The Impact of Synthetic Data on Membership Inference Attacks

Md Sakib Nizam Khan(✉) and Sonja Buchegger

KTH Royal Institute of Technology, Stockholm, Sweden
{msnkhan,buc}@kth.se

Abstract. Privacy of machine learning on Big Data has become a prominent issue in recent years due to the increased availability and usage of sensitive personal data to train the models. Membership inference attacks are one such issue that has been identified as a major privacy threat against machine learning models. Several techniques including applying differential privacy have been advocated to mitigate the effectiveness of inference attacks, however, they come at a cost of reduced utility/accuracy. Synthetic data is one approach that has been widely studied as a tool for privacy preservation recently but not much yet in the context of membership inference attacks. In this work, we aim to deepen the understanding of the impact of synthetic data on membership inference attacks. We compare models trained on original versus synthetic data, evaluate different synthetic data generation methods, and study the effect of overfitting in terms of membership inference attacks. Our investigation reveals that training on synthetic data can significantly reduce the effectiveness of membership inference attacks compared to models trained directly on the original data. This also holds for highly overfitted models that have been shown to increase the success rate of membership inference attacks. We also find that different synthetic data generation methods do not differ much in terms of membership inference attack accuracy but they do differ in terms of utility (i.e., observed based on train/test accuracy). Since synthetic data shows promising results for binary classification-based membership inference attacks on classification models explored in this work, exploring the impact on other attack types, models, and attribute inference attacks can be of worth.

Keywords: Synthetic Data · Machine Learning · Membership Inference Attack · Accuracy

1 Introduction

Machine learning has become one of the most essential elements of many technological solutions in recent years due to its huge benefits. The increasingly common applications of machine learning include image and speech recognition, predictive analytics, natural language processing, behavioral analysis, recommender systems, etc. In the majority of these scenarios, the data that the

B. Arief et al. (Eds.): SocialSec 2023, LNCS 14097, pp. 93–108, 2023.
https://doi.org/10.1007/978-981-99-5177-2_6

machine learning models build upon contains privacy-sensitive data of individuals. Privacy of machine learning has become a concerning issue recently due to the rapid increase in the use of personal and thus potentially privacy-sensitive data to train machine learning models. The problem of models with sufficient capacity (i.e., especially deep neural networks (DNN)) is that they tend to memorize the training data [35]. Several attacks have been developed that are capable of revealing information about the training data by exploiting the memorization capability of machine learning models.

Membership inference is one of the most prominent privacy vulnerabilities of machine learning models [24]. The goal of the attacker in the case of membership inference is to identify whether a given record is used to train the machine learning model or not which can eventually leak privacy-sensitive information. For instance, just revealing membership in a set used to train a target model related to a certain disease can reveal that a person has the disease which is a severe violation of privacy. Moreover, the attack can be carried out with minimal information in a black-box manner which increases the severity of the attack further. Shokri et al. [25] first proposed the black-box membership inference attack (MIA) which utilized a neural network-based binary classifier for detecting membership. Since then a plethora of attacks both black-box and white-box have been proposed for membership inference [10].

Model overfitting has been identified as one of the prime reasons for membership inference [25]. The machine learning models overfit when the model accuracy on the training data is significantly better compared to the accuracy on the unseen test data. This difference in accuracy is also termed as generalization error. Many different mitigation techniques have been proposed to solve the problem of membership inference mainly to achieve indistinguishability between the model's behavior on training and unseen test data. The mitigation strategies include the use of differential privacy [20], adding inference attack as a regularization term during model training [17], adding perturbed noise to models prediction outputs [13], etc. Nonetheless, it has been shown in some recent works that these defense strategies are not effective enough for some novel membership inference attacks (MIAs) [24].

The use of synthetic data for disclosure protection has gained significant attention in recent years. It has been widely studied as a measure of privacy protection for data release and analysis. The benefit of synthetic data is that there is no given direct linkability between the records and individuals since the data is not real. Furthermore, several works have shown that in terms of utility for machine learning tasks, synthetic data can also achieve acceptable accuracy close to original data, e.g., for tabular data [14] and image data [30]. However, more research is needed to understand what happens to MIAs if we train machine learning models using synthetic data instead of original data.

In this work, we thus investigate the impact of synthetic data on membership inference attacks against machine learning models where the target models are mainly classification models with supervised learning. First, we compare the membership inference attack accuracy between models trained on synthetic data

and original data using four publicly available datasets. Second, we compare different synthetic data generation methods in terms of MIAs. Third, we also study the effect of overfitting on models trained on original and synthetic data concerning membership inference attack accuracy by reducing the number of training data to intentionally overfit the model.

