IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 June 2024, accepted 12 August 2024, date of publication 21 August 2024, date of current version 3 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3446913

RESEARCH ARTICLE

Benchmarking Evaluation Protocols for Classifiers Trained on Differentially Private Synthetic Data

PARISA MOVAHEDI[1], **(Member, IEEE), VALTTERI NIEMINEN**[1,2], **ILEANA MONTOYA PEREZ**[1], **HIBA DAAFANE**[1], **DISHANT SUKHWAL**[1], **TAPIO PAHIKKALA**[1], **AND ANTTI AIROLA**[1]

[1]Department of Computing, Turku University, 20014 Turku, Finland
[2]Helsinki University Hospital (HUS), 00290 Helsinki, Finland

Corresponding author: Parisa Movahedi (parmov@utu.fi)

This work was supported in part by the Business Finland under Grant 37428/31/2020, and in part by European Union’s Horizon Europe Research and Innovation Programme under Grant 101095384.

ABSTRACT Differentially private (DP) synthetic data has emerged as a potential solution for sharing sensitive individual-level biomedical data. DP generative models offer a promising approach for generating realistic synthetic data that aims to maintain the original data’s central statistical properties while ensuring privacy by limiting the risk of disclosing sensitive information about individuals. However, the issue regarding how to assess the expected real-world prediction performance of machine learning models trained on synthetic data remains an open question. In this study, we experimentally evaluate two different model evaluation protocols for classifiers trained on synthetic data. The first protocol employs solely synthetic data for downstream model evaluation, whereas the second protocol assumes limited DP access to a private test set consisting of real data managed by a data curator. We also propose a metric for assessing how well the evaluation results of the proposed protocols match the real-world prediction performance of the models. The assessment measures both the systematic error component indicating how optimistic or pessimistic the protocol is on average and the random error component indicating the variability of the protocol’s error. The results of our study suggest that employing the second protocol is advantageous, particularly in biomedical health studies where the precision of the research is of utmost importance. Our comprehensive empirical study offers new insights into the practical feasibility and usefulness of different evaluation protocols for classifiers trained on DP-synthetic data.

INDEX TERMS Biomedical data, classification, differential privacy, generative AI, model evaluation, synthetic data.

I. INTRODUCTION

Sharing medical data for research purposes is challenging due to privacy concerns and strict regulation such as GDPR [1]. This is a major limiting factor when developing machine learning based medical applications such as diagnostic classifiers, prognostic models or patient monitoring systems that require individual-level patient data for training and validation [2].

Synthetic data generation has been proposed as a method for enabling medical data sharing without compromising patient privacy [3]. However, it has been repeatedly shown, that synthetic data is not inherently private [4], [5]. The outputs of machine learning models can leak information about their training data [6], and generative models used to create synthetic data are no different in this respect. This, in turn, creates vulnerabilities to attacks such as membership and attribute inference [4], [7]. The most widely accepted solution to fix this shortcoming is to combine generative models with differential privacy (DP). DP is a mathematical framework proposed by Dwork et al. [8], which allows quantifying and enforcing privacy for computations done on sensitive data. It provides a probabilistic guarantee on how much information could be revealed about any individual from the result of the computation, that could not be inferred if the individual was not present in the data. Numerous

The associate editor coordinating the review of this manuscript and approving it for publication was Rupak Kharel.

methods have been proposed in the recent years to generate DP-synthetic tabular data which includes marginal-based approaches [9], [10] often achieving higher performances in comparison to the alternative approaches such as generative adversarial networks and diffusion models [11], [12], [13]

Even though DP is widely regarded as the leading standard for privacy protection, several challenges remain when deploying it in real healthcare applications. Achieving a balance between utility and privacy requires careful selection of the privacy parameters and other implementation considerations [14]. However, there is limited understanding of what an appropriate value of the privacy parameter is for a specific system, purpose, or dataset, and little guidance on how to determine it. Other issues to consider in DP implementation are ensuring regulatory compliance like the GDPR [1], and effectively communicating DP guarantees to build trust. Addressing these challenges requires collaboration between researchers, regulators, and industry practitioners to facilitate broader adoption of DP in practical healthcare applications. Finally, once the privacy level is agreed on, one should leverage advanced and well tested DP-algorithm in order to achieve state-of-the-art utility level.

