.

In most cases, the defense was successful, as the overall accuracy decreased after the defense. In Figure 4.6, this is the case if the bar is below the dashed line. The exact values for the attack accuracy differences before and after the attack can be found in Table 4.4. Cases where the defense caused an increase in attack accuracy, and therefore made data more vulnerable, are colored in red. The defense is unsuccessful for the Caesarian data generated from a Copula (+0.23 accuracy), for the heart data generated with a Bayesian Network (+0.05 accuracy), as well as for the Thyroid data with the TVAE as a synthesizer (+0.11 accuracy). For all other scenarios, the defense results in a decrease in MIA accuracy between -0.05 (Heart data with TVAE), and -0.45 (Thyroid

data with Bayes). Here, the more the attack accuracy decreases, the more successful the defense.

Table 4.4: Defense success measured by MIA accuracy (shadow model approach)

	Caesarian	Heart	Breast Cancer	Thyroid
Bayes	-0.32	0.05	-0.14	-0.45
CTGAN	-0.12	-0.14	-0.18	-0.12
TVAE	-0.19	-0.05	-0.33	0.11
Copula	0.23	-0.31	-0.28	-0.27

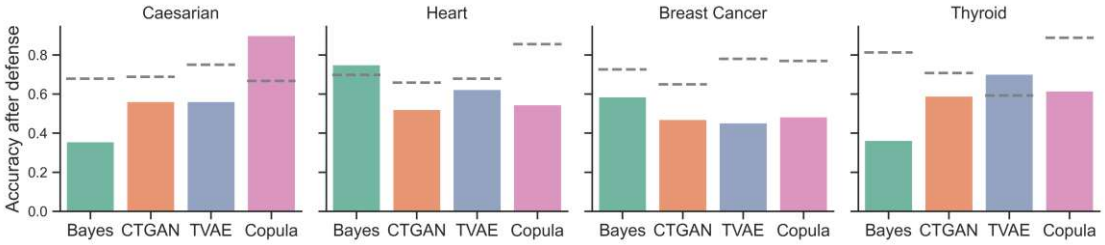

Figure 4.6: MIA accuracy before and after defense for shadow model approach: bars indicate the attack accuracy after the defense, dashed gray lines show the accuracy before.

We then analyze the defense evaluation in two parts. At first, we only analyze records at risk, followed by those not at risk. Figure 4.7 shows the accuracies for records at risk only. As before, the bars indicate the accuracies for all records at risk **after** the defense. The accuracy before the defense (gray dashed line) is always one in this case, as records at risk are always predicted correctly by the MIA by definition. We see that for the records at risk, the defense is highly effective, with the only exception being the Copula on the Caesarian data set, where the accuracy stays at one even after the defense. On average the defense led to a decrease in attack accuracy of 0.48 for CTGANs, 0.38 for Bayesian Networks, 0.35 for TVAEs, and 0.29 for Copula. However, the accuracies after the defense vary highly with accuracy reduction between zero (Caesarian data) and 0.66 (Breast Cancer data). The defense success for Bayesian Networks and TVAEs varies less, with a decrease in accuracy between 0.19 and 0.56, and 0.25 and 0.6 respectively. The values for the CTGANs are rather stable and range from 0.43 to 0.53. The defense is highly beneficial to the Breast Cancer data, where the attack accuracy decreases by 0.52 on average. For the Thyroid data, there is an average decrease of 0.37. For both the Caesarian and Heart data the accuracy decreases by 0.3. For most scenarios that show an increase in accuracy, these values are still below 0.6. This means that even though the attack caused the remaining data to be more vulnerable against the attack, there is still no serious threat to these records.

Although the defense can also decrease the risk for records **not** at risk, the opposite holds for the majority of the cases seen in Figure 4.8. Here, for 7 out of 14 scenarios, the

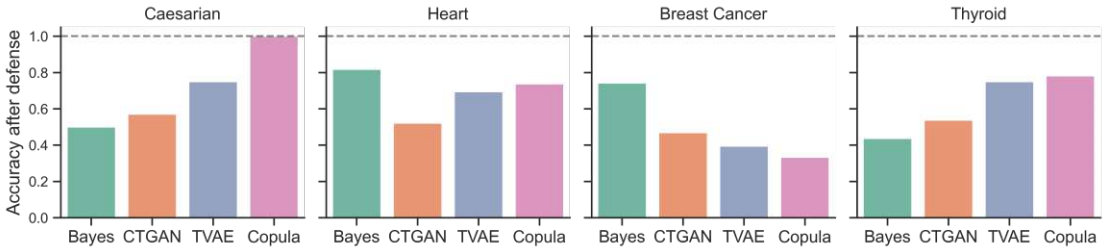

Figure 4.7: MIA accuracy before and after defense for shadow model approach for records at risk

defense causes an increase in attack accuracy. On average, the defense causes the attack accuracy to increase by 0.05. This means that the defense exposes the remaining records to a higher degree. For the Thyroid data using TVAE and Caesarian data using Copula, for example, the defense introduces a major privacy risk for the remaining records with an accuracy increase of 0.47 and 0.43 respectively. Furthermore, all four data sets in combination with the CTGAN show increased accuracies between 0.13 and 0.23.

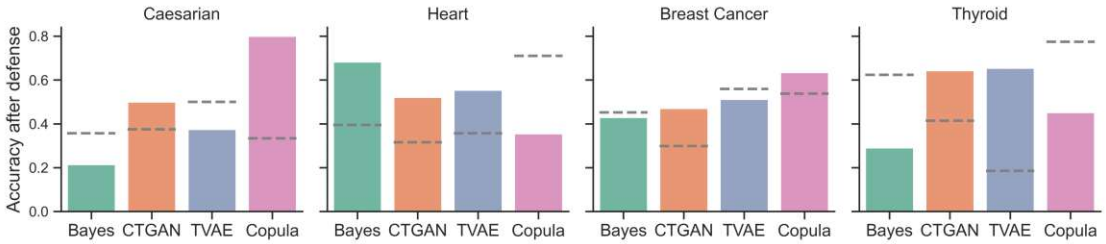

Figure 4.8: MIA accuracy before and after defense for shadow model approach for records not at risk

Next, we look at the recomputed risk scores of the remaining data that was not removed during the defense process and calculate the percentage of records newly exposed at risk as a consequence by the defense (Figure 4.9). We find that by the defense, new vulnerable records can be created. The percent of records at risk for the original data before the defense is shown by the dashed line in Figure 4.9. Up to 19% and on average 7.8% of records previously not at risk can be turned to be at risk by the defense process. However, only the defense on the Breast Cancer data generated by a Copula produces a higher percentage of records at risk than before the defense. The percent of records at risk for the Heart data decreased by 54% on average. For the Thyroid data, there are on average 76% fewer records at risk after the defense. For the Caesarian and Breast Cancer data, the average decrease in the amount of records at risk is 65% and 55% respectively.

Figure 4.10 shows the distributions of risk scores before and after the defense. Although there are no major differences, some distributions change slightly. In combination with Figure 4.6 and Figure 4.9 we can see the connection between defense success, new records at risk, and risk score distribution. The defense for the Caesarian data generated with a

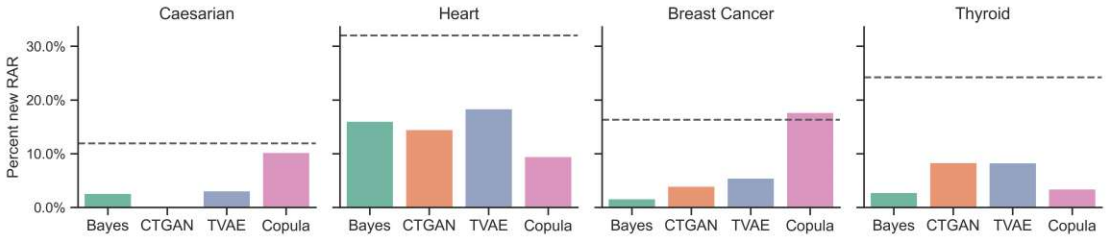

Figure 4.9: Percent of new records at risk caused by defense for shadow model approach

Copula, for example, caused the overall accuracy of the attack to increase (see Table 4.4). Although there are 1.6% fewer records at risk after the defense, the risk score increased. This means that even though there are fewer records at risk, these records' risk scores are generally higher, which makes them more vulnerable to the attacks, and the overall MIA accuracy increases. For the Breast Cancer data with Copula, we observe an increase in risk score and records at risk, as well as attack performance after the defense. For the cases where the defense caused the desired outcome and decreased the overall attack accuracy, we generally observe a decline in the percent records at risk and similar or slightly lower risk scores.

Lastly, we look at the new outliers detected for the remaining data after the defense and compare the risk scores of these outliers to the inliers' risk scores. In Figure 4.11 we see that the risk scores for outliers and inliers do not differ much. Only for the outliers detected using the Local Outlier Factor algorithm, we detect slightly higher risk scores for outliers with Copula, TVAE, and Bayesian Networks. However, these differences are very minimal. The outliers detected using iForest with synthetic data generated by CTGAN and TVAE tend to have slightly lower risk scores. Although, again, this difference is not significant.

4.1.4 Utility Assessment

For the utility assessment, we train five classifiers, namely Logistic Regression, Support Vector Machine, k-Nearest-Neighbor, Random Forest, and Naive Bayes, on the original data, the synthetic data, and the synthetic data generated by a synthesizer that was trained only on records not at risk of each of the four data sets described in Section 3.2. We then measure the data utility via the prediction accuracy of each classifier to compare the original and the two synthetic data sets.

Figure 4.12 shows the accuracies for every synthesizer-data set pair. Here, *synthetic* denotes the synthetic data that was sampled from a model trained on the entire training data, while *synthetic without RAR* stands for the synthetic data where the synthesizer was only trained on records not at risk. Although the utility, here measured via accuracy, is lower for synthetic data than for the original, there is mostly no substantial difference for the synthetic data generated with and without records at risk inside the training set. Only when using a Copula as a synthesizer, did the synthetic data without RAR achieve

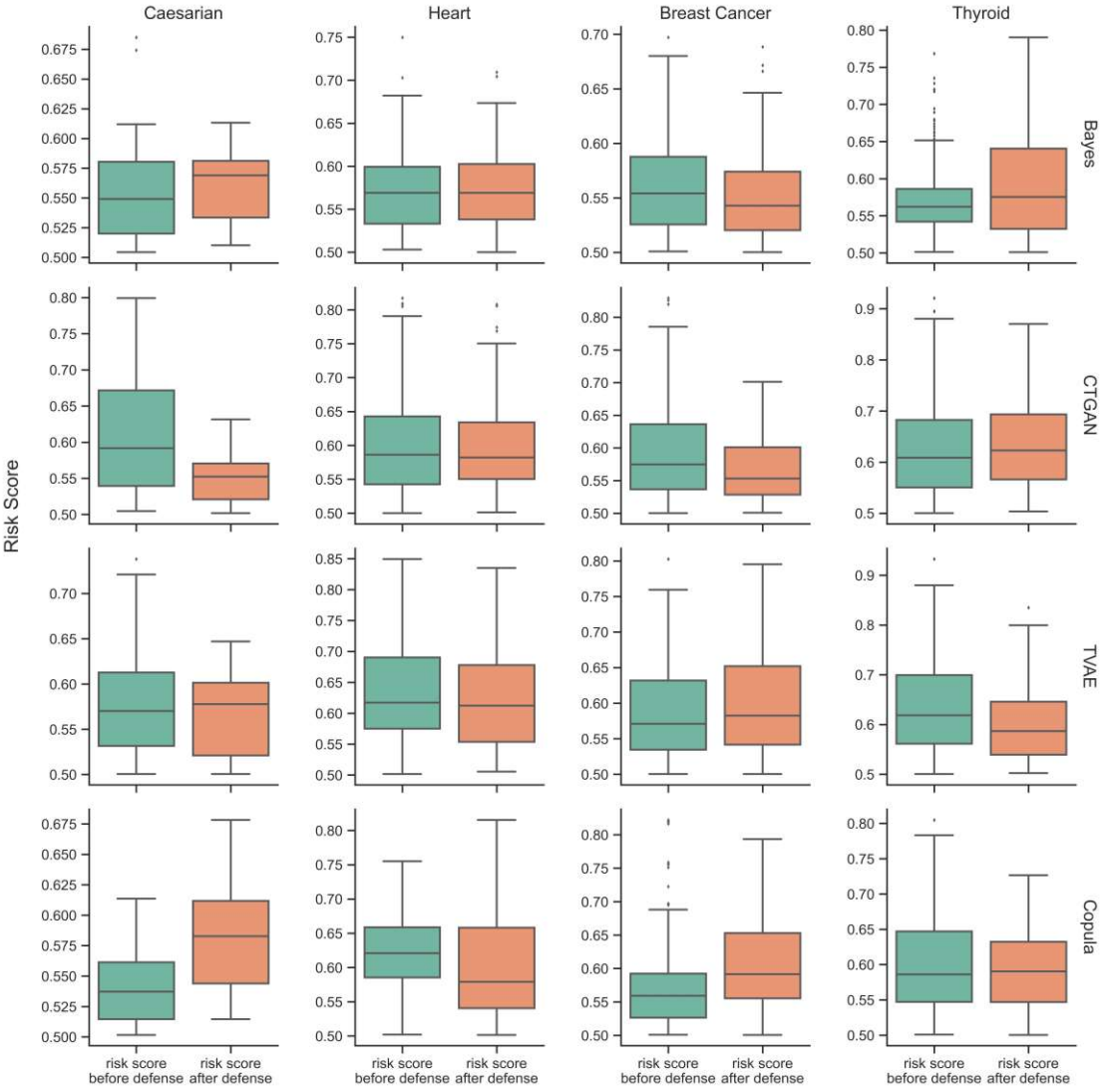

Figure 4.10: Distribution of risk score before and after defense for the shadow model approach