Organization. The rest of the paper is organized as follows. In Sect. 2, we discuss the related work, followed by an overview of the membership inference attack in Sect. 3. In Sect. 4, we briefly discuss synthetic data generation methods. We then present our experimental setup in Sect. 5 followed by the experimental results in Sect. 6. Finally, we conclude our work in Sect. 7.

2 Related Work

Since the inception of membership inference attacks against machine learning models, there has been a lot of work focusing on the mitigation of such attacks. Overfitting has been one of the prime reasons for membership inference. Thus, several works [17,23,25,27,33] focused on reducing the overfitting of machine learning models to defend against membership inference attacks. The works mainly utilized various regularization techniques such as l2 regularization [25], dropout [23], and adversarial regularization [17] to reduce the overfitting of machine learning models. Besides this, several mechanisms leveraged differential privacy [1,12,29,34] for mitigating the risk of membership inference attacks. The problem with employing these mitigation strategies is that besides reducing the privacy risks typically they reduce the performance of the target model as well.

We also consider some related works that look into different aspects of synthetic data concerning inference attacks. For instance, in a recent work, Stadler et al. [28] performed a quantitative evaluation of the privacy gain of synthetic data publishing and compared it with other anonymization techniques. The authors first empirically evaluated whether synthetic data generated by a wide range of generative models without any additional privacy measures provide robust protection against linkage attacks for all target records or not. Based on their experiments, the authors concluded that synthetic data does not provide uniform protection for all records and cannot protect some outlier records from linkage attacks. Next, the authors also show that differentially private synthetic data can protect such records from inference attacks but at a high utility cost. According to the author's findings, synthetic data cannot provide transparency about the privacy-utility tradeoff, unlike traditional anonymization techniques. It is impossible to predict what features of the original data will be preserved or suppressed in the synthetic data. Lastly, the authors also provide a framework as an open-source library to quantify the privacy gain of synthetic data publishing and compare the quality with different anonymization techniques.

Slokom et al. [26] investigated whether training a classifier on synthetic data instead of original data can mitigate the effectiveness of attribute inference attacks. The authors first demonstrate with a model trained on original data that by performing the attack an attacker can learn sensitive attributes both

about individuals present in the training data and also about previously unseen individuals. Then they replicated the attack on a model trained on synthetic data instead of the original data and found that the synthetic model is also as susceptible to attribute inference as the original model. According to the authors, this finding relates to the success of an attack inferring sensitive information from individuals using priors and not the machine learning model itself.

Zhang et al. [36] proposed a novel approach for membership inference against synthetic health data that tries to infer whether specific records are used for generating synthetic data or not. The authors evaluate fully synthetic and partially synthetic data based on their proposed approach for membership inference. According to the authors, their experimental results show that partially synthetic data are susceptible to membership inference whereas fully synthetic data are substantially more resilient against such inference attacks. The authors believe that their method can be used for preliminary risk evaluation of releasing any partially synthetic data.

Hu et al. [11] proposed to use data generated by Generative Adversarial Network (GAN) based on original data (i.e., synthetic data) for training machine learning models to defend against membership inference attacks. To ensure high utility, the authors utilize two different GAN structures with special training techniques for image and tabular data types respectively. For the generation of image data, the truncation technique is used whereas for tabular data clustering is used to ensure the quality of the generated data. Their empirical evaluation show that the proposed approach is effective against existing attack schemes and more efficient than existing defense mechanisms.

Though not focusing directly on membership attacks, another relevant related work for the potential privacy-protecting properties of synthetic data for data analysis, Ruiz et al. [22] investigate the linkability of synthetic to original data and argue that, based on a scenario where the attacker has access to the original dataset in its entirety, individuals' representations in synthetic and original datasets remain linked by the information they convey. Nonetheless, a contrasting finding was reported by Giomi et al. in [8] where the authors evaluated three types of privacy risk namely singling out, linkability, and inference risks of synthetic data using their proposed framework called Anonymneter. From their experiments, the authors observed that synthetic data is the least vulnerable to linkability. The findings indicate that one-to-one relationships between the original and generated records are not preserved in synthetic data.

The main difference between these works and our work is that none of the works perform a thorough investigation of the effect of synthetic data on the overall accuracy of membership inference attacks compared to the original data when using the synthetic data as training data, in contrast to publishing synthetic data in an effort to protect the privacy of the original data. For example, [28] investigates whether synthetic data can protect outliers equally as other records from inference attacks, [26] looks into attribute inference attacks, [36] focused on inference attacks against synthetic data generation process, and [11] only focused on GAN based methods for synthetic data generation and did not

investigate the effect of different synthesizers or overfitting on membership inference attack accuracy in the context of synthetic data. There exists a research gap regarding the impact of training machine learning models on synthetic data instead of original data on membership inference attacks which we try to bridge in this work.