It is an open question how to evaluate the real-world prediction performance of machine learning models trained on DP-synthetic data. For example, DP-synthetic datasets can introduce or amplify biases in machine learning models [15], [16]. Enforcing privacy through DP always comes with a cost, with DP-synthetic data by definition being a distorted version of the real data that may not yield the same results as the original would [8], [17], [18], and [19]. Consequently, even if a classifier trained on synthetic data performs well on synthetic test data, it might not make as accurate predictions when applied to real-world data. A common alternative approach to evaluating the quality of classifiers trained on synthetic data is to calculate the prediction performance on a test set consisting of real data [20], [21], [22]. This is a useful protocol in scientific studies where the goal is to compare the quality of different data synthesizing methods, constituting the so-called downstream evaluation approach. However, these approaches may not be directly implementable in sensitive real-world applications. The primary justification for the use of DP-synthetic data arises from the issue that unrestricted access to private data cannot be granted due to privacy concerns.

The objective of this study is to address the existing void in literature on assessing the predictive performance of classifiers trained using DP-synthetic data through empirical investigation of two alternative downstream evaluation protocols. In the first protocol the analyst has access only to DP-synthetic data to test the model, whereas, in the second protocol a DP-query is allowed to be performed on the real held-out data. This data is on the curator's side and it is not publicly available, but the analyst is able to submit the models trained on DP-synthetic data and receive the DP-wise calculated statistics such as classifier performance. Furthermore, to asses the performance of each of these protocols, we have introduced a protocol evaluation criteria by taking into account both the systematic and random aspects of the error within each protocol.

This paper extends our previous work [23], where preliminary results with a more limited evaluation were presented. In this study, we introduce a novel measure for assessing the quality of the protocols in evaluating the downstream classifiers. Further, we conduct a comprehensive empirical assessment of the two protocols across five medical datasets and five different DP-synthetic data generation methods.

II. BACKGROUND

A. DIFFERENTIAL PRIVACY

A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-differential private if for all datasets D_1 and D_2 that differ in at most one record, and for all measurable sets S of outputs, the following inequality holds:

$$\Pr[\mathcal{A}(D_1) \in S] \leq e^{\epsilon} \Pr[\mathcal{A}(D_2) \in S] + \delta. \qquad (1)$$

where ϵ represents the user-defined upper bound on the privacy loss, which affects the amount of noise added to the algorithm and δ is a small constant representing a extremely unlikely event, where the guarantee does not hold [24]. (ϵ)-DP is a special case of (ϵ, δ)-DP where $\delta = 0$. A smaller value of ϵ provides a stronger guarantee. (ϵ, δ)-differential privacy guarantees that an outcome of the algorithm is nearly equally probable on any two datasets that differ in at most one record, with privacy loss bounded by epsilon ϵ with probability at least 1-δ.

Generally, in DP methods privacy is protected at a given level of ϵ if the algorithm's output does not heavily depend on the input data of any single individual present in the dataset. The results should be almost the same whether or not an individual's information is included in the dataset. This is achieved by adding randomization to the algorithm applied on the data utilizing a noise mechanism calibrated based on the level of privacy (ϵ) and the algorithm to be privatized.

Although a singular, universally acceptable value for the privacy budget ϵ cannot be established due to its contextual dependency, in the literature, values of $\epsilon \; \leq 1$ have been considered to provide strong guarantee [25], [26], and depending on the type of data and task, values of $\epsilon \leq 10$ have been observed to still result in meaningful guarantees [27].

B. DP-SYNTHETIC DATA

The goal of generating DP-synthetic healthcare data is to create synthetic records that closely resemble the original data and yield comparable results in analyses, all while ensuring privacy protection. This often involves optimizing the level of introduced noise in the generation process by carefully tuning the privacy budget ϵ and the choice of privacy mechanisms to achieve the desired privacy-utility trade-off [18].

For all DP-synthetic data generators (assuming $\delta = 0$ for the sake of simplicity), the DP-guarantee can be expressed as follows. If the data corresponding to a single individual is added to the original data used to train the synthetic data generator, the log-probability of a certain set of synthetic data being generated can change at most ϵ. Therefore, the existence of any single individual's data in the curators database could be determined based on the synthetic data only with negligible probability.

The state of the art approaches for tabular data generation are the marginal-based generative methods which are a class of techniques used in machine learning and data analysis to generate data that matches the marginal distributions of variables in a given dataset [28].

Generative adversarial networks (GAN) represent another mainstream approach to synthetic data generation [29], but while these methods are well suited to image generation tasks on large datasets, they have been shown to often fail to surpass simple baselines when applied to tabular data generation [15], [22], [30].