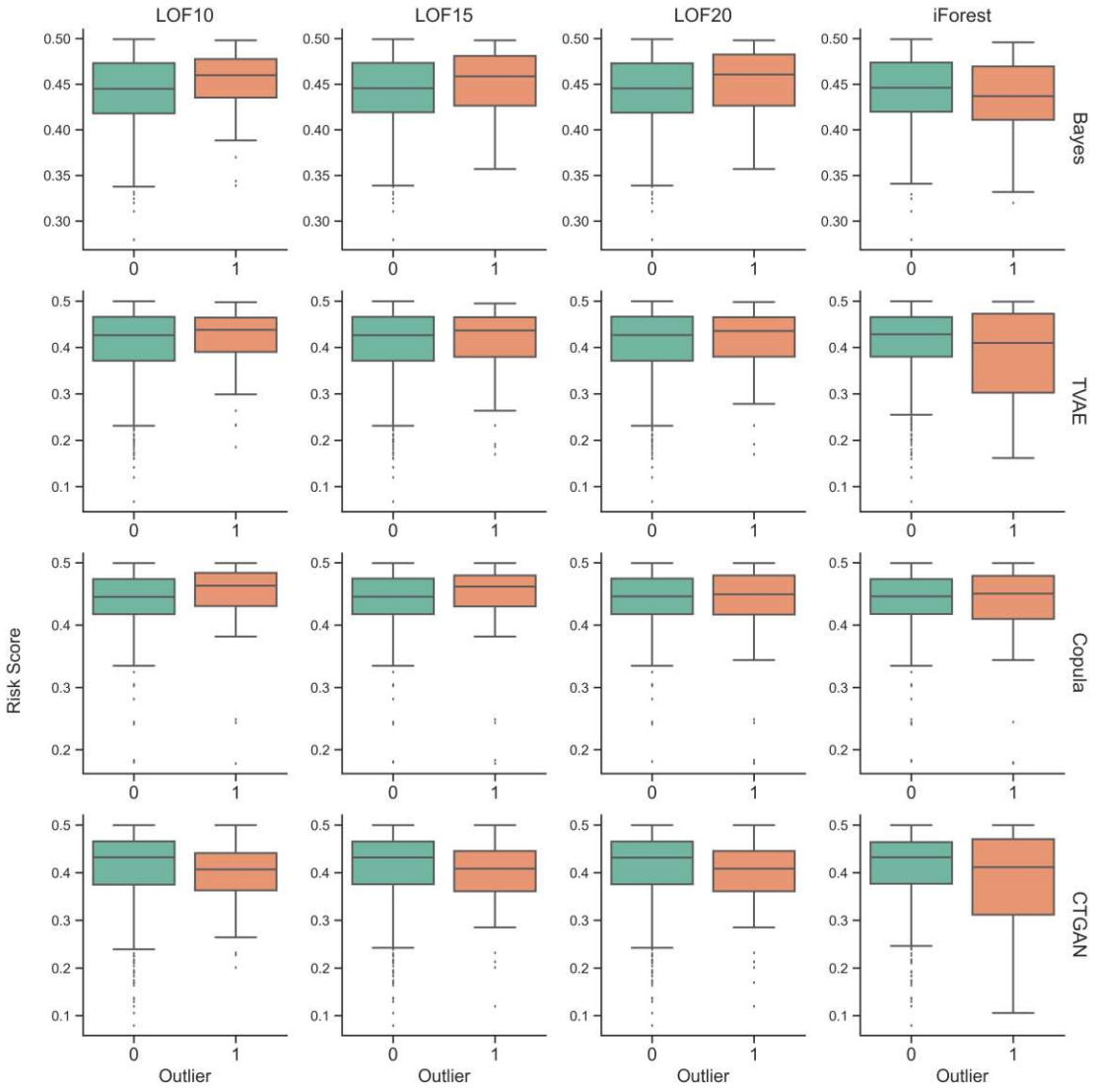

Figure 4.11: Distribution of risk scores for new outliers

notably lower accuracy scores than the synthetic data where the Copula was trained on the entire training set. A Copula trained on the entire data set produces synthetic data with a loss of accuracy between 0.02 and 0.16. This is in line with what the authors of [81] show, where the decrease in accuracy for data sets with 1,500 records or less ranges between 0.15 and 0.25. Other work shows a smaller loss of accuracy, ranging from 0.02 to 0.1 [19]. The data generated using the Bayesian Network suffers none to minimal loss of utility of around 0.02, with the exception of the Heart data, where the accuracy decreases by 0.07 for the synthetic data set and 0.06 for the synthetic data trained on a Bayesian Network excluding records at risk. The empirical study conducted in [19] finds a decrease in accuracy between zero and 0.03.

The synthetic data generated by a CTGAN shows a large decrease in accuracy for the Caesarian data (-0.06), and an even bigger decrease for the Heart (-0.31) and the Breast Cancer data (-0.41). This is also what the authors of [81] found, where the decrease in accuracy for data sets of up to 2,200 records ranges between 0.15 and 0.9 when using the CTGAN. The maximum loss of accuracy for the TVAE is 0.1, the minimum is 0.01. This is also in line with the study done by [82], where the accuracy loss for VAEs ranges from zero to around 0.2.

Even though we mostly could not observe any substantial utility loss when comparing the two synthetic data sets (except Heart and Breast Cancer with Copula and Bayesian Networks with Caesarian), we also need to look at the per-class utility for the imbalanced Thyroid data set. For this data set, classes one and two are the minority classes, whereas class three is the majority class with around 95% of records. When looking at Figure 4.13, we see that the accuracy of the majority class, class three, seems quite stable over the original as well as the two synthetic data sets. Classes one and two, however, suffer substantial utility loss with most accuracy values of almost zero. Only the Bayesian Network is able to maintain a high level of utility for class one, for class two this is not the case. This is, again, mostly the case for both of the two synthetic data sets. Additionally, the TVAE data seems to be less affected by the defense, as its generated data achieves higher accuracy values for almost all synthetic data sets generated without RAR than synthetic data generated from a TVAE with all training samples available. Additionally, the recall values for the minority classes are noticeably higher for TVAE than for other synthesizers, with the exception of Bayesian Networks for class one. Overall, the TVAE preserves the utility best for all three classes. The CTGAN and Copula are not able to preserve the utility for any of the minority classes.

In Figure 4.14 we look at the target variable distribution of the entire data set and compare it to the distributions of detected outliers. We find that the outliers' distributions are very similar to the original, especially for LOF15 and LOF20. Here the difference is 0.04 and 0.02 at largest, respectively. For the iForest and quantile method, Class 1 is a lot larger than for the original, whereas Class 2 is a lot smaller.

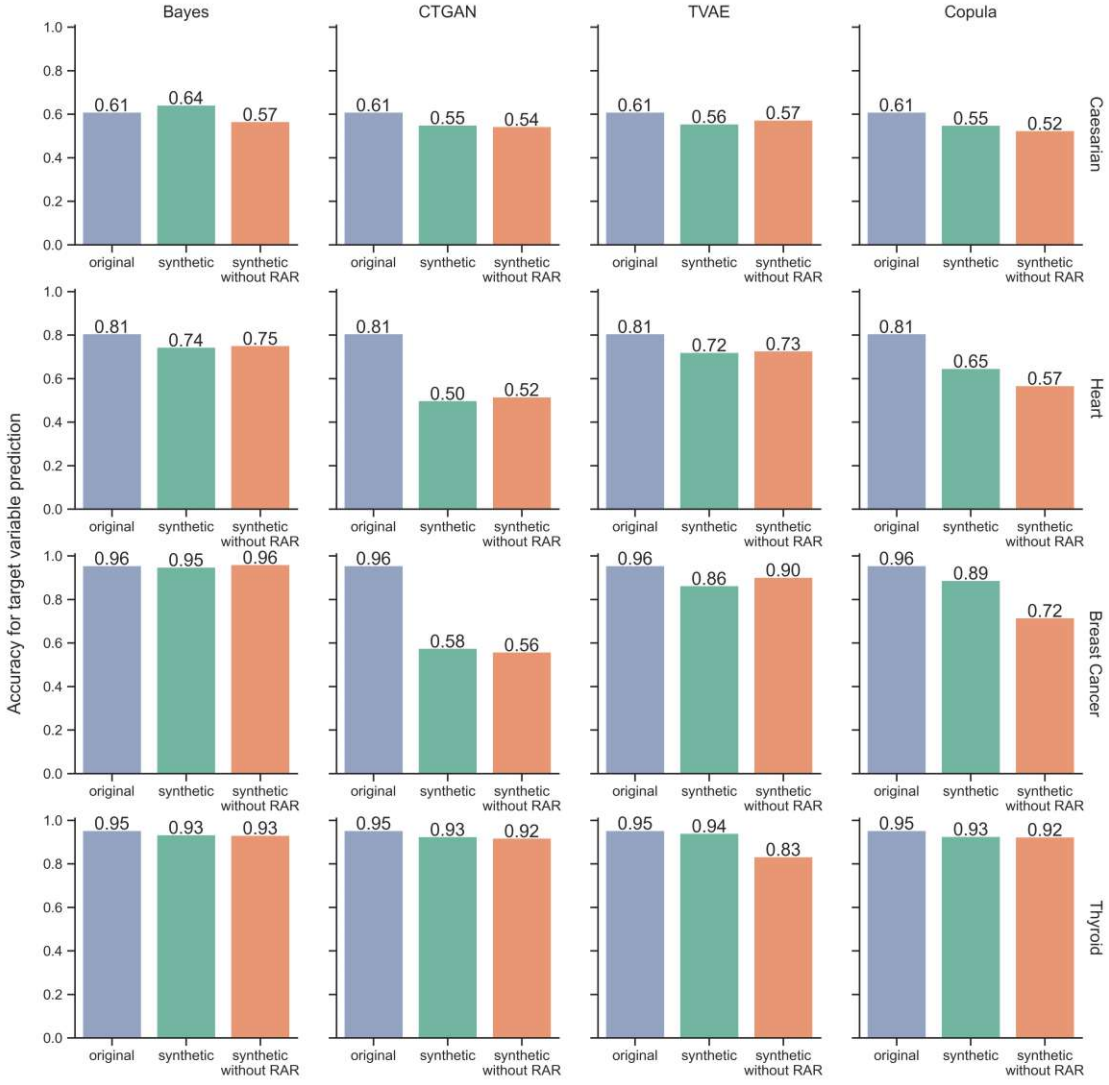

Figure 4.12: Utility comparison by synthesizer for shadow model approach

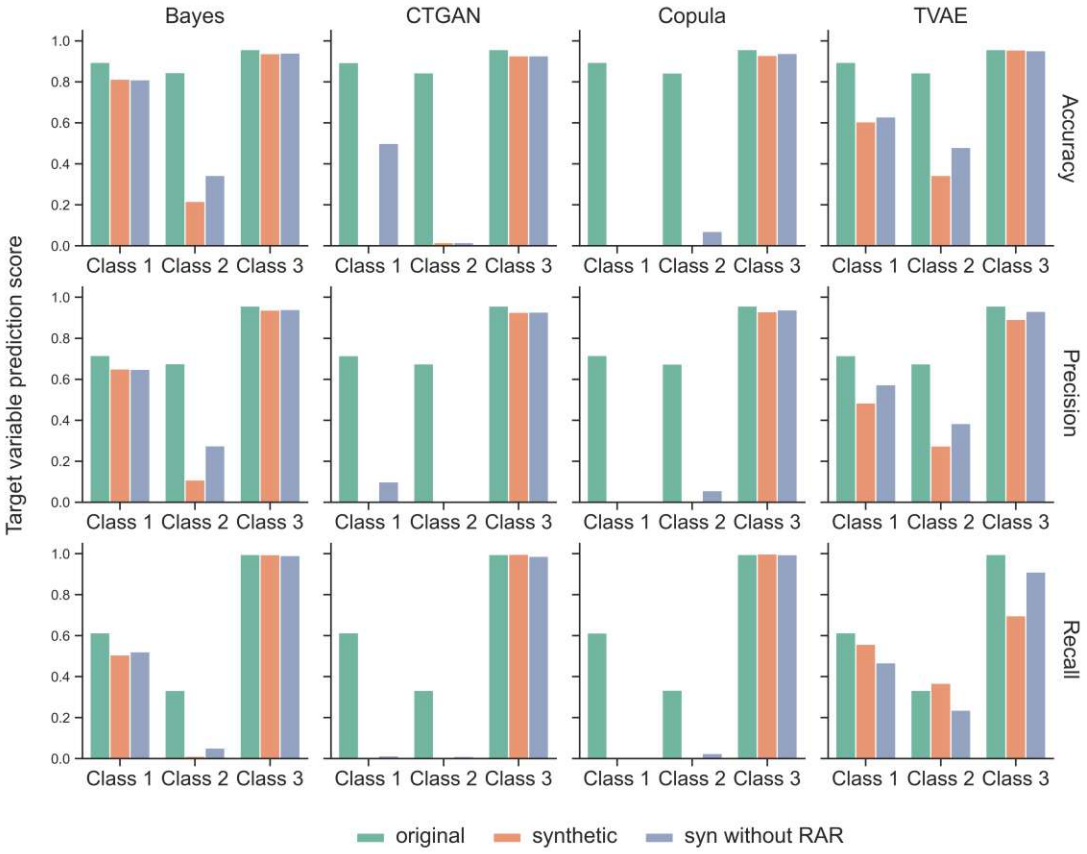

Figure 4.13: Utility comparison by class for shadow model approach

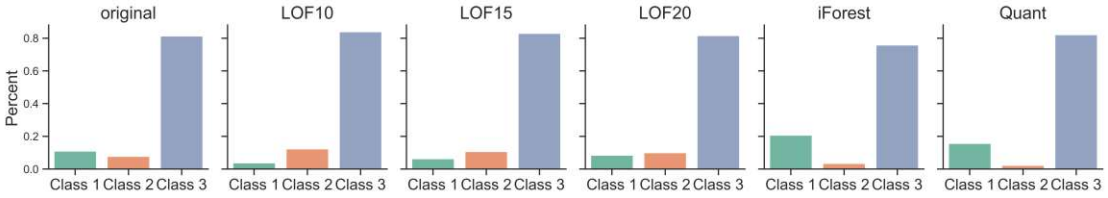

Figure 4.14: Outliers detected per class for Thyroid data: original distribution of the target variable compared to the target variable distribution per outlier detection algorithm.

4.2 Distance-based Approach

For the distance approach, we repeat the experiment five times, using five different seeds and aggregating the results obtained from these repetitions. Note, however, that the split of raw and reference data remains the same for each repetition, but we train a new synthesizer, and generate data from it, in each repetition. We show the results as averages of the evaluation metrics over the five experiment repetitions. All parameters used for this approach, plus their description are given in Table 4.5. As MIA defense, records with a distance d larger than the quantile $q = \{0.8, 0.85, 0.9, 0.95\}$ are removed from the training data.

4.2.1 Overall Attack evaluation

Table 4.5: Description of Parameters used for the Distance-Based Approach

Parameter	Description
d	Distance from a real record to its closest synthetic record
λ	Threshold value for risk identification
p	Maximum number of parents for Bayesian Networks (Data Synthesizer)
α	Cutoff for risk scores. As defense, records with risk scores larger than the $1 - \alpha = \lambda$ quantile will be removed from the training set.