3 Membership Inference Attacks (MIA)

The attack mechanism for membership inference can vary depending on the adversarial knowledge of the attacker. In this section, we provide a brief overview of adversarial knowledge and attack types and then discuss the attack mechanism used in this work in detail.

Adversarial Knowledge. The knowledge of an attacker can vary depending on how much information the attacker has access to about the machine learning model they are trying to attack. With regards to membership inference, there are two types of information that are beneficial for the attacker, namely information about the training data and information about the target model. The information about the training data refers to knowing the distribution of the training data. It is assumed in most of the membership inference attack settings that the distribution of the training data is known to the attacker which means that the attacker can obtain a so-called shadow dataset from the same distribution as the training data. To be realistic, it is also assumed that the training dataset and the shadow dataset are disjoint. The information about the target model refers to knowing the learning algorithm, model architecture, and parameters of the model. Depending on the knowledge of the attacker, MIAs are divided into two categories which are white-box attacks and black-box attacks.

White-Box Attack. In this type of attack, it is assumed that the attacker has knowledge about the training data and also about the target model. The knowledge that the attacker possesses in this setting includes information about the training data distribution, learning algorithm, model architecture, and parameters of the target model.

Black-Box Attack. In the case of black-box, it is assumed that the attacker only has information about the training data distribution and can query the target model in a black-box manner. For instance, in the case of the classification model, the attacker can provide an input record as a query to the model and can obtain the corresponding prediction output from the model. Nonetheless, the attacker does not have any information about the target model architecture, parameters, or the learning algorithm.

Approaches of Attack. Due to overparameterization, machine learning models such as DNN sometimes achieve the capacity to memorize features about the data that they are trained on [4]. As a result, the models behave differently on training data (i.e., members) versus the test data (i.e., nonmembers) that they have never seen. For instance, if the target model is a classification model, then it

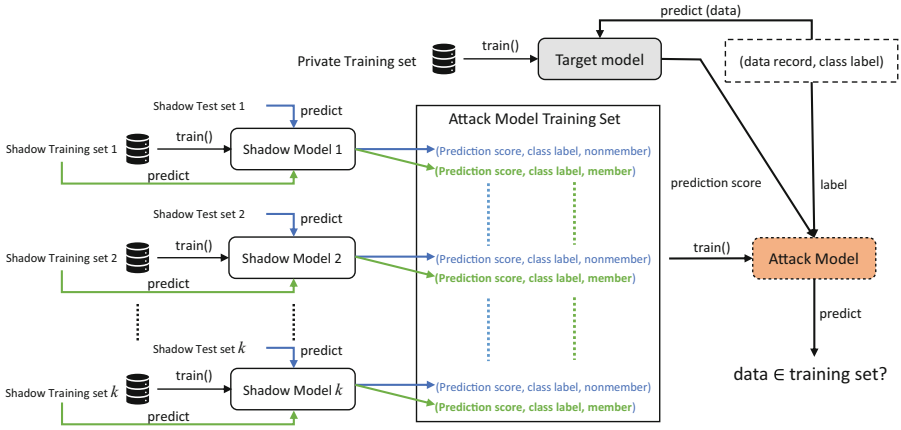

Fig. 1. Binary Classifier-based Membership Inference Attack

would classify the true class of training data with a higher confidence score than it would do for the unseen test data. This differentiation allows an attacker to build an attack model that can distinguish between members and nonmembers of a training dataset. Depending on how an attack model is created, the approach of membership inference attack can be divided into two major categories which are, binary classifier-based approach and metric-based approach. In this work, we use a binary classifier-based membership inference attack on classification models which is discussed in the following section. For more details on membership inference attacks based on other approaches and other models see [10].

3.1 Binary Classifier-Based Black-Box Attack

The basic idea of this approach for MIA is to train a binary classifier (i.e., attack model) capable of distinguishing the behavior of a target model on the members of the training set from nonmembers. The very first and most popular binary classifier-based membership inference attack technique (also termed shadow training) was proposed by Shokri et al. [25]. The main idea of the attack technique proposed by the authors is that the attacker trains multiple shadow models to imitate the behavior of the target model. Figure 1 shows an overview of the attack process. The shadow models are trained using shadow datasets which are drawn from a similar distribution as the training dataset of the target model. The assumption here is that the attacker knows the distribution of the training data which can be used to generate the shadow datasets. It is also assumed for non-triviality that the target model's training dataset and the shadow datasets are disjoint.

For each shadow model, the shadow dataset is divided into a shadow train set and a shadow test set. The shadow models are then trained using their corresponding train sets. Once the models are trained, prediction outputs are generated using each trained model for both their own train set and the unseen test set. The obtained output vectors for each shadow model are then labeled

as *members* for the model's own train set and as *non-members* for the unseen test set. The resulting labeled data make up the training data for the attack model which is the binary classifier inferring membership status. It is important to note that for each class of the target model, a separate attack model is trained to infer membership for the particular class. Once the attack model is trained, for the membership inference, the attacker first queries the target model with a particular record and obtains the prediction vector. Then the attacker passes the prediction vector value to the trained attack model with its true class to obtain the membership status.

4 Synthetic Data Generation

The idea of synthetic data as a confidentiality measure was introduced first in 1993 by Rubin [21], where the proposal was to use multiple imputation on all variables to generate fully synthetic data such that no original data is released. Since then multiple approaches such as parametric, non-parametric (e.g., classification and regression trees (CART) [3], random forest [2], etc.), saturated model, and so on have emerged for generating synthetic data. Recently, due to the advancement of machine learning, deep learning-based methods such as Generative Adversarial Networks (GANs) [9] are also getting widely used for generating fully synthetic data.

Synthpop is an open-source R package developed by Nowok et al. [18] for generating synthetic data based on the original data. The package provides the possibility of choosing parametric and non-parametric methods for synthetic data generation. The non-parametric method is mainly based on classification and regression tree (CART) which is capable of handling any type of data. For parametric the available methods include linear regression and predictive mean matching for numerical variables and logistic regression for categorical variables.

Synthetic Data Vault (SDV) is a python-based library also for generating synthetic data developed by Patki et al. [19]. Besides statistical approaches, SDV also includes GAN-based approaches to generate synthetic data. Conditional Tabular GAN (CTGAN) developed by Xu et al. [31] is a popular GAN-based approach capable of handling and achieving good performance for the mixed type of data. Apart from CTGAN, SDV also has another GAN-based approach called CopulaGAN which is a variation of CTGAN that utilizes cumulative distribution function (CDF) based transformations for making the learning process easier. In this work, we use different methods from *Synthpop* and *SDV* libraries for generating synthetic datasets.

5 Experimental Setup

This Section discusses the experimental setup. Table 1 provides a brief overview of the datasets used in this work for the empirical evacuation.

The goal of the experiments is to find out whether, given the synthetic data, the attacker can infer whether a particular record (e.g. an individuals data) was

used to generate the synthetic data that, in turn, was used to train the model as opposed to whether the record was a member of the dataset that was used to train the model directly. The inferred membership is thus an indirect one. Our codes for the experiments are available on GitHub[1].

As shown in Table 1, we use four publicly available datasets for our experiments which are Adult [15] and Avila [6] from UCI Machine Learning Repository [7], Polish quality of life dataset (SD2011) [5] from Synthpop example datasets, and Location-30 dataset created by Shokri et al. [25] based on location check-ins in the Foursquare social network, restricted to the Bangkok area and collected from April 2012 to September 2013 [32]. The classification task for the Adult dataset is to classify if the income of a person is above or below 50K and for Avila, it is to classify the author based on the patterns. For Location-30, the task is to predict the user's geosocial type based on their geographical profile. For the Polish dataset, there is no such common classification task. However, since it is a census dataset similar to Adult and has an income variable, we predict whether the income of a person is above or below 1K.

For the target model architecture of Adult, Avila, and Polish datasets, we use a fully connected deep neural network (DNN) model with layer sizes 600, 512, 256, and 128 before the final output layer. We use Adam optimizer for the learning with 200 epochs. For the Location-30 dataset, we also use a fully connected DNN but with layer sizes 512, 248, 128, and 64 before the output layer. Adam optimizer with 100 epochs is used for learning. For Location-30, we use l2 regularization with a weight of 0.0007 whereas for the other three datasets weight of 0.00003 is used. Tanh is used as the activation function for all of the datasets. We train 20 shadow models for each of the datasets.

For synthetic models, we generate synthetic data using the original records sequentially starting from 1. For instance, in the case of Adult, we use original records from 1 to 10000 for synthetic data generation. Similarly, for Polish original records 1–2500 is used for generating synthetic data. The data for the shadow model and nonmember test set for the attack model is randomly drawn from the remaining original dataset whereas the member test set for the attack model is drawn randomly from the original records used for synthetic data generation.

6 Experimental Results

6.1 Membership Inference Accuracy Comparison

In this experiment, we evaluate the impact of synthetic data on membership inference by comparing membership inference attack accuracy between the model trained on original data versus the model trained on synthetic data. As mentioned previously, we use the binary classification-based attack technique proposed by Shokri et al. [25] for membership inference.

For this experiment, for each of the four datasets, we train an original target model and a synthetic target model. We divide the original dataset by drawing

[1] https://github.com/sakib570/mia-synthetic-data.