As the distance-based approach builds on the assumption that records in the training set have smaller values for distance d than records outside the training set, we test if the assumption holds. For this, we conduct a t-test, where under the null hypothesis (H0) the mean distance for members (d_1) is smaller than the mean distance of non-members (d_0), i.e.

$$H_{0_{U,1}} : \mu_{d_1} \geq \mu_{d_0} \tag{4.9}$$

$$H_{1_{U,1}} : \mu_{d_1} < \mu_{d_0} \tag{4.10}$$

Similarly to the shadow model approach, we here denote $H_{0_{U,1}}$ the Null-Hypothesis for the first statistical test for this unsupervised approach. $H_{1_{U,1}}$ denotes the alternative hypothesis. We test this for every data set-synthesizer combination and find that, with a significance level of 0.05, 71% of our experiments show significantly smaller distances d for records used during training, which is the assumption the methods build on. None of the tests can conclude the opposite, that records used for training have significantly larger distances than records that are not in the training data. Figure 4.15 shows the p-values by synthesizer. The generally small p-values hint that the assumption that members have smaller distances to their nearest synthetic data record holds. We observe that records used to train a Bayesian Network have significantly smaller distances than records that were not used during training. This could imply a possible vulnerability for Bayesian Networks.

Next, we evaluate the overall MIA by the metrics accuracy, precision, recall, F1, and ROC-AUC. Figure 4.16 shows the corresponding results. The recall values for MIA are

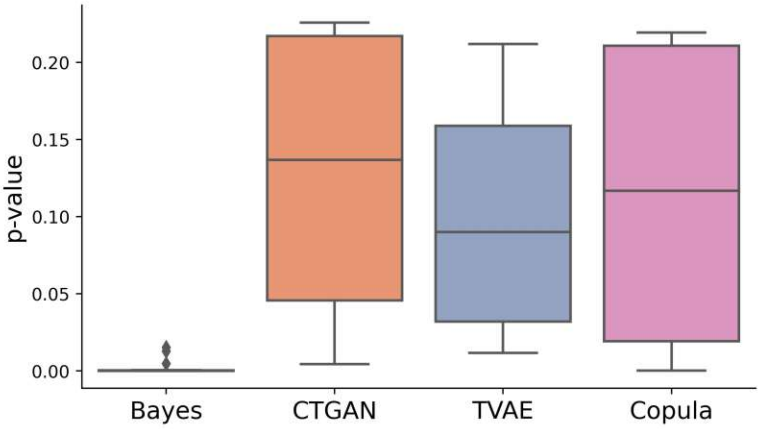

Figure 4.15: P-value-distribution for members vs. non-members: the box-plots show the distribution of p-values for the hypothesis that the mean distance μ_d is smaller for records inside the training data (Equations (4.9) and (4.10)).

particularly important, as they describe the proportion of members that were labeled as such. Although the threat is small for values around 0.5, some cases produce recall values of 0.7 and higher. For synthetic data generated by a Copula trained on the heart data, the MIA works notably well with a recall value of 0.82. This means that the adversary is able to correctly infer membership on 82% of members of the training set. As Bayesian Networks obtain, with the exception of recall for Heart data, the highest values for every evaluation metric and data set, we conclude that these models are the most vulnerable to MIA. Attacks on synthetic data generated by CTGAN or TVAE show very low attack success for all metrics and are therefore little to no threat to the data. The ROC-AUC values for CTGAN and TVAE range between 0.52 and 0.56, and 0.51 and 0.72 respectively. The authors of [9] show ROC-AUC scores of up to 0.7 with the CTGAN and 0.77 with the TVAE. Especially for the CTGAN, their results are higher than ours. However, their attack assumptions are more relaxed, as they assume an adversary to have access to the target model, which they can use to generate synthetic data. For our attack, we assume the adversary is more limited as we assume the adversary only has access to the synthetic data set a data holder publishes. Additionally, we find that the data set size seems to affect the attack: The bigger the data set, the less accurate the attack's membership predictions. N.b. that in Figure 4.16, the data sets are arranged from smallest to largest.

We explicitly look at the results of the Bayesian Network synthesizer and the relation between parameter p, the maximum number of parents, and the MIA outcome. The results are presented in Figure 4.17, from which we can observe a distinct trend: for larger p, the attack is more accurate. The two smallest data sets, Caesarian and Heart, in combination with a high number of maximum parents, are especially vulnerable. The Bayesian Network trained on the Caesarian data with $p = 4$, for example, obtains an accuracy of 0.78. For the Heart data, this value is 0.75. With these values, the attacks

can pose a serious privacy threat.

In Table 4.6, we also analyze the differences in accuracies between members and non-members, as we did for the shadow model approach in Table 4.1. In 12 out of 16 settings, the non-members were predicted more accurately – which is the exact opposite of the outcome of the shadow model approach. Here, however, the differences between members and non-members are smaller than for the shadow model approach. Notably, for synthetic Thyroid data generated with a Copula, no member was predicted as such, while all non-members were labeled correctly – this means that all records were labeled as non-members.

Table 4.6: Proportion of correctly labeled records by membership (distance-based appraoch)

		member	non-member
Bayes	Caesarian	0.76	0.70
	Heart	0.67	0.72
	Breast Cancer	0.57	0.66
	Thyroid	0.56	0.60
CTGAN	Caesarian	0.60	0.50
	Heart	0.49	0.60
	Breast Cancer	0.46	0.57
	Thyroid	0.53	0.51
TVAE	Caesarian	0.55	0.64
	Heart	0.45	0.67
	Breast Cancer	0.43	0.60
	Thyroid	0.49	0.55
Copula	Caesarian	0.54	0.62
	Heart	0.82	0.24
	Breast Cancer	0.52	0.53
	Thyroid	0.00	1.00

4.2.2 Risk Identification

We compute the risk score for each member of the original training data set as described in Section 3.4.6.

First, we evaluate if the **the attack risk is different for outlying versus inlying records**. As previously mentioned we use the distance d as the risk score for this attack method. This is done similarly as in Equations (4.7) and (4.8), where we used the confidence to compare the vulnerability for records in a shadow model attack. Unlike the shadow model approach, no research on whether outliers are more at risk for this attack than inliers has been published. This is why we first use a two-sided t-test to assess a

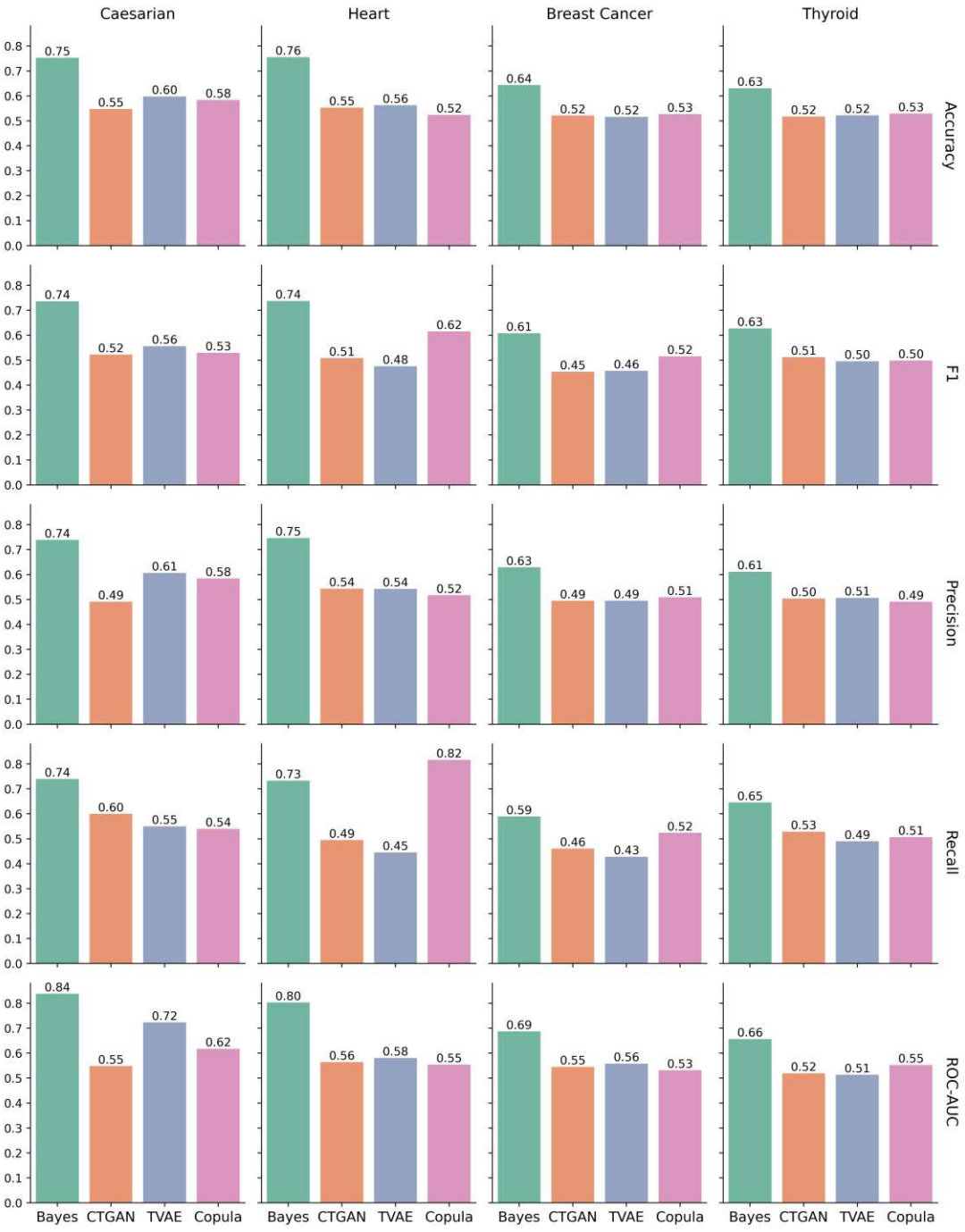

Figure 4.16: Overall MIA evaluation for the distance-based approach: the evaluation metrics accuracy, F1-score, precision, recall, and ROC-AUC are visualized for each data set and synthesizer.

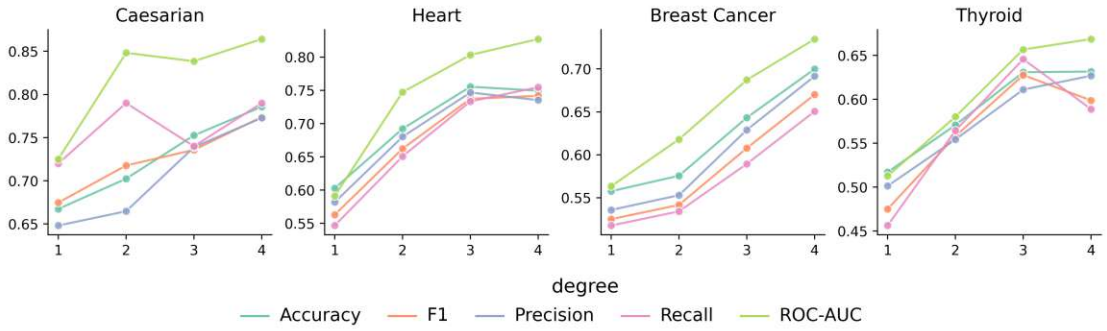

Figure 4.17: Overall MIA evaluation for Bayesian Networks with the distance approach: the evaluation metrics accuracy, precision, recall, F1, and ROC-AUC are visualized for different values for parameter p (maximum number of parents).

possible difference between the two groups. The hypotheses are defined as the following:

$$H_{0_{U,2}} : \mu_{d_{out}} = \mu_{d_{in}} \tag{4.11}$$

$$H_{1_{U,2}} : \mu_{d_{out}} \neq \mu_{d_{in}} \tag{4.12}$$

Here $\mu_{d_{out}}$ and $\mu_{d_{in}}$ describe the mean distances for out- and inliers respectively. The distribution of the resulting p-values can be seen in Figure 4.18. We find that mostly for outliers detected with the Local Outlier Factor algorithm the $H_{0_{U,2}} : \mu_{d_{out}} = \mu_{d_{in}}$ is rejected, meaning that for those detected outliers, the distances differ from inliers. High p-values, like for TVAE and LOF20, point to a test that accepts the null hypothesis and hence conclude that there is no difference for mean distances for out- and inliers. The tests show small p-values for outliers detected with LOF10 for all synthesizers and LOF15 with Bayesian Network and CTGAN. This implies that for these cases the distance d is smaller for outliers than inliers. Overall, tests on data produced by CTGANs obtain smaller p-values for all outlier detection algorithms. For Copula the opposite holds: p-values are larger throughout all detection algorithms. P-values for tests using LOF20 and TVAE accumulate around 0.5, this signifies that there is no difference in distance d for in- and outliers.

Since the two-sided t-tests only test whether the mean distances for the two groups are equal or not, we now conduct one-sided t-tests. With this, we can gain insight into whether the distance for outliers is significantly smaller or larger than the distance for inliers. Recall that smaller distances present a greater risk of being correctly labeled as a member. We thus use a t-test with the following hypotheses:

$$H_{0_{U,3}} : \mu_{d_{out}} \geq \mu_{d_{in}} \tag{4.13}$$

$$H_{1_{U,3}} : \mu_{d_{out}} < \mu_{d_{in}} \tag{4.14}$$

If we reject the null hypothesis, we conclude that outliers have smaller distances d, and are therefore more at risk for MIA. Overall, only 10% of these tests are significant on an

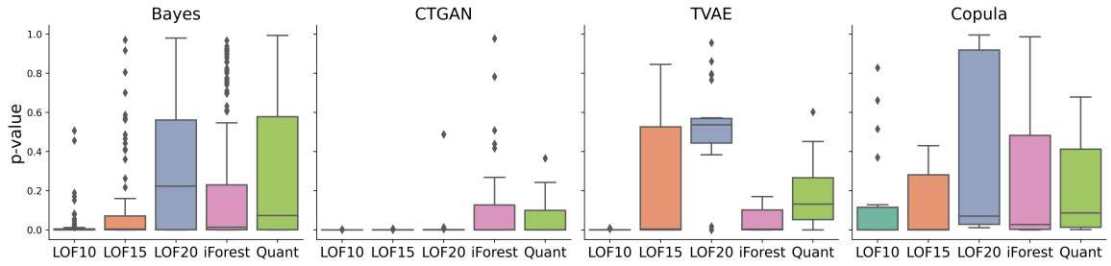

Figure 4.18: P-value-distribution for outliers vs. inliers: the box-plots show the distribution of p-values for the two-sided t-test with the hypothesis that the mean distance μ_d is different for outlying and inlying records eqs. (4.11) and (4.12).

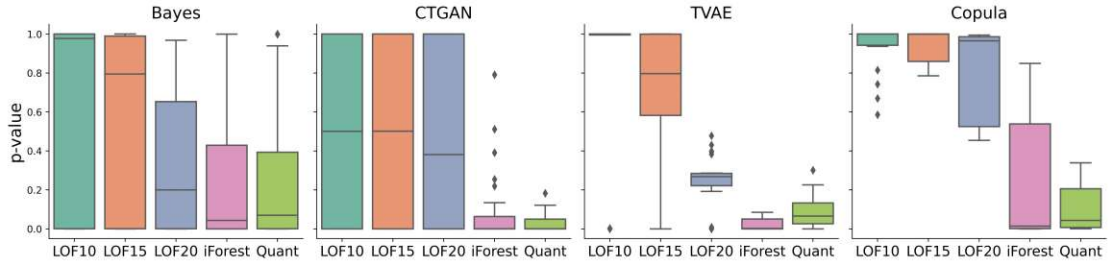

Figure 4.19: P-value-distribution under $H_{0U,3} : \mu_{d_{out}} \geq \mu_{d_{in}}$: the box-plots show the distributions of p-values of a one-sided t-test, testing if the mean distance μ_d for outliers is larger than for inliers Equations (4.13) and (4.14).