Table 1. Dataset Description

Dataset	Total Instances	No. Classes	Model Type	Data Type	Synthetic Data Gen.	Target Model		Shadow Train Size	Attack Model	
						Train Set	Test Set		Member	Non-member
Adult	48842	2	Original	Original	-	7000	3000	7000	2500	2500
			Synthetic	Original	10000	-	-	7000	2500	2500
				Synthetic	-	7000	3000	-	-	-
Polish	5000	2	Original	Original	-	840	360	840	600	600
			Synthetic	Original	2500	-	-	840	600	600
				Synthetic	-	840	360	-	-	-
Location-30	5010	30	Original	Original	-	840	360	840	600	600
			Synthetic	Original	2500	-	-	840	600	600
				Synthetic	-	7000	3000	-	-	-
Avila	20867	12	Original	Original	-	7000	3000	7000	2500	2500
			Synthetic	Original	10000	-	-	7000	2500	2500
				Synthetic	-	7000	3000	-	-	-

three disjoint datasets where one dataset is used for training and testing the target model, one is for training and testing the shadow model, and the remaining one is for testing the attack model. However, for the synthetic model, we first generate a synthetic dataset based on the original dataset. The portion of the dataset used for generating synthetic data is similarly drawn from the original data as in the case of the original model. For generating synthetic data, we use *Synthpop* method *CART with Catall* for all of the datasets. In the case of the synthetic model, the data for the shadow model and testing the attack model are drawn from the original dataset similar to the original model. Thus, the only difference between the original and synthetic model is that, for synthetic, the target model is trained using the synthetic data that is generated based on original data instead of directly training on original data which is the case for the original model. This introduces an extra layer of indirection to the original data for the synthetic model and the goal of this experiment is to study the impact of this indirection on the membership inference accuracy. Table 1 depicts for each dataset the number of records used for the target, shadow, and attack model for both the original and the synthetic scenario.

For the evaluation, we measure the train and test accuracy of the target model and the attack accuracy and precision of the attack model. We use an equal number of members and nonmembers (i.e., 50% members and 50% nonmembers) for the validation of the attack model. Thus, attack accuracy close to 0.5 would indicate that the attack performance is as good as a random guess. The train and test accuracy comparison between the original and synthetic model provides an intuition of whether the synthetic model is behaving similarly to the original model or not. Table 2 depicts the results obtained from this experiment for each of the four datasets.

As shown in Table 2, concerning train and test accuracy, for all of the datasets, the original and synthetic models achieve similar results. The deviation in accuracy remains within ∼0%–5% except for the test accuracy in the case of the Location-30 dataset. The synthetic model for the Location-30 dataset achieves much better test accuracy (i.e., ∼14%) than the original model,

Table 2. Accuracy Comparison between Original and Synthetic Model

Dataset	Target Model	Train Accuracy	Test Accuracy	Attack Accuracy	Attack Precision
Adult	Original	92.84	80.73	0.545	0.524
	Synthetic	93.057	83.26	0.5042	0.5023
Polish	Original	97.38	55.83	0.66	0.59
	Synthetic	98.33	60.84	0.53	0.52
Location-30	Original	100	48.61	0.83	0.75
	Synthetic	100	64.44	0.54	0.57
Avila	Original	99.92	98.66	0.5108	0.5054
	Synthetic	99.95	99.15	0.4991	0.4992

however, both models achieve the same train accuracy (i.e., 100%). This indicates that the synthetic model for the Location-30 dataset generalizes better than the original model.

In terms of attack accuracy, we see that the synthetic models achieve lower attack accuracy than the corresponding original models as well as the accuracy values are close to the baseline accuracy of 0.5 for all of the datasets. The most significant reduction happens in the case of the Location-30 dataset where the accuracy drops from 0.83 for the original model to 0.54 for the synthetic model. Similarly, for the Polish dataset, we see that attack accuracy reduces from 0.66 to 0.53 for the synthetic model. In the case of Adult and Avila, even though the attack accuracy for the original model is already close to the baseline accuracy of 0.5, we still see that the synthetic model brings the accuracy further close to the baseline. One reason behind the significant reduction of attack accuracy for the synthetic model of the Location-30 dataset can be a better generalization. Since overfitting is one of the prime reasons for membership inference and the synthetic model for Location-30 on top of the indirection layer of synthetic data achieves better generalization (i.e., less overfitting), it is able to reduce the attack accuracy further. The attack precision scores also show a similar trend as the attack accuracy.

In summary, the experiment reveals that training on synthetic data instead of original data can significantly reduce the effectiveness of membership inference attacks. For all of the datasets we see that training on synthetic data is able to bring down the attack accuracy close to random guessing. The exact results for attack accuracy have been shown to not only depend on the dataset but also the bias of the samples used [24], and should thus be taken as approximations though we did not influence the sample selection in our experiments.