0.05 significance level. Each cell in Table 4.7 represents various tests for different cutoff values and experiment repetitions. The values represent the percent of significant tests for each synthesizer, data set, and outlier algorithm. Data sets for which no single test returned significance are excluded from Table 4.7; specifically, for Heart and Thyroid, no test returned significance, while for Caesarian, some tests for Bayesian Networks were significant. For the Breast Cancer data, all synthesizers returned significant tests. Most of the significant tests thus come from the Breast Cancer data. The results suggest that outliers detected with the LOF algorithm in the Breast Cancer data set are more at risk than inliers. The outliers detected with the quantile method and iForest have significantly smaller values for d in combination with the Caesarian data. The p-value distributions for these tests are shown in Figure 4.19.

On the other hand, for the opposite test with the hypotheses:

$$H_{0_{U,4}} : \mu_{d_{out}} < \mu_{d_{in}} \tag{4.15}$$

$$H_{1_{U,4}} : \mu_{d_{out}} \geq \mu_{d_{in}}, \tag{4.16}$$

high p-values in Figure 4.19 are significant, as these hypotheses are symmetric to Equations (4.13) and (4.14). 29% of tests under this H_1 return significant p-values. This indicates that **inlying records are at higher risk** for these cases. The percent of

Table 4.7: Percent of significant tests for $H_{1_{U,3}} : \mu_{d_{out}} < \mu_{d_{in}}$

		LOF10	LOF15	LOF20	iForest	Quant
Caesarian	Bayes	0.0	0.0	0.0	37.5	12.5
Breast Cancer	Bayes	88.8	62.5	5.0	0.0	0.0
	CTGAN	100.0	100.0	93.3	0.0	0.0
	TVAE	100.0	53.3	0.0	0.0	0.0
	Copula	93.3	66.7	66.7	0.0	0.0

significant tests per data, synthesizer, and outlier detection algorithm combination can be seen in Table 4.8. Especially for the Thyroid data, the hypothesis that the distance d is larger for inliers seems to hold. The results also suggest that outliers detected using iForest or the quantile method are at a lower risk than inliers.

Table 4.8: Percent of significant tests for $H_{1_{U,4}} : \mu_{d_{out}} \geq \mu_{d_{in}}$ for each data set, synthesizer and outlier detection algorithm combination

		LOF10	LOF15	LOF20	iForest	Quant
Caesarian	Bayes	0.0	0.0	0.0	2.5	0.0
	CTGAN	0.0	0.0	0.0	3.3	6.7
	TVAE	0.0	0.0	0.0	21.7	10.0
	Copula	0.0	0.0	0.0	6.7	6.7
Heart	Bayes	0.0	0.0	0.0	28.8	36.2
	CTGAN	0.0	0.0	0.0	70.9	43.6
	TVAE	0.0	0.0	0.0	73.3	40.0
	Copula	0.0	0.0	0.0	60.0	53.3
Breast Cancer	Bayes	0.0	0.0	1.2	80.0	0.0
	CTGAN	0.0	0.0	0.0	100.0	8.3
	TVAE	0.0	0.0	0.0	100.0	0.0
	Copula	0.0	0.0	0.0	100.0	0.0
Thyroid	Bayes	100.0	100.0	100.0	100.0	100.0
	CTGAN	100.0	100.0	100.0	100.0	100.0
	TVAE	100.0	100.0	100.0	100.0	100.0
	Copula	0.0	0.0	0.0	0.0	75.0

As the assumption of members having a smaller distance d holds (see Figure 4.15), the smaller d for members, the higher their risk for the attacks. Non-members are more likely to have larger values for d and are therefore more likely to be labeled as such. Next, we look at the **relation between outliers and records for which membership was correctly predicted**. As recent studies claim outliers are more at risk for MIA [8, 16], we test if outliers' memberships can be predicted more accurately. For this, we again design a t-test that tests if the attack accuracy is higher for outliers than for inliers.

With $\mu_{c_{out}}$ and $\mu_{c_{in}}$ being the accuracy for out- and inliers respectively, the hypotheses are defined as follows:

$$H_{0_{U,5}} : \mu_{c_{out}} \leq \mu_{c_{in}} \tag{4.17}$$
$$H_{1_{U,5}} : \mu_{c_{out}} > \mu_{c_{in}} \tag{4.18}$$

We find that under this hypothesis, less than 2% of tests turn out to be significant.

If we reverse the hypothesis and test whether the attack accuracy is significantly higher for inliers, i.e.

$$H_{0_{U,6}} : \mu_{c_{out}} \geq \mu_{c_{in}} \tag{4.19}$$
$$H_{1_{U,6}} : \mu_{c_{out}} < \mu_{c_{in}}, \tag{4.20}$$

over 32% of tests return p-values smaller than the significance level of 0.05. Figure 4.20 shows the p-value distribution per synthesizer and outlier detection algorithm. Our results imply that for the distance-based approach, **outliers seem to be harder to predict membership for**. The average accuracies over all outlier detection algorithms for each synthesizer are visualized in Figure 4.21. We can see that inliers (orange bars) are more likely to have their membership predicted correctly in all cases. Overall, the maximum accuracy for outliers is 0.55, which is for outliers detected with the quantile method and Bayesian Network as a synthesizer. The accuracies for LOF outliers with Bayesian Networks are all around 0.4, iForest has an accuracy of 0.45. Especially for data generated by a TVAE, the difference in accuracy between in- and outliers is substantial. With accuracy values between 0.05 and 0.2, there is no real risk for the detected outliers. The values for CTGAN range from 0.15 to 0.25, which is also way too low to make reliable predictions.

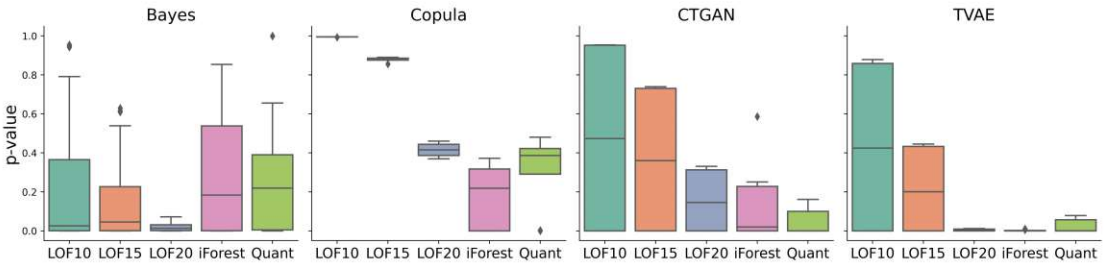

Figure 4.20: P-value-distribution for correctly predicted records: the box-plots visualize the distribution of p-values for the hypothesis that the accuracy μ_c is smaller for outlying than for inlying records (see Equations (4.19) and (4.20)).

We now look at the accuracies for out- and inliers for non-members only. Contrary to the members only (Figure 4.21, the accuracies for outliers are higher than for inliers. The only exception for this is the Copula with LOF10, where the outliers' is minimally smaller than the inliers', by 0.002. All other synthesizer-algorithm combinations show higher

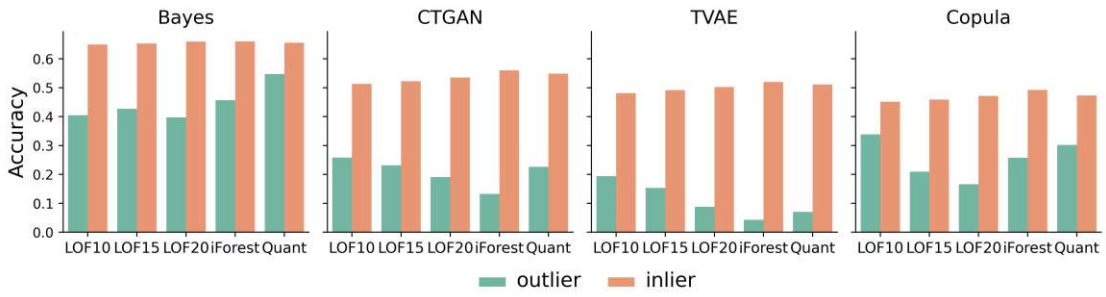

Figure 4.21: Inlier vs. Outlier Accuracy for distance approach for members. For each data set and synthesizer, we visualize the average accuracy for inliers and outliers.

accuracies, mostly with more than 0.1 increase, for outliers than inliers. This shows that outliers that were not used in the training data are more likely to be correctly inferred as non-members. Additionally, we conclude that inliers not included in the training data are more likely to be incorrectly labeled as members. This conclusion can be drawn from the low accuracy values in Figure 4.21. With this, and the results shown in Figure 4.21, we now conclude that the distance approach is more likely to label inliers as members.

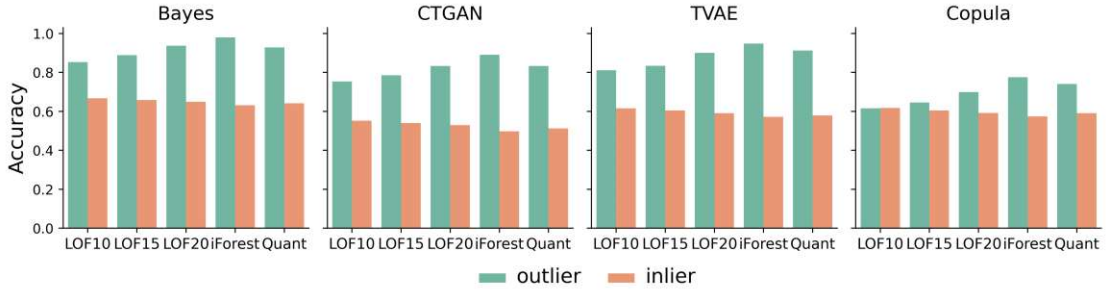

Figure 4.22: Inlier vs. Outlier Accuracy for distance approach for non-members only. For each data set and synthesizer, we visualize the average accuracy for inliers and outliers.

As we did for the shadow model approach in Section 4.1.2, we again test how well **outlier detection algorithms predict the records at risk** by computing precision and recall from the variable *at_risk* (ground truth) and detected outliers (prediction). In Figure 4.23, we see that outlier detection algorithms are not able to identify the records at risk. With the majority of recall values below 0.1, only a small number of records at risk were identified as outliers. Precision values are even lower, meaning that only a few detected outliers are also found to be at risk. For the TVAE, no outliers were predicted correctly. The recall values for Copula are highest, with values from 0.03 to 0.32. The precision values, however, are all at or below 0.062. For the Copula and LOF10, 32% of correctly predicted records were outliers. However, only 2% percent of the outliers could be predicted correctly. The outcome for Bayes and CTGAN is similar, with both recall and precision values ranging between 0.004 and 0.1.

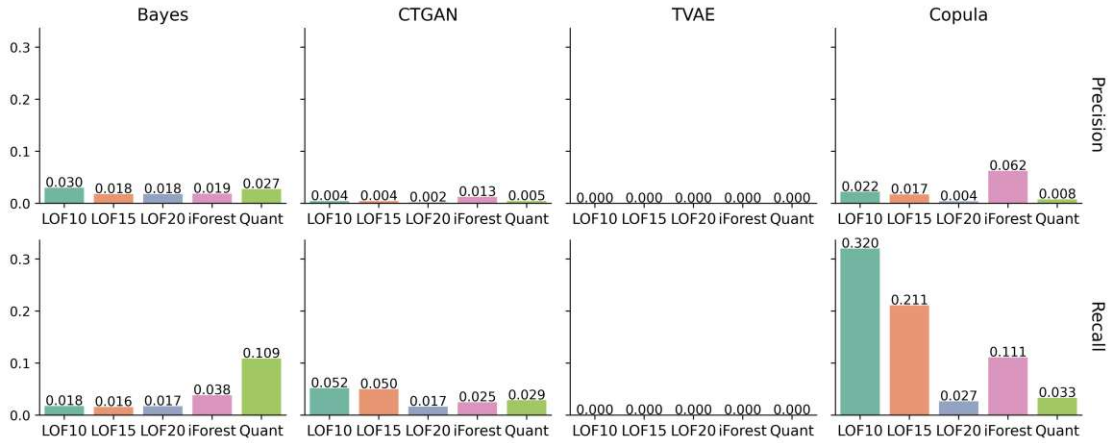

Figure 4.23: Relation between outliers and records at risk for the distance approach: the plot shows how well outlier detection algorithms can predict records at risk.

With the above-presented results in Figure 4.21 and Figure 4.23, we conclude that for the distance approach, there is no greater risk for outliers than for inliers. Outlier detection algorithms are also not suitable to predict the records at risk for this attack – inliers could be predicted with higher accuracy (see Figure 4.21). Recalling that this approach labels target records that are close to synthetic records as members, this conclusion makes sense. The distance from an inlier to its closest synthetic record is most likely smaller than this distance for an outlier.

4.2.3 Defense Evaluation

For the defense we follow the same process as for the shadow model approach: we remove the records at risk from the data the synthesizers train on. We then repeat the attack and reevaluate the MIA. Figure 4.24 shows the attack accuracy before and after the defense. Turquoise bars below the dashed grey lines illustrate a smaller attack accuracy after the defense. This is the desirable outcome. Bars larger than the dashed grey line represent the cases where the attack defense did not work and made the data even more vulnerable. Again, all values for the Thyroid data generated with a Copula are zero, since the attack could not label any members correctly, hence there were no records at risk found nor removed from the original training data. We observe a slight decrease for most attacks. Some attacks, however, seem to be even more accurate after the defense, causing the opposite of the intended result, e.g. Caesarian data with Copula. There is no apparent trend on how the cutoff value α influences the defense outcome. We observe that for synthetic data created with Bayesian Networks, the defense only causes an average decrease in attack performance of 0.01. The defense works best with Thyroid data and TVAE as well as the Breast Cancer data with CTGAN, where there is an average decrease in accuracy of 0.12 and 0.1 respectively.

Table 4.9 shows the exact numbers displayed in Figure 4.24. Scenarios where the defense caused an increase in accuracy, and therefore made the data more vulnerable, are highlighted in red. It seems that for smaller data sets (Caesarian and Heart), as well as for data using the TVAE as a synthesizer, the defense does not achieve the desired outcome of decreasing the MIA performance. To further inspect this effect, we analyze the attack accuracies separately for records at risk only (Figure 4.25) and accuracies for records labeled not at risk (Figure 4.26). The two plots show that although the defense works well for the records at risk, as they are less likely to be predicted correctly, it can also slightly increase the risk for the remaining records in the original data.

To further explore the effect of the defense making the remaining records more vulnerable, we compare the number of records that were incorrectly labeled before and correctly labeled after the defense. The results can be seen in Figure 4.27. Recall that for this approach, we use the risk score quantiles of 5, 10, 15, and 20 percent to define the records at risk. Therefore, between 5% and 20% of the records are at risk. These values are visualized by the gray dashed lines. We can see that there are up to 14% new records at risk resulting from our defense. While smaller data sets, i.e. Caesarian and Heart, are very affected with 6.7% and 7.7% new records at risk respectively, the Breast Cancer and Thyroid data seem to be less affected by this, with 1.8% and 4.7% new records at risk. Generally, the parameter α, i.e. the number of records removed, does not influence too much how many new records at risk are a result of the defense. On average, for $\alpha = 5$, there are 4.5% new records at risk. For $\alpha = 10, 15, 20$, there are 4.9, 5.1 and 5.2% new records at risk respectively. We conclude that there is a slight upward trend, where for larger α, the defense causes more vulnerable records.

The risk score distributions are shown in Figure 4.28. The distributions before and

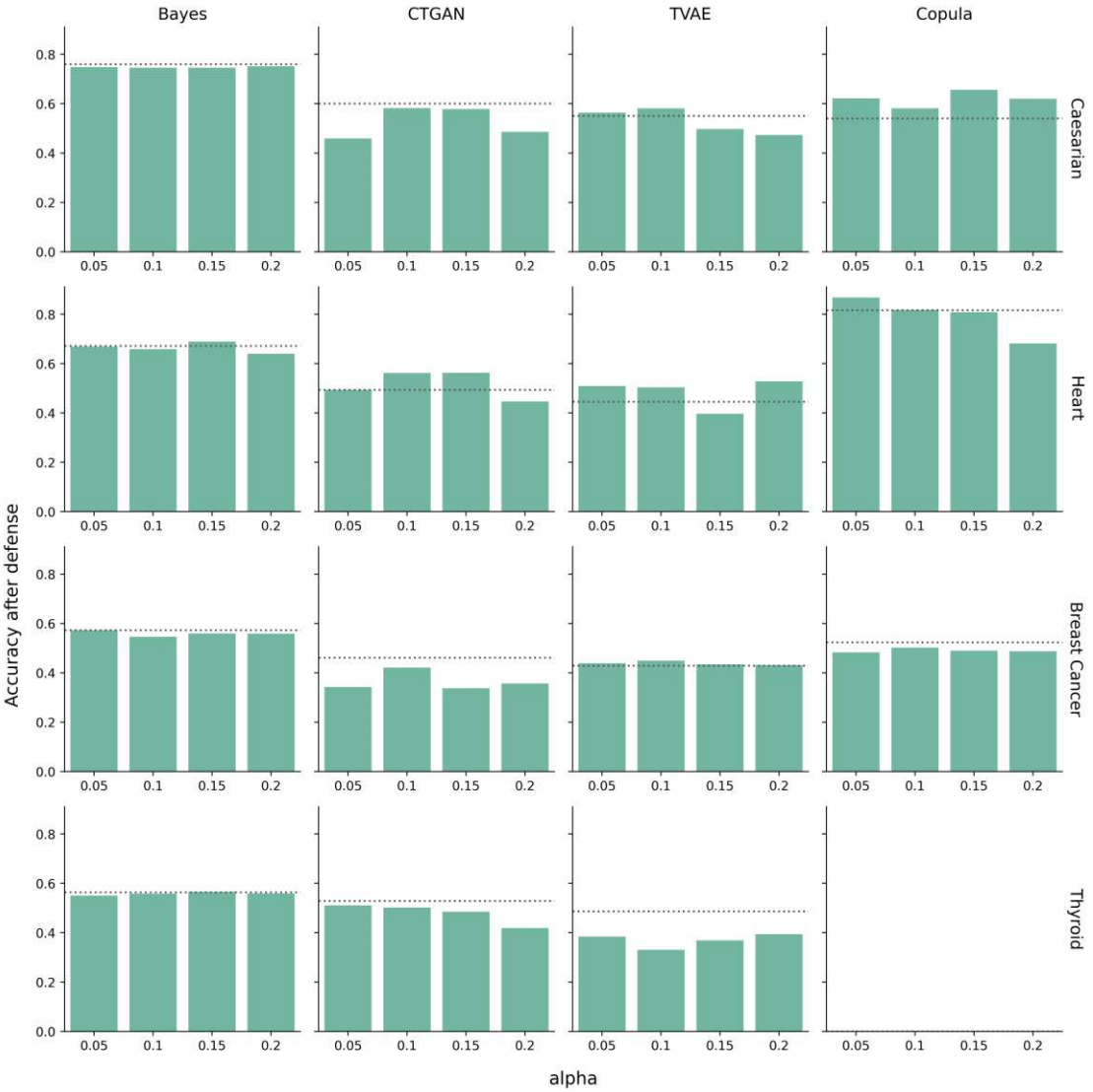

Figure 4.24: MIA Accuracy before and after defense: the bars show the attack accuracy after the defense for different alpha values. We use gray lines to indicate the attack accuracy before the defense.

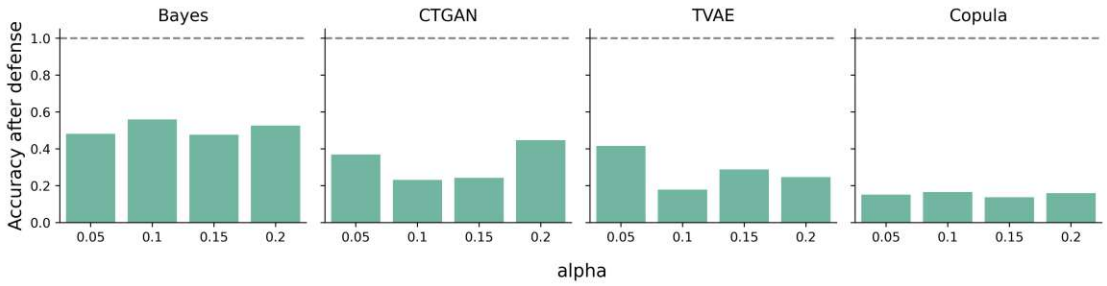

Figure 4.25: MIA Accuracy before and after defense for Records at risk: the gray line indicates the accuracy before the defense. This value is always one for records at risk.

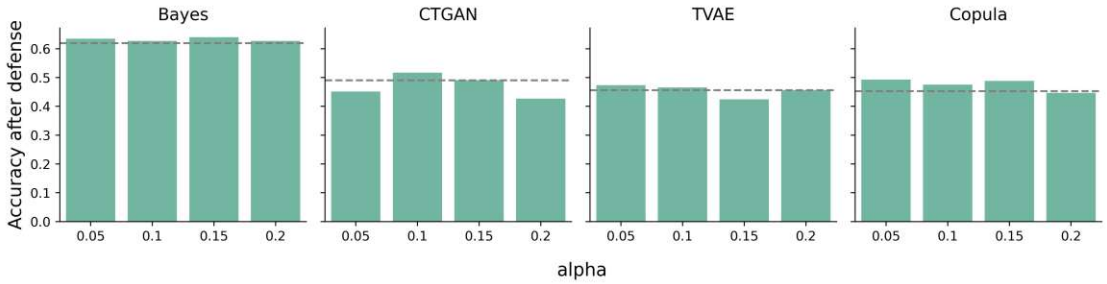

Figure 4.26: MIA Accuracy before and after defense for records not at risk: the accuracy before the defense is shown by the gray lines.

Table 4.9: Defense success measured by MIA accuracy (distance approach)

		Breast Cancer	**Caesarian**	**Heart**	**Thyroid**
	α				
Bayes	0.05	0,00	-0,01	-0,00	-0,01
	0.1	-0,03	-0,01	-0,01	-0,01
	0.15	-0,01	-0,01	0,02	0,00
	0.2	-0,01	-0,01	-0,03	-0,01
CTGAN	0.05	-0,12	-0,14	0,00	-0,02
	0.1	-0,04	-0,02	0,07	-0,03
	0.15	-0,12	-0,02	0,07	-0,04
	0.2	-0,10	-0,12	-0,05	-0,11
TVAE	0.05	0,01	0,01	0,06	-0,10
	0.1	0,02	0,03	0,06	-0,18
	0.15	0,01	-0,05	-0,05	-0,20
	0.2	0,00	-0,08	0,08	0,00
Copula	0.05	-0,04	0,08	0,05	0,00
	0.1	-0,02	0,04	0,00	0,00
	0.15	-0,03	0,12	-0,01	0,00
	0.2	-0,04	0,08	-0,13	0,00

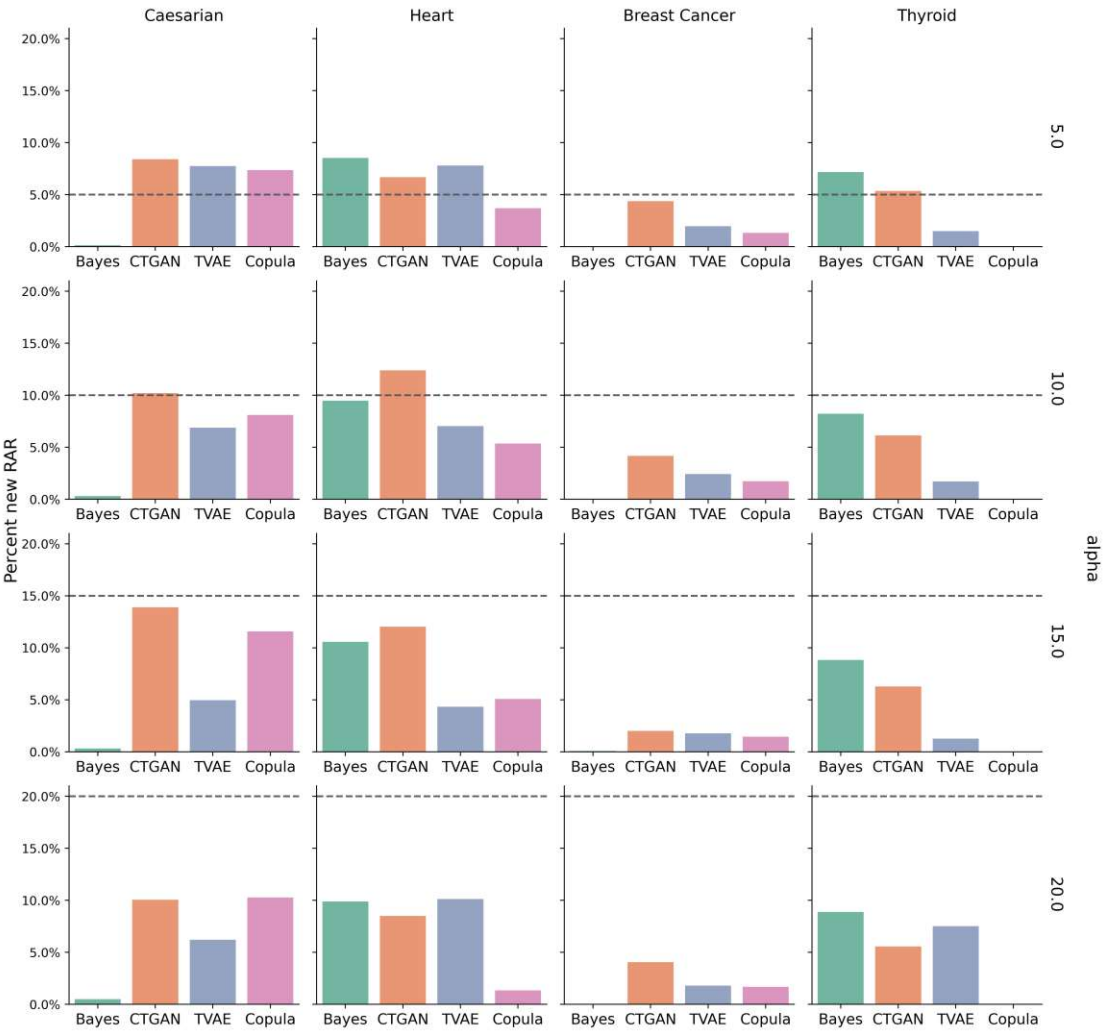

Figure 4.27: Percent of new records at risk caused by defense for distance approach

after the defense are very similar. Small differences can be seen for the Caesarian data, where the difference when using Copula is largest. The risk scores before and after the attack are highly similar for the Breast Cancer and Heart data and show only minor distributional differences. For the Thyroid data, the risk scores, both before and after the defense, tend to accumulate around 0.2 for the Copula and around 0 for the rest of the synthesizers.