6.2 Evaluation of Synthetic Data Generation Methods

In this experiment, we evaluate different synthetic data generation methods concerning membership inference. We compare four different synthetic data

Table 3. Comparison of Synthetic Data Generation Methods

Dataset	Model	Train Accuracy	Test Accuracy	Generalization Error	Attack Accuracy
Adult	Original	92.84	80.73	12.11	0.545
	Synthpop CART+Catall	93.057	83.26	9.797	0.5042
	Synthpop Parametric	87.44	82.8	4.64	0.49
	SDV CTGAN	92.72	77.16	15.56	0.49
	SDV Copula GAN	84.11	81.86	2.25	0.498
Polish	Original	97.38	55.83	41.55	0.66
	Synthpop CART+Catall	98.33	60.84	37.49	0.53
	Synthpop Parametric	92.85	55.96	36.89	0.51
	SDV CTGAN	99.88	61.26	38.62	0.505
	SDV Copula GAN	99.76	50.48	49.28	0.49
Location-30	Original	100	48.61	51.39	0.83
	Synthpop CART+Catall	100	64.44	35.56	0.54
	Synthpop Parametric	100	24.16	75.84	0.534
	SDV CTGAN	100	8.33	91.67	0.506
	SDV Copula GAN	100	4.44	95.56	0.491
Avila	Original	99.92	98.66	1.26	0.5108
	Synthpop CART+Catall	99.95	99.15	0.8	0.4991
	Synthpop Parametric	98.04	39.6	58.44	0.45
	SDV CTGAN	99.028	64.033	34.995	0.5074
	SDV Copula GAN	96.785	35.06	61.725	0.4974

generation methods which are *Synthpop CART with Catall, Synthpop Parametric, SDV CTGAN,* and *SDV Copula GAN*. The reason behind choosing these methods is that we want to cover traditional approaches such as CART and parametric as well as more recent GAN-based approaches. For this experiment, we use the same attack technique and training/testing methods used in Sect. 6.1.

Table 3 shows the train and test accuracy, generalization error, and attack accuracy obtained for different synthetic data generation methods for each of the four datasets. For the Adult dataset, in terms of train/test accuracy and generalization error, all of the synthetic data generation methods obtain similar results as the original model. The attack accuracy for all of the methods is also very close to the baseline accuracy of 0.5 with negligible differences. *Synthpop CART with Catall* obtained the closest train/test accuracy and generalization error to the original model. For the Polish dataset, we also see a similar trend where all of the methods obtained similar results to the original model for train/test accuracy and generalization error. However, in terms of attack accuracy, we see more significant differences between the methods than in the case of Adult. For Polish, *SDV CTGAN* obtained the closest train/test accuracy to the original model and also achieved an attack accuracy value (i.e., 0.505) close to baseline accuracy.

For Location-30, the methods perform very differently than what we saw for Adult and Polish. In terms of test accuracy, *SDV CTGAN* and *Copula GAN* obtain only 8.33 and 4.44 respectively which are very poor compared to the test accuracy of 48.61 of the original model. *Synthpop Parametric* perform a bit better than the GAN methods and obtain a test accuracy value of 24.16 which is still far from the test accuracy of the original model. This can be an indication that the synthetic datasets generated by these methods are unable to capture the complete features of the original data. Nonetheless, *Synthpop CART with Catall* perform much better in terms of test accuracy than the other methods and obtain even better test accuracy value (i.e., 64.44) than the original model which indicates that the synthetic model generalizes better than the original model. In terms of attack accuracy even though *Synthpop CART with Catall* obtain 0.54 which is the highest attack accuracy value compared to other methods, it is still close to the baseline accuracy and not significantly far away from other methods. Thus, overall *Synthpop CART with Catall* performs best for the Location-30 dataset. For the Avila dataset, we also see a similar trend as the Location-30 dataset where *Synthpop Parametric, SDV CTGAN,* and *SDV Copula GAN* perform poorly in terms of test accuracy compared to the original model. Also, *Synthpop CART with Catall* performs best overall and achieves both train/test accuracy close to the original model and attack accuracy close to the baseline.

In summary, from this experiment, we see that the synthetic data generation method can perform differently depending on the dataset. The performance of the synthetic model can also vary depending on the method and some methods can imitate the original model better than others in terms of train and test accuracy. For attack accuracy, we do not see any significant differences between the methods. Finally, *Synthpop CART with Catall* performs best for the combination of train/test accuracy and attack accuracy for all of the datasets.

6.3 Effect of Overfitting

Model overfitting has been identified as one of the most common causes of membership inference by many studies [16,25]. Hence, in this experiment, we want to study whether synthetic data has any effect on membership inference of overfitted models and whether the reduction of the attack accuracy achieved remains robust even when the training process is biased in favor of the attacker. Since overfitting occurs when models have sufficient memorizing capacity, reducing the size (number of records) of the training dataset increases the memorizing capability which in turn increases the overfitting of a model. Thus, for this experiment, we varied the size of training datasets to intentionally overfit both the original and synthetic models and then measured the attack accuracy.