Lastly, we analyze the risk scores for new outliers for the remaining data. We therefore repeat the outlier detection on the remaining data only and compare the risk score distributions for out- and inliers. The results can be seen in Figure 4.29. We find that for synthetic data generated with Bayesian Networks and CTGAN, the outliers obtained higher risk scores for the attacks. For the Copula this is not the case, as risk scores

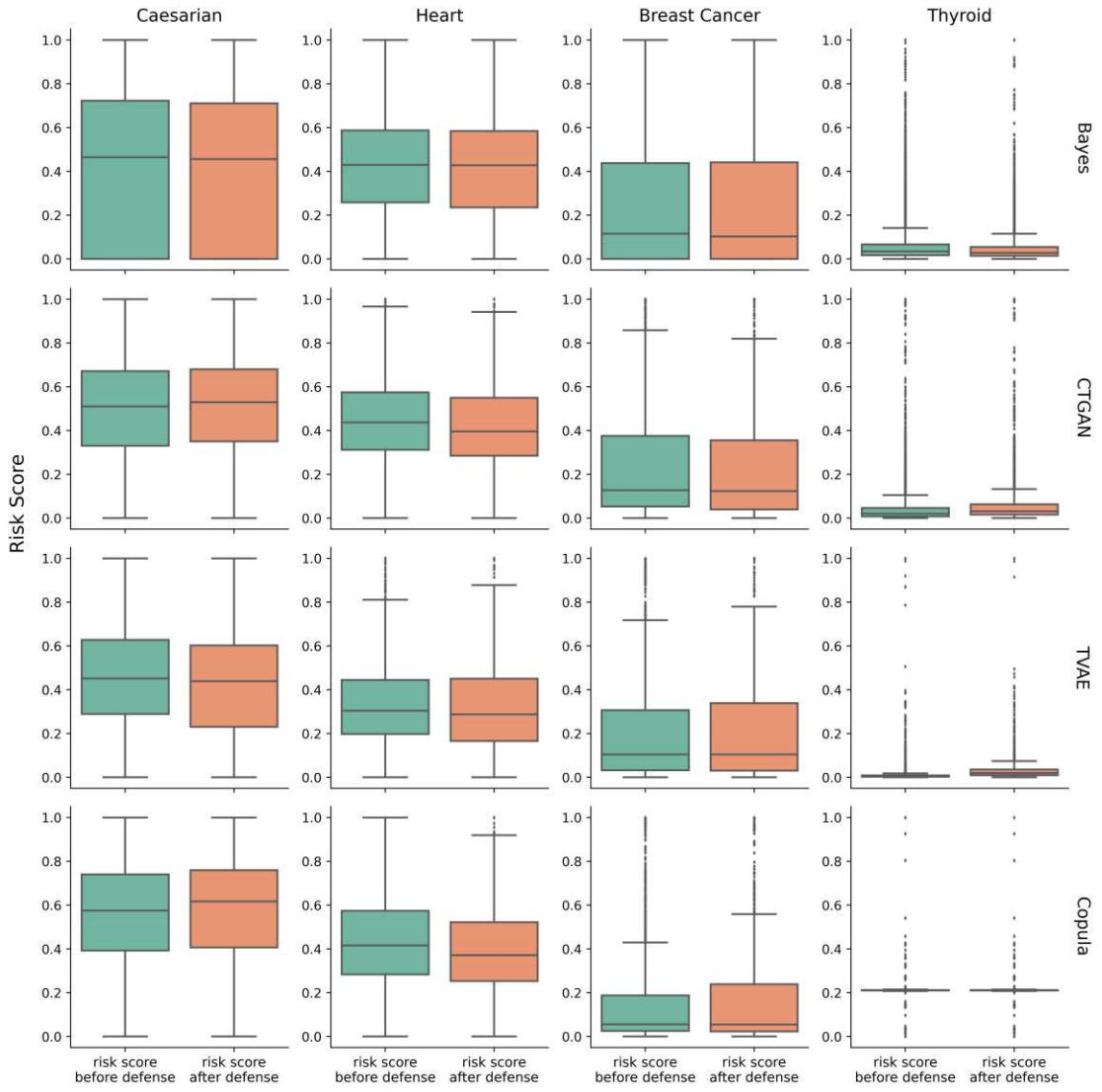

Figure 4.28: Distribution of risk scores before and after defense for the for the distance approach

are similarly distributed for in- and outliers. For the TVAE, the outliers detected using LOF10 and LOF15 obtain higher risk scores than inliers.

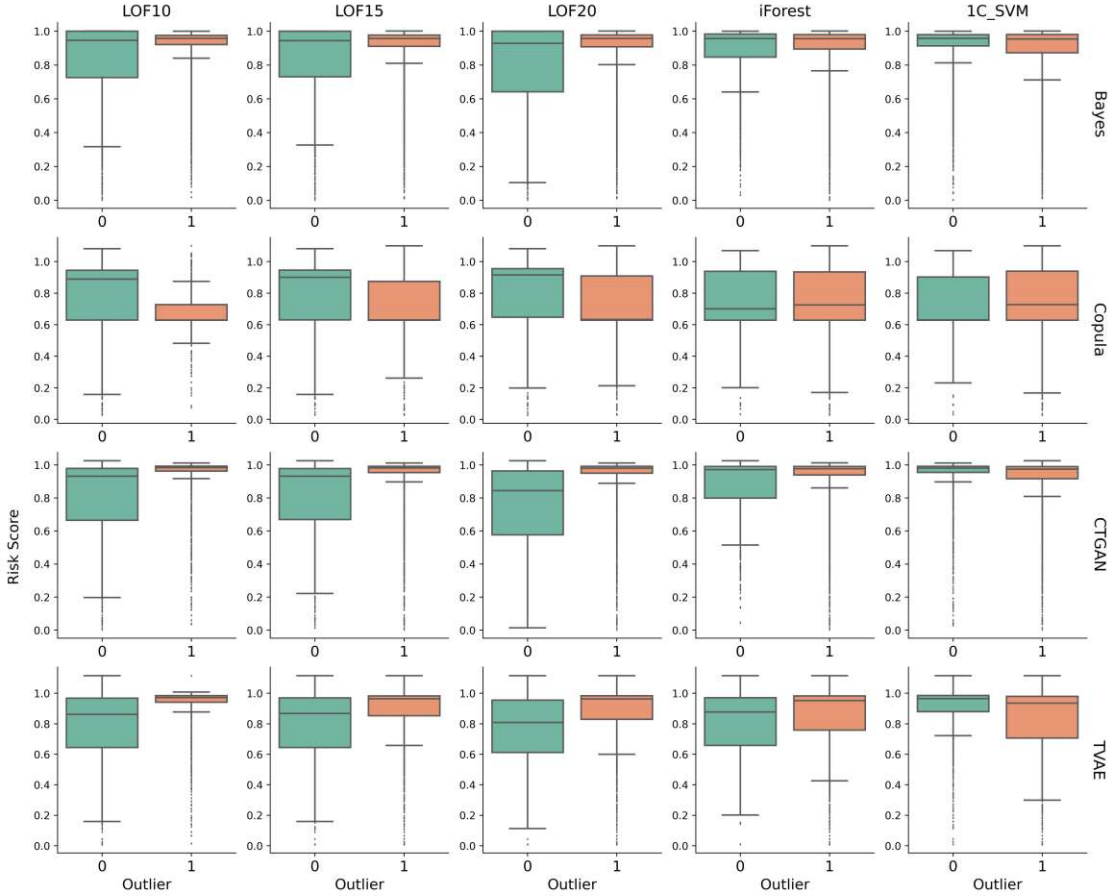

Figure 4.29: Distribution of risk scores for new outliers

4.2.4 Utility Assessment

For the utility assessment, we follow the same procedure as in Section 4.1.4. Figure 4.30 shows the difference in accuracies by the synthesizer. We observe that although there is a difference between synthetic and original data, the difference between the two synthetic data sets seems to be minimal. The Thyroid data, which is the largest data set used in our experiments, suffers substantial utility loss. However, only for the TVAE-generated data, the synthetic data generated without RAR further drops significantly by 0.26. For the TVAE the largest decrease in accuracy for the synthetic data where all records were used during training, ranges between 0.07 and 0.13. This is also within the range of what the authors of [82] found. Furthermore, the synthetic data generated by CTGAN suffers a higher utility loss, with a decrease in accuracy ranging from 0.06 to 0.44. This is the highest decrease in accuracy over all synthesizers. Previous work has already shown that the CTGAN can exhibit substantial utility disadvantages [81]. The accuracy for Copula drops between 0.1 and 0.35 for the synthetic data before the defense. This is also within the range presented by [81]. The accuracy for Copula-generated data after the defense changes only minimally from the synthetic data before the defense, with differences of accuracy between zero and 0.03. For the Bayesian Networks, the accuracy drops by 0.16 at most (Thyroid data). For the Caesarian data, the accuracy stays the same for the original, the synthetic, and the synthetic data without records at risk.

Still, the overall differences between synthetic data with and without RAR are minimal.

Next, we look at the effect of parameter α on the utility. As α describes the cutoff for risk scores, a higher α results in more records being removed from the original data. Naturally, one would expect less utility for a higher α, i.e. smaller training data. However, when looking at the experiment results in Figure 4.31, we discover that this is not necessarily the case. Generally, the utility does not seem to be affected much by the α value. One possible explanation for this is that the defense removes data records that are similar to other data points and therefore makes the data more general and keeps the synthesizer from overfitting.

Although the synthetic data resulting from the defense seems to not cause more overall utility loss than synthesizing data from the entire training data as seen in Figure 4.30, we now want to look at the per-class-utility for the Thyroid data. We again look at this data set, since its target variable is highly imbalanced. Class 3 is the majority class with around 93% of records belonging to this class. Class 1 and 2 are the minority classes (2% and 5% respectively). Figure 4.32 shows the accuracy, precision, and recall per class, and highlights the extent of the utility loss for the minority classes: While the evaluation metrics for class 3 are always around the same for the synthetic data sets and the original, the values for the synthetic data sets are significantly smaller for the minority classes. However, with the exception of TVAE, the synthetic data generated using the entire training data and the synthetic data resulting from the defense still have comparable evaluation measures.

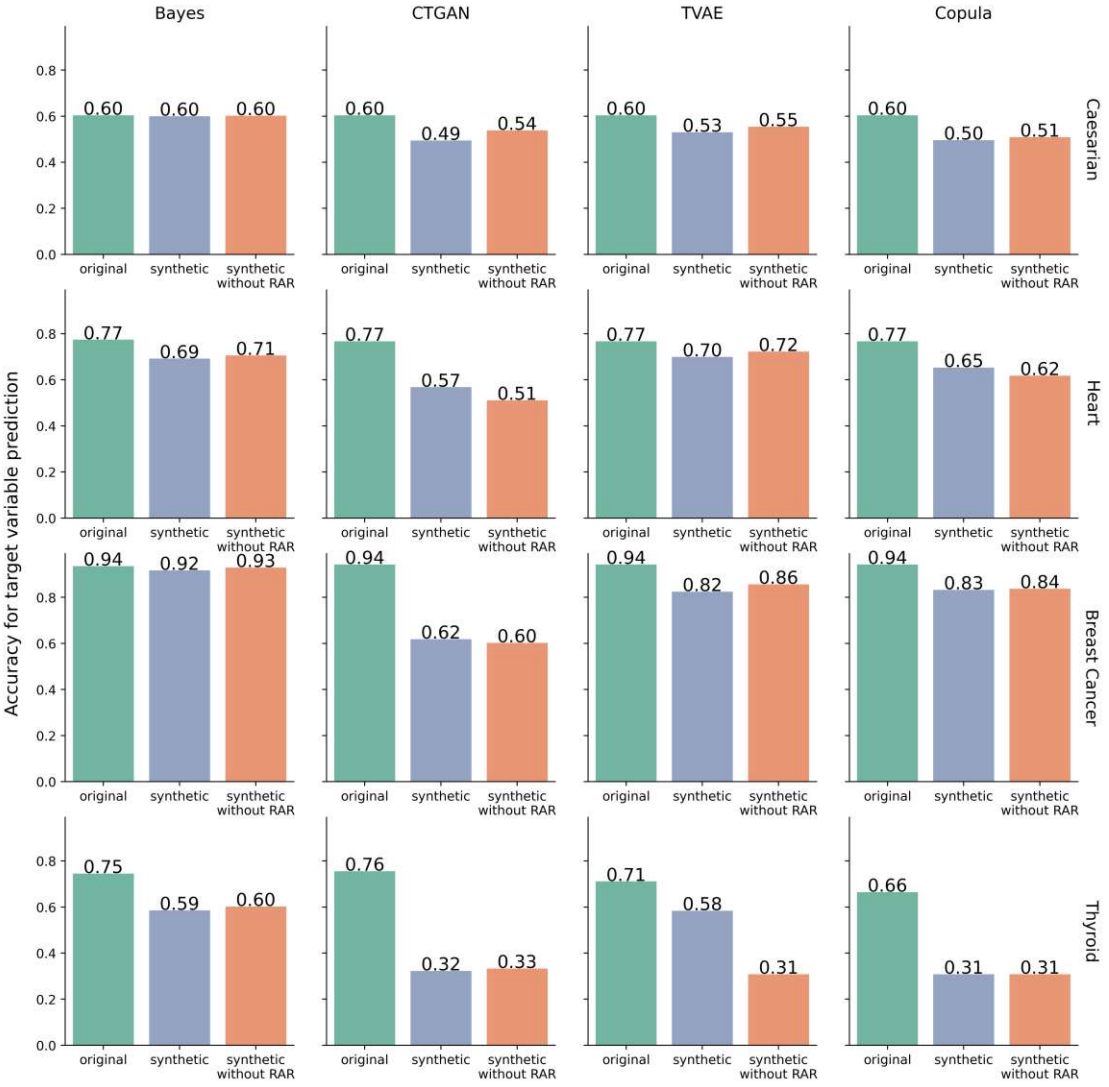

Figure 4.30: Utility comparison by synthesizer: the utility difference between the original and the two synthetic data sets is visualized using the prediction accuracy.

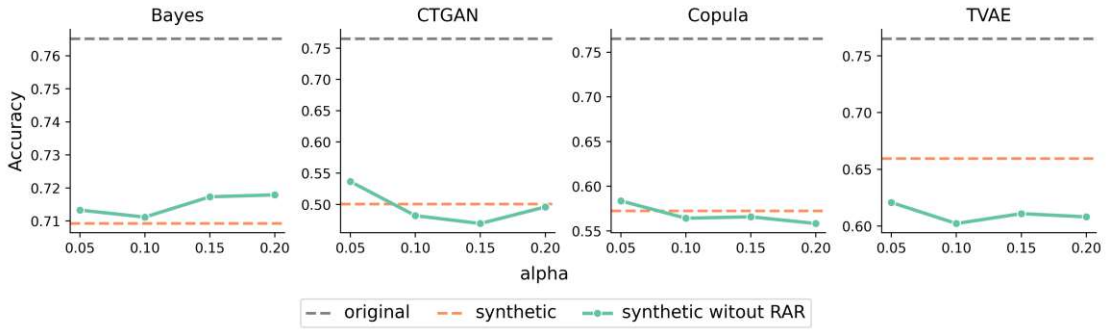

Figure 4.31: Utility comparison for different α values: the plot visualizes the difference in utility for synthetic and original data sets for every α value. The gray line shows the average accuracy for the original data, the orange line represents the synthetic data from a synthesizer trained on the entire original data and the turquoise line displays the results for synthetic data where records at risk were excluded from the training data.

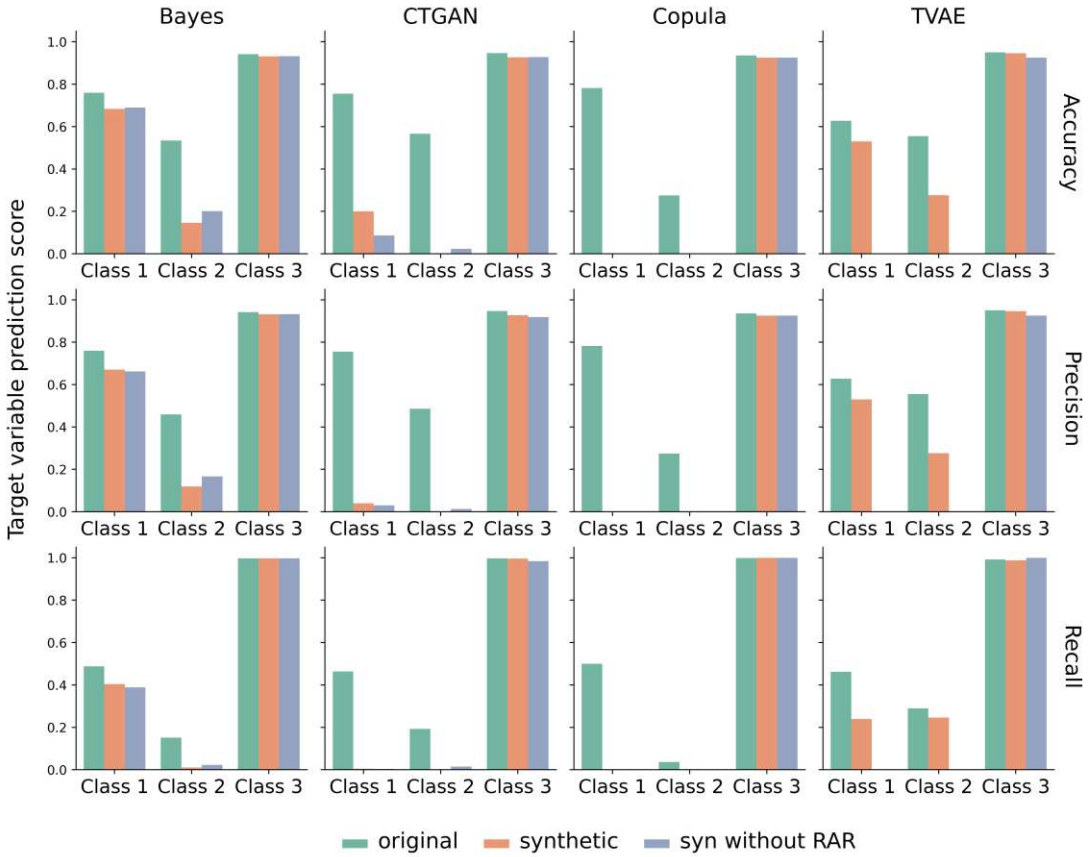

Figure 4.32: Utility per class for Thyroid data: accuracy, precision, and recall are displayed per class. Classes 1 and 2 are minority classes. Over 90% of records belong to class 3.

4.3 Comparison of the two approaches

Lastly, we want to show the overall attack performance by visualizing the accuracy values for the shadow model and distance approach side by side in Figure 4.33. These values are the aggregated results from Figure 4.16 and Figure 4.1. For the distance approach, for Bayesian Networks, we only consider the results with a maximum of three parents per node, as this is also the maximum number of parents used for all Bayesian Networks in the shadow model approach. Although the attack accuracies are quite similar for both approaches, the distance approach achieves higher scores in most cases, the only exception being Copula with Breast Cancer data, where the shadow model approach is 0.03 higher than the distance approach, as well as Copula and CTGAN with the Thyroid data, where the shadow model approach is 0.07 and 0.005 higher than the distance approach respectively. Especially for Bayesian Networks, the distance approach seems to be able to infer membership a lot more accurately than the shadow model approach. On average, the accuracy of the distance approach obtains accuracies larger by 0.07 than the shadow model approach. Generally, the smaller data sets, Caesarian and Heart, are seemingly more prone to MIA.

Figure 4.33 again highlights that the attack does not pose an overall privacy threat to the dataset as a whole, with accuracies between 0.4 and 0.6, which is close to the results one would obtain with random guessing. The only exception here is the distance approach with Bayesian Networks, where the accuracies range from 0.63 to 0.75. However, as we found in our analysis, some records are easier to predict membership on, and therefore more at risk for such attacks. But, as these vulnerable records do not seem to follow a certain trend, e.g. it is not only outliers, an adversary can never know if a target record belongs to the group of records at risk. Therefore, the adversary can still not know if their membership prediction is reliable.

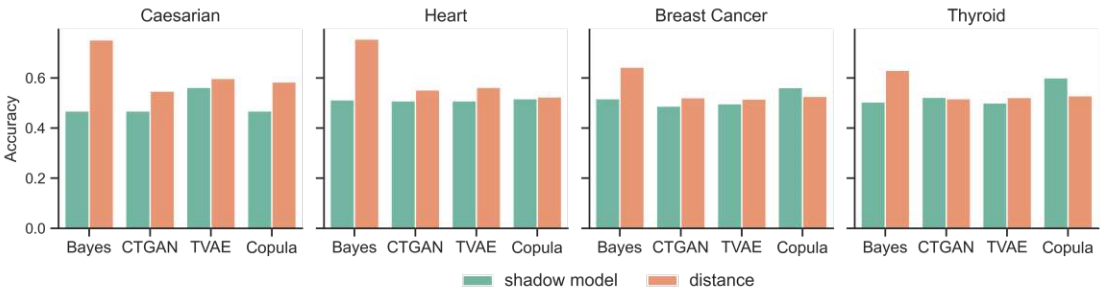

Figure 4.33: Comparison of the two approaches with MIA accuracy

In Figure 4.34 we compute the accuracies for the two approaches for members and non-members separately. The top row shows the shadow model approach, and the bottom row displays the results for the distance-based approach. The distance-based approach predicts non-members more accurately than members in most scenarios. For the shadow model approach, on the other hand, the opposite holds: For most scenarios, the members are predicted more accurately than non-members. The reason for this could be that an

influential record will have a greater effect on the generated data if it is included in the training data, than when it is not.

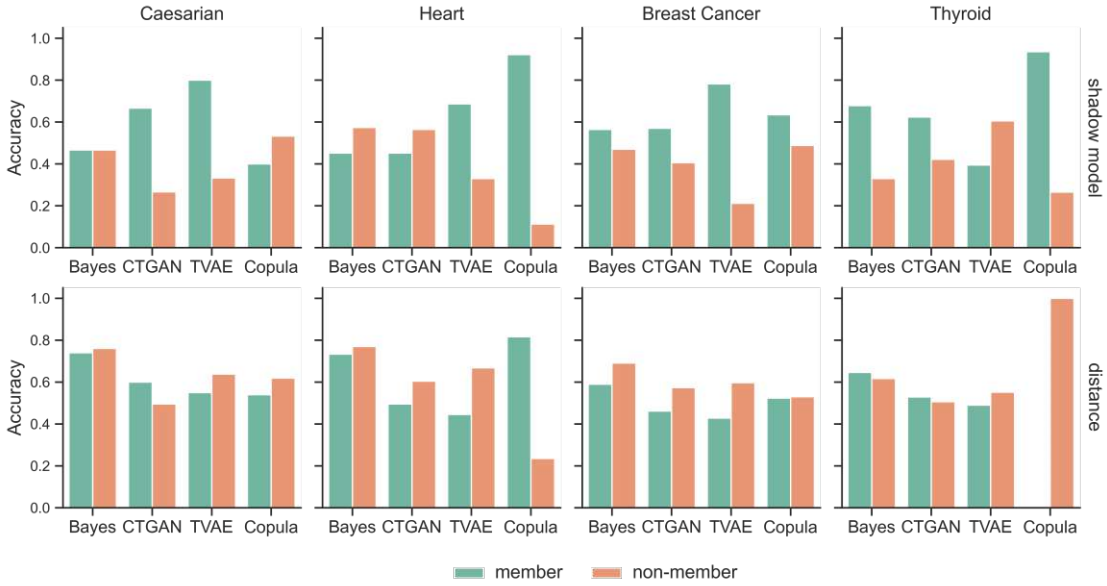

Figure 4.34: Comparison of the two approaches with MIA accuracy by membership

Lastly, we look at the risk record identification for both approaches. In Table 4.10 we list the percent of records that were labeled at risk and not at risk by both approaches. In the following, we will call this percentage the agreement rate. We find that the agreement rate of records labeled not at risk is substantially larger than the one for records at risk. This makes sense, as a lot more records are labeled not at risk than at risk. The highest agreement rate occurs for the Breast Cancer data with Bayesian Networks, with 84.5%. Overall, mostly the Heart and Breast Cancer data show the highest agreement rates for records not at risk. For the Copula, however, the Thyroid data has an agreement rate of almost 10% for records at risk. Meanwhile, the agreement rate for records at risk is very low for the Caesarian data, where only the CTGAN obtains a rate that is not zero, of 4%. All other rates for records at risk are below 4.5%.

This analysis indicates that the records' risks vary across all attack approaches and need to be considered for risk identification. We want to highlight that even when considering most or all possible attack approaches for risk identification, we can never be certain about possible attacks being developed in the future. A record that might not be at risk for one attack might not be with another.

4.4 Summary

This chapter contained the experimental analysis and illustrations of the obtained results. We presented the overall attack performances, risk identification, defense evaluation,

Table 4.10: Agreement rates in percent for records at risk and records not at risk

		records at risk	records not at risk
Bayes	Caesarian	0.0	80.8
	Heart	2.5	73.0
	Breast Cancer	0.4	84.5
	Thyroid	0.5	43.7
CTGAN	Caesarian	4.0	56.6
	Heart	1.2	61.8
	Breast Cancer	2.5	73.4
	Thyroid	1.1	82.7
TVAE	Caesarian	0.0	72.7
	Heart	4.3	61.8
	Breast Cancer	0.8	57.2
	Thyroid	0.5	73.1
Copula	Caesarian	0.0	72.7
	Heart	2.5	76.0
	Breast Cancer	2.7	35.9
	Thyroid	9.6	59.7

and utility assessment for the shadow model and distance-based approach. We found that the risk for each record depends on the attack approach and cannot be generalized. Furthermore, we found that the defense has a positive effect on the records at risk and does not negatively affect the data utility of balanced data. For unbalanced data, the defense can pose a problem to the minority classes' utility. We finished our analysis with a comparison of the two attack approaches and highlighted the differences in their results.

CHAPTER 5

Conclusion

In the following, we present the main contributions of this work, as well as a summary of the results obtained, by answering the research questions defined in Chapter 1.

5.1 Contributions

In our work we implement two approaches, shadow modeling and distance-based, to conduct membership inference attacks on synthetic tabular data. We use four different synthesizers (Bayesian Networks, CTGAN, TVAE, and Copula), to generate the synthetic data. We analyze the overall power of the two attack approaches by computing evaluation scores and comparing them.

[8] and [16] both claim that outliers are most at risk for MIA. They test their hypothesis on five and ten outliers respectively, using the shadow model approach. In this thesis, we consider all records in the training set and compute the risk by evaluating how accurately each record's membership can be inferred. In addition to the shadow model attack, we also do this for the distance-based attack by [9].

As the work of Stadler et al. [8] showed, membership inference attacks with a shadow model approach are no threat to the entirety data records, but rather affect single records. This is also what our experiments show. However, they conclude that outliers are more at risk than inliers. We we detected outliers using detection methods such as the Local Outlier Factor, Isolation Forests, or through finding records with rare values, and analyze the relationship between these outliers and the records at risk for membership inference attacks, to find out if outliers are significantly more vulnerable to the attacks. We find that there is no significant correlation between outliers and records at risk that can be generalized for all data sets and synthesizers. Nonetheless, there are settings, e.g. Heart data and Copula with outliers detected using LOF and iForest, where the outliers' accuracies for the attack were significantly larger than the inliers.

The distance-based approach for tabular data by Hayeong et al. [9] assumes that members have smaller distances to their closest synthetic data record than non-members. We test this assumption and conclude that, albeit not always statistically significant, this assumption holds. In contrast to their work, we design the attack as a no-box attack rather than black- and white-box attacks. Although our attacks using CTGAN and TVAE obtain slightly lower overall ROC-AUC scores, this is to be expected due to our restrictive model assumptions. We find that the no-box attack only seems to be a serious threat to small data sets and for the Bayesian Network synthesizer. We extend the work of [9] and design a method for finding the decision threshold necessary for conducting the attack. With this, we conclude that the no-box, distance-based attack only poses a possible threat to the data generated using Bayesian Networks, especially when training on small data sets. We further extend the work of [9] by identifying the records at risk, and, like for the shadow model approach, looking into the relationship between these records at risk and outliers. For this distance approach, we find that, unlike recent work suggests [8, 16], inliers can be predicted with much higher accuracy than outliers. This seems to be especially true when using CTGAN or TVAE as a synthesizer.

As a simple baseline defense against membership inference attacks, we propose to remove the records at risk from the training data and evaluate the defense's influence on the removed and remaining data records by recalculating each record's risk. Furthermore, we assess the data utility of the synthetic data before and after the defense as well as the original data and compare them. We lay special focus on per-class utility for imbalanced data in our evaluation.

5.2 Summary

We summarize our results and findings in regard to the Research Questions defined in Section 1.3:

1. **To what extent can we predict records at risk for membership inference attacks by detecting outliers with algorithms like Local Outlier Factor and Isolation Forest?**
   We find that the results are highly dependent on the MIA approach: While outliers can be predicted more accurately with the shadow model attack, the opposite holds for the distance-based attack. For the shadow model approach, up to 40% of correctly inferred members are outliers. Additionally, 24% of outlying members are labeled as members by the shadow model approach. Although the attack accuracy is mostly higher for outliers, the statistical tests show that this is not significant in most cases. Attacks using the distance approach show that outliers are, contrary to recent assumptions, less likely to be inferred correctly than inliers.

   a) **By removing these records from the original training data, to what extent does the overall success of the MIA on the records at risk suffer? To what degree do the membership predictions change for**

the removed records?
By removing the records at risk as a defense measure, we show that although the overall attack accuracy mostly decreases with the defense, an increase in the overall attack accuracy is also possible. We were able to show that consequently to the defense an attacker is less likely to predict them correctly, which makes the attack model more unreliable.

b) **To what extent are the remaining data records affected by removing the records at risk?**
For the shadow model approach, we observe that the risk for the remaining records showed, on average, an increase of 0.05 attack accuracy. This however seems highly dependent on the data set and synthesizer. When using the distance-based attack, the attack accuracy increases by 0.015 on average. Although this is a larger increase, the results are more stable across synthesizers and data sets, compared to the shadow model results.

2. **To what extent does the data utility suffer when synthesis models learn from the original data excluding the records at risk?**
We find that although there is a noticeable decrease in utility for synthetic data compared to the original data, the utility of synthetic data generated from a synthesizer that was only trained on records not at risk, and a synthesizer that was trained on the entire training data, are mostly the same for well-balanced data sets. Because of that, there is no apparent disadvantage to using the defense with balanced data sets.

   a) **By how much does the utility for synthetic data learned from imbalanced data sets decrease compared to synthetic data generated from a balanced data set?**
   The defense does not seem to affect the overall accuracy of imbalanced data sets, as it is around the same as the utility of synthetic data before the defense. However, this is caused by the high accuracy on the majority class.

   b) **To what extent does the utility on the minority class of the imbalanced data suffer compared to the majority class?**
   While the majority classes suffer no to minimal utility loss, the minority classes can suffer substantial utility loss. This, however, again concerns both the synthetic data with and without records at risk. The only exception is the TVAE which shows substantially smaller recall and precision values for the data excluding records at risk compared to the synthetic data generated from the entire training set.

3. **Which data synthesizing models generate synthetic data that is more vulnerable to MIA?**
For the shadow model attack, TVAE and Copula turn out to be the most vulnerable to MIA. When using the distance-based attack, Bayesian Networks generate data that is a lot more susceptible to the attacks. With an increasing number of maximum

parents for Bayesian Networks, the attack accuracy increases as well. Bayesian Networks are, however, also the synthesizers that create synthetic data with the least utility loss. This applies to both, the synthetic data before and after the defense.