Table 4 shows the obtained results. For this experiment, we use the Adult and Location-30 datasets and reduce the number of training data to 100 for both the datasets to achieve overfitting. We compare the results of train size 100 with train size 7000 for Adult and 840 for Location-30, the same train sizes used in previous experiments. For the Adult dataset, when we reduce the train size from

7000 to 100 the generalization error increases from 12.11 to 28.01 respectively for the original model. This indicates that the original model with train size 100 overfits more on the training data. Due to this overfitting, the attack accuracy for the original model increases to 0.63 (train size 100) from 0.545 (train size 7000). However, when we look at the synthetic model attack accuracy, it only increases to 0.53 from 0.5042 for the same. The generalization error for the synthetic model also does not increase as much as the original model.

Table 4. Effect of Overfitting

Dataset	Train Size	Model Type	Train Accuracy	Test Accuracy	Generalization Error	Attack Accuracy
Adult	7000	Original	92.84	80.73	12.11	0.545
		Synthetic	93.057	83.26	9.797	0.5042
	100	Original	98	69.99	28.01	0.63
		Synthetic	95.99	75.99	20	0.53
Location-30	840	Original	100	48.61	51.39	0.83
		Synthetic	100	64.44	35.56	0.54
	100	Original	100	28.45	71.55	1
		Synthetic	100	56	44	0.55

For the Location-30 dataset, synthetic data shows an even more significant effect on attack accuracy. In the case of the original model, the generalization error increases to 71.55 (train size 100) from 51.39 (train size 840). Similarly, the attack accuracy increases to 1 (train size 100) from 0.83 (train size 840). The attack accuracy 1 for the model with train size 100 means that the attack model can successfully infer members and nonmembers with 100% accuracy. Nonetheless, for synthetic data, the attack accuracy for train size 100 is 0.55 which is not far from the baseline accuracy. The synthetic model significantly reduces the attack accuracy compared to the original model. In terms of the generalization error, we see a similar trend as the Adult dataset.

In summary, the experiment reveals that in scenarios where training on original data results in a highly overfitted model, training on synthetic data instead can significantly reduce the possibility of membership inference attacks.

7 Discussion and Conclusion

In this work, we investigate the impact of synthetic data on membership inference attacks. We also compare different synthetic data generation methods in terms of membership inference attack accuracy and study the effect of overfitting on the synthetic and original models. Our investigation reveals that training on synthetic data can effectively reduce the membership inference attack accuracy compared to the models trained on original data. The synthetic model can bring

down the attack accuracy close to baseline accuracy which is as good as a random guess even for datasets that have significantly high attack accuracy. In the case of synthetic data generation methods, our experiments reveal that some methods generate synthetic data such that the models train on them imitate the original model better than others in terms of train and test accuracy. However, the attack accuracy does not vary significantly for any of the methods. Thus, the choice of synthetic data generation method should depend on the utility and one should choose a method that provides the best utility close to the original data. Furthermore, our investigation on overfitting reveals that training on synthetic data can significantly reduce the possibility of membership inference in scenarios where original data produces highly overfitted models. In this work, we just look at binary classification-based membership inference attacks. Nonetheless, there are other metric-based and also more advanced membership inference attacks. In future work, we, therefore, want to investigate whether synthetic data has a similar impact on such attacks. Additionally, further investigation can be done to understand if synthetic data can mitigate or reduce the effectiveness of attribute inference attacks on machine learning models.

Acknowledgment. This work was partially supported by the Wallenberg AI, Autonomous Systems & Software Program (WASP) funded by Knut & Alice Wallenberg Foundation.

References

1. Abadi, M., et al.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 308–318 (2016)
2. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
3. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees, vol. 37, no. 15, pp. 237–251. Wadsworth International Group (1984)
4. Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., Song, D.: The secret sharer: evaluating and testing unintended memorization in neural networks. In: 28th USENIX Security Symposium (USENIX Security 2019), pp. 267–284 (2019)
5. Czapiński, J., Panek, T.: Social diagnosis 2011. Objective and subjective quality of life in Poland. Contemp. Econ. **5**(3) (2011)
6. De Stefano, C., Maniaci, M., Fontanella, F., di Freca, A.S.: Reliable writer identification in medieval manuscripts through page layout features: the "Avila" bible case. Eng. Appl. Artif. Intell. **72**, 99–110 (2018)
7. Dua, D., Graff, C.: UCI Machine Learning Repository (2017). http://archive.ics.uci.edu/ml
8. Giomi, M., Boenisch, F., Wehmeyer, C., Tasnádi, B.: A unified framework for quantifying privacy risk in synthetic data. In: Proceedings of Privacy Enhancing Technologies Symposium (2023). https://doi.org/10.56553/popets-2023-0055
9. Goodfellow, I., et al.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020)
10. Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P.S., Zhang, X.: Membership inference attacks on machine learning: a survey. ACM Comput. Surv. (CSUR) **54**(11s), 1–37 (2022)