Summarizing our contributions, we have shown that different attack approaches identify different records at risk. While outliers are more likely to be predicted for the shadow model approach, the opposite holds for the distance approach. As data owners do not know which attack approach an adversary might use, it is nearly impossible to identify all records at risk in one data set. We find that the attacks do not perform well overall, and only certain records are at higher risk. We therefore do not see a serious privacy risk, as an adversary has no way to find out if their target record is at higher risk and therefore able to be inferred with high confidence. Furthermore, we find that the attack approaches perform differently on the various synthesizers. We discover that although removing the records at risk from the data decreases the risk for these records, it will, in most cases, create a new layer of records at risk in the remaining data.

5.3 Future Work

In this thesis, we have identified records at risk for membership inference attacks on synthesizers, and their relation to outliers and analyzed the efficiency of removing vulnerable records as a defense measure. However, several aspects for future research that can extend our understanding of membership inference attacks against synthetic data remain:

- As more MIA approaches on synthesizers are presented, they can be used for identifying records at risk and their underlying patterns.
- We design the MIA as a no-box attack. Risk identification and defense success have yet to be studied on white- and black-box attacks. Within these settings, an adversary would have access to the synthesizer and would therefore be able to collect an arbitrary amount of synthetic records. With this, the influence of the synthetic data set's size on the attack success can be studied.
- As our attack assumptions are quite strict, e.g. having access to a reference data set and knowledge about the synthesizer used, the risk identification and attack outcome in a setting with more relaxed assumptions has yet to be explored.
- Additionally, future work could include attack evaluation and risk identification on differentially private synthetic data to compare the overall attack performance to non-differential private data, plus the comparison of records at risk found.

List of Figures

List of Tables

Bibliography

[1] L. Sweeney, "Guaranteeing anonymity when sharing medical data, the datafly system.," in *Proceedings of the AMIA Annual Fall Symposium*, p. 51, American Medical Informatics Association, 1997.

[2] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific reports*, vol. 3, no. 1, pp. 1–5, 2013.

[3] C. Culnane, B. I. Rubinstein, and V. Teague, "Health data in an open world," *arXiv preprint arXiv:1712.05627*, 2017.

[4] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125, IEEE, 2008.

[5] Y.-A. De Montjoye, L. Radaelli, V. K. Singh, and A. S. Pentland, "Unique in the shopping mall: On the reidentifiability of credit card metadata," *Science*, vol. 347, no. 6221, pp. 536–539, 2015.

[6] G. Cormode, C. M. Procopiuc, E. Shen, D. Srivastava, and T. Yu, "Empirical privacy and empirical utility of anonymized data," in *2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, pp. 77–82, IEEE, 2013.

[7] J. Brickell and V. Shmatikov, "The cost of privacy: destruction of data-mining utility in anonymized data publishing," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 70–78, 2008.

[8] T. Stadler, B. Oprisanu, and C. Troncoso, "Synthetic data–anonymisation groundhog day," in *31st USENIX Security Symposium (USENIX Security 22)*, pp. 1451–1468, 2022.

[9] J. Hyeong, J. Kim, N. Park, and S. Jajodia, "An empirical study on the membership inference attack against tabular data synthesis models," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 4064–4068, 2022.

[10] Z. Zhang, C. Yan, and B. A. Malin, "Membership inference attacks against synthetic health data," *Journal of biomedical informatics*, vol. 125, p. 103977, 2022.

[11] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18, IEEE, 2017.

[12] J. Hayes, L. Melis, G. Danezis, and E. De Cristofaro, "Logan: Membership inference attacks against generative models," *arXiv preprint arXiv:1705.07663*, 2017.

[13] D. Chen, N. Yu, Y. Zhang, and M. Fritz, "Gan-leaks: A taxonomy of membership inference attacks against generative models," in *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pp. 343–362, 2020.

[14] R. Yang, J. Ma, Y. Miao, and X. Ma, "Privacy-preserving generative framework against membership inference attacks," *arXiv preprint arXiv:2202.05469*, 2022.

[15] B. van Breugel, H. Sun, Z. Qian, and M. van der Schaar, "Membership inference attacks against synthetic data through overfitting detection," *arXiv preprint arXiv:2302.12580*, 2023.

[16] M. Meeus, F. Guepin, A.-M. Cretu, and Y.-A. de Montjoye, "Achilles' heels: Vulnerable record identification in synthetic data publishing," *arXiv preprint arXiv:2306.10308*, 2023.

[17] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104, 2000.

[18] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 eighth ieee international conference on data mining*, pp. 413–422, IEEE, 2008.

[19] M. Hittmeir, A. Ekelhart, and R. Mayer, "On the utility of synthetic data: An empirical evaluation on machine learning tasks," in *Proceedings of the 14th International Conference on Availability, Reliability and Security*, pp. 1–6, 2019.

[20] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering–a systematic literature review," *Information and software technology*, vol. 51, no. 1, pp. 7–15, 2009.

[21] D. Lambert, "Measures of disclosure risk and harm," *Journal of Official Statistics-Stockholm-*, vol. 9, pp. 313–313, 1993.

[22] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," *IEEE Security and Privacy*, 1998.

[23] N. R. Adam and J. C. Worthmann, "Security-control methods for statistical databases: a comparative study," *ACM Computing Surveys (CSUR)*, vol. 21, no. 4, pp. 515–556, 1989.

[24] L. H. Cox, "Suppression methodology and statistical disclosure control," *Journal of the American Statistical Association*, vol. 75, no. 370, pp. 377–385, 1980.

[25] T. Dalenius, "Finding a needle in a haystack or identifying anonymous census records," *Journal of official statistics*, vol. 2, no. 3, p. 329, 1986.

[26] L. Sweeney, "k-anonymity: A model for protecting privacy," *International journal of uncertainty, fuzziness and knowledge-based systems*, vol. 10, no. 05, pp. 557–570, 2002.

[27] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, pp. 3–es, 2007.

[28] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd international conference on data engineering*, pp. 106–115, IEEE, 2006.

[29] P. Ohm, "Broken promises of privacy: Responding to the surprising failure of anonymization," *UCLA l. Rev.*, vol. 57, p. 1701, 2009.

[30] C. Dwork, "Differential privacy," in *International colloquium on automata, languages, and programming*, pp. 1–12, Springer, 2006.

[31] A. Goncalves, P. Ray, B. Soper, J. Stevens, L. Coyle, and A. P. Sales, "Generation and evaluation of synthetic patient data," *BMC medical research methodology*, vol. 20, no. 1, pp. 1–40, 2020.

[32] B. Nowok, "Utility of synthetic microdata generated using tree-based methods," *UNECE Statistical Data Confidentiality Work Session*, 2015.

[33] K. El Emam, L. Mosquera, and J. Bass, "Evaluating identity disclosure risk in fully synthetic health data: model development and validation," *Journal of medical Internet research*, vol. 22, no. 11, p. e23139, 2020.

[34] M. Hittmeir, R. Mayer, and A. Ekelhart, "Efficient bayesian network construction for increased privacy on synthetic data," in *2022 IEEE International Conference on Big Data (Big Data)*, pp. 5721–5730, IEEE, 2022.

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

[36] R. J. Little *et al.*, "Statistical analysis of masked data," *JOURNAL OF OFFICIAL STATISTICS-STOCKHOLM-*, vol. 9, pp. 407–407, 1993.

[37] D. B. Rubin, *Multiple imputation for nonresponse in surveys*, vol. 81. John Wiley & Sons, 2004.

[38] S. E. Fienberg and R. J. Steele, "Disclosure limitation using perturbation and related methods for categorical data," *Journal of Official Statistics*, vol. 14, no. 4, p. 485, 1998.

[39] G. Caiola and J. P. Reiter, "Random forests for generating partially synthetic, categorical data.," *Trans. Data Priv.*, vol. 3, no. 1, pp. 27–42, 2010.

[40] J. P. Reiter, "Using cart to generate partially synthetic public use microdata," *Journal of official statistics*, vol. 21, no. 3, p. 441, 2005.

[41] S. Hawala, "Producing partially synthetic data to avoid disclosure," in *Proceedings of the Joint Statistical Meetings. Alexandria, VA: American Statistical Association*, 2008.

[42] J. Drechsler, S. Bender, and S. Rässler, "Comparing fully and partially synthetic data sets for statistical disclosure control in the german iab establishment panel: supporting paper für die work session on data confdentiality 2007 in manchester," *EUNECE/Programmes*, 2007.

[43] J. Drechsler and J. P. Reiter, "Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data," in *Privacy in Statistical Databases*, vol. 5262, pp. 227–238, Springer, 2008.

[44] J. Drechsler and J. Reiter, "Disclosure risk and data utility for partially synthetic data: An empirical study using the german iab establishment survey," *Journal of Official Statistics*, vol. 25, no. 4, p. 589, 2009.

[45] J. P. Reiter and R. Mitra, "Estimating risks of identification disclosure in partially synthetic data," *Journal of Privacy and Confidentiality*, vol. 1, no. 1, 2009.

[46] A. R. Benaim, R. Almog, Y. Gorelik, I. Hochberg, L. Nassar, T. Mashiach, M. Khamaisi, Y. Lurie, Z. S. Azzam, J. Khoury, *et al.*, "Analyzing medical research results based on synthetic data and their relation to real data results: systematic comparison from five observational studies," *JMIR medical informatics*, vol. 8, no. 2, p. e16492, 2020.

[47] F. K. Dankar, M. K. Ibrahim, and L. Ismail, "A multi-dimensional evaluation of synthetic data generators," *IEEE Access*, vol. 10, pp. 11147–11158, 2022.

[48] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, "Data synthesis based on generative adversarial networks," *arXiv preprint arXiv:1806.03384*, 2018.

[49] Z. Wan, Y. Zhang, and H. He, "Variational autoencoder based synthetic data generation for imbalanced learning," in *2017 IEEE symposium series on computational intelligence (SSCI)*, pp. 1–7, IEEE, 2017.

[50] N. Patki, R. Wedge, and K. Veeramachaneni, "The synthetic data vault," in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 399–410, IEEE, 2016.

[51] J. Young, P. Graham, and R. Penny, "Using bayesian networks to create synthetic data," *Journal of Official Statistics*, vol. 25, no. 4, p. 549, 2009.

[52] B. Nowok, G. M. Raab, and C. Dibben, "synthpop: Bespoke creation of synthetic data in r," *Journal of statistical software*, vol. 74, pp. 1–26, 2016.

[53] J. Geweke, "Bayesian inference in econometric models using monte carlo integration," *Econometrica: Journal of the Econometric Society*, pp. 1317–1339, 1989.

[54] C. Romano, "Applying copula function to risk management," in *Capitalia, Italy. http://www. icer. it/workshop/Romano. pdf*, 2002.

[55] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems* (Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, eds.), vol. 27, p. 2672–2680, Curran Associates, Inc., 2014.

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

[57] D. E. Rumelhart, G. E. Hinton, R. J. Williams, *et al.*, "Learning internal representations by error propagation," 1985.

[58] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[59] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282, 2018.

[60] K. Leino and M. Fredrikson, "Stolen memories: Leveraging model memorization for calibrated white-box membership inference," in *29th USENIX Security Symposium*, 2020.

[61] S. Yeom, I. Giacomelli, A. Menaged, M. Fredrikson, and S. Jha, "Overfitting, robustness, and malicious algorithms: A study of potential causes of privacy risk in machine learning," *Journal of Computer Security*, vol. 28, no. 1, pp. 35–70, 2020.

[62] K. Leino and M. Fredrikson, "Stolen memories: Leveraging model memorization for calibrated {White-Box} membership inference," in *29th USENIX security symposium (USENIX Security 20)*, pp. 1605–1622, 2020.

[63] L. Song and P. Mittal, "Systematic evaluation of privacy risks of machine learning models," in *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2615–2632, 2021.

[64] Q. Li, Y. Guo, and H. Chen, "Practical no-box adversarial attacks against dnns," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12849–12860, 2020.

[65] F. Guépin, M. Meeus, A.-M. Cretu, and Y.-A. de Montjoye, "Synthetic is all you need: removing the auxiliary data assumption for membership inference attacks against synthetic data," *arXiv preprint arXiv:2307.01701*, 2023.

[66] A. Sablayrolles, M. Douze, C. Schmid, Y. Ollivier, and H. Jégou, "White-box vs black-box: Bayes optimal strategies for membership inference," in *International Conference on Machine Learning*, pp. 5558–5567, PMLR, 2019.

[67] D. M. Hawkins, *Identification of outliers*, vol. 11. Springer, 1980.

[68] I. Guttman, "Care and handling of univariate or multivariate outliners in detecting spuriosity—a bayesian approach," *Technometrics*, vol. 15, no. 4, pp. 723–738, 1973.

[69] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Optics-of: Identifying local outliers," in *Principles of Data Mining and Knowledge Discovery: Third European Conference, PKDD'99, Prague, Czech Republic, September 15-18, 1999. Proceedings 3*, pp. 262–270, Springer, 1999.

[70] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[71] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial intelligence review*, vol. 22, pp. 85–126, 2004.

[72] J. W. Tukey *et al.*, *Exploratory data analysis*, vol. 2. Reading, MA, 1977.

[73] R. Wirth and J. Hipp, "Crisp-dm: Towards a standard process model for data mining," in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, vol. 1, pp. 29–39, Manchester, 2000.

[74] H. Ping, J. Stoyanovich, and B. Howe, "Datasynthesizer: Privacy-preserving synthetic datasets," in *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, pp. 1–5, 2017.

[75] J. Pearl, "Bayesian netwcrks: A model cf self-activated memory for evidential reasoning," in *Proceedings of the 7th conference of the Cognitive Science Society, University of California, Irvine, CA, USA*, pp. 15–17, 1985.

[76] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao, "Privbayes: Private data release via bayesian networks," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 4, pp. 1–41, 2017.

[77] M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, I. J. Goodfellow, and J. Pouget-Abadie, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, pp. 2672–2680, 2014.

[78] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2154–2156, 2018.

[79] P. Irolla and G. Châtel, "Demystifying the membership inference attack," in *2019 12th CMI Conference on Cybersecurity and Privacy (CMI)*, pp. 1–7, IEEE, 2019.

[80] R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern classification second edition john wiley & sons," *New York*, vol. 58, p. 16, 2001.

[81] M. Hernadez, G. Epelde, A. Alberdi, R. Cilla, and D. Rankin, "Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimensions," *Methods of Information in Medicine*, 2023.

[82] B.-C. Tai, S.-C. Li, Y. Huang, and P.-C. Wang, "Examining the utility of differentially private synthetic data generated using variational autoencoder with tensorflow privacy," in *2022 IEEE 27th Pacific Rim International Symposium on Dependable Computing (PRDC)*, pp. 236–241, IEEE, 2022.