11. Hu, L., et al.: Defending against membership inference attacks with high utility by GAN. IEEE Trans. Dependable Secure Comput. **20**(3), 2144–2157 (2023). https://doi.org/10.1109/TDSC.2022.3174569
12. Jayaraman, B., Evans, D.: Evaluating differentially private machine learning in practice. In: 28th USENIX Security Symposium (USENIX Security 2019), pp. 1895–1912 (2019)
13. Jia, J., Salem, A., Backes, M., Zhang, Y., Gong, N.Z.: Memguard: defending against black-box membership inference attacks via adversarial examples. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, pp. 259–274 (2019)
14. Khan, M.S.N., Reje, N., Buchegger, S.: Utility assessment of synthetic data generation methods. In: Privacy in Statistical Databases USB Proceedings (2022)
15. Kohavi, R., Becker, B.: Adult Dataset. UCI Machine Learning Repository, vol. 5, p. 2093 (1996)
16. Li, J., Li, N., Ribeiro, B.: Membership inference attacks and defenses in classification models. In: Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy, pp. 5–16 (2021)
17. Nasr, M., Shokri, R., Houmansadr, A.: Machine learning with membership privacy using adversarial regularization. In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, pp. 634–646 (2018)
18. Nowok, B., Raab, G.M., Dibben, C.: Synthpop: bespoke creation of synthetic data in R. J. Stat. Softw. **74**, 1–26 (2016)
19. Patki, N., Wedge, R., Veeramachaneni, K.: The synthetic data vault. In: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 399–410. IEEE (2016)
20. Rahman, M.A., Rahman, T., Laganière, R., Mohammed, N., Wang, Y.: Membership inference attack against differentially private deep learning model. Trans. Data Priv. **11**(1), 61–79 (2018)
21. Rubin, D.B.: Statistical disclosure limitation. J. Off. Stat. **9**(2), 461–468 (1993)
22. Ruiz, N., Muralidhar, K., Domingo-Ferrer, J.: On the privacy guarantees of synthetic data: a reassessment from the maximum-knowledge attacker perspective. In: Domingo-Ferrer, J., Montes, F. (eds.) PSD 2018. LNCS, vol. 11126, pp. 59–74. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99771-1_5
23. Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., Backes, M.: Ml-leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. arXiv Preprint arXiv:1806.01246 (2018)
24. Senavirathne, N., Torra, V.: Dissecting membership inference risk in machine learning. In: Meng, W., Conti, M. (eds.) CSS 2021. LNCS, vol. 13172, pp. 36–54. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-94029-4_3
25. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 3–18. IEEE (2017)
26. Slokom, M., de Wolf, P.P., Larson, M.: When machine learning models leak: an exploration of synthetic training data. In: Domingo-Ferrer, J., Laurent, M. (eds.) PSD 2022. LNCS, vol. 13463, pp. 283–296. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-13945-1_20
27. Song, L., Mittal, P.: Systematic evaluation of privacy risks of machine learning models. In: 30th USENIX Security Symposium (USENIX Security 2021), pp. 2615–2632 (2021)

28. Stadler, T., Oprisanu, B., Troncoso, C.: Synthetic data-anonymisation groundhog day. In: 31st USENIX Security Symposium (USENIX Security 2022), pp. 1451–1468 (2022)
29. Wang, D., Ye, M., Xu, J.: Differentially private empirical risk minimization revisited: faster and more general. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
30. Wood, E., Baltrušaitis, T., Hewitt, C., Dziadzio, S., Cashman, T.J., Shotton, J.: Fake it till you make it: face analysis in the wild using synthetic data alone. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3681–3691 (2021)
31. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional GAN. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
32. Yang, D., Zhang, D., Qu, B.: Participatory cultural mapping based on collective behavior data in location-based social networks. ACM Trans. Intell. Syst. Technol. (TIST) **7**(3), 1–23 (2016)
33. Yin, Y., Chen, K., Shou, L., Chen, G.: Defending privacy against more knowledgeable membership inference attackers. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 2026–2036 (2021)
34. Yu, L., Liu, L., Pu, C., Gursoy, M.E., Truex, S.: Differentially private model publishing for deep learning. In: 2019 IEEE Symposium on Security and Privacy (SP), pp. 332–349. IEEE (2019)
35. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning (still) requires rethinking generalization. Commun. ACM **64**(3), 107–115 (2021)
36. Zhang, Z., Yan, C., Malin, B.A.: Membership inference attacks against synthetic health data. J. Biomed. Inform. **125**, 103977 (2022)