Advancements in machine learning and algorithmic decision support systems have made it possible for predictive models to enhance or even fully automate human decision-making across a diverse range of healthcare related scenarios. When deployed thoughtfully, these technologies hold the promise of enhancing precision and predictive power. Model evaluation is a critical component of machine learning, determining how well a trained model performs on unseen data. The utility of DP synthetic data can be assessed by how well a machine learning model, trained on this data, performs. Typically, the model is trained using the synthetic data and its accuracy is subsequently evaluated by testing it with real-world data [20], [21], [22]. However, in real-world scenarios a researcher utilizing the synthetic data can not assume unrestricted access to real test data due to privacy concerns. One possibility is to use synthetic data also for testing. Protocols commonly used in crowd-sourced machine learning competitions [31], [32] suggest another possible approach, where a limited access to a hidden test set is provided in order to ensure fair and unbiased evaluation of the developed models' generalization capabilities.

III. RESEARCH CONTRIBUTION

A. PROPOSED EVALUATION PROTOCOLS

Two distinct evaluation protocols are considered enabling data analysts to validate a classifier trained on DP-synthetic data, utilizing either DP-synthetic or sensitive real test data. In real-world biomedical data sharing scenarios, the data holder may choose to release only the DP-synthetic data to the public, without providing any additional access or services related to the actual data. In this setting, the analyst is typically restricted to using only the synthetic data for constructing and assessing a machine learning model. However, if the data holder were to provide limited access to a distinct subset of real data for testing purposes, the analyst would then have the opportunity to submit their model for evaluation. This evaluation would include applying the constructed model to the real data subset within a secure and private environment, which might result in valuable and less biased evaluation metrics. The first setting is simpler, more cost-effective, and demands less infrastructure. However, the second protocol grants access to a real test dataset which may yield a more realistic evaluation of the model's performance.

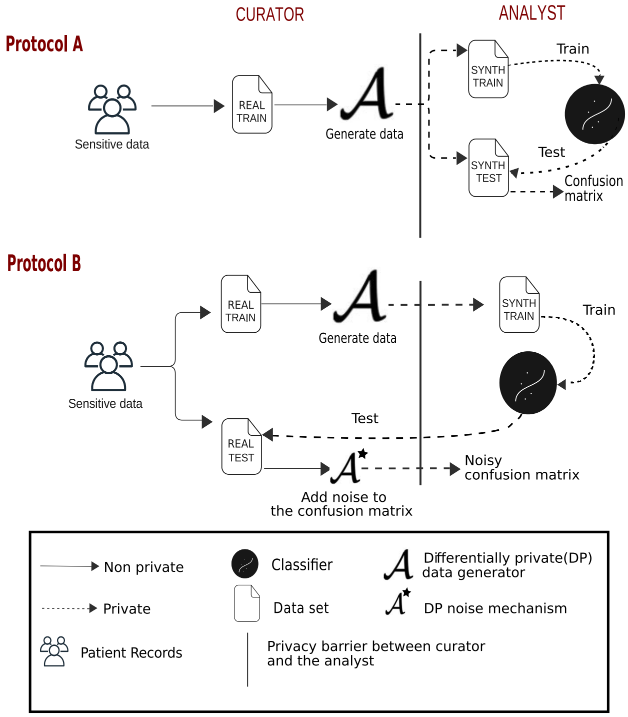

FIGURE 1. Data flow diagram for the proposed evaluation protocols.

Figure 1 illustrates the data flow for protocol A in top and protocol B in the bottom where for both protocols the curator side (the private environment, e.g. a hospital server) is divided by a privacy barrier (vertical line) from the analyst (non-private environment). The flow of the data is shown with the non-dashed line if the data is not privatized. If the data is differentially private the flow is illustrated with dashed lines. In the following we present the protocols in detail, where the steps can be followed by the arrows in the figure.

- **Protocol A: Syn-Only**. As shown in Figure 1 (top):
  1) Trusted data curator creates a training dataset from the sensitive data.
  2) A DP-generative model is trained with the sensitive training dataset given the pre-specified privacy budget ϵ.
  3) The curator generates synthetic dataset with the DP-generative model. The synthetic data is released to the analyst operating outside the private environment. The analyst divides the data to train and test sets.
  4) The analyst trains a classifier with the synthetic training data.
  5) The analyst calculates the predictions of the classifier on the synthetic test set.

6) The confusion matrix is computed by comparing true and predicted classes for the synthetic test set.

- **Protocol B: Syn-Real**. Illustrated in Figure 1 (bottom):
  1) Trusted data curator creates disjoint training and test datasets from the sensitive data
  2) A DP-generative model is trained with the sensitive training dataset given the pre-specified privacy budget ϵ.
  3) Curator generates synthetic dataset with the DP-generative model. The synthetic data is released to the analyst operating outside the private environment.
  4) The analyst trains a classifier with the synthetic training data.
  5) The analyst sends the classifier to the curator.
  6) The curator calculates the predictions of the classifier on the private real test set.
  7) The confusion matrix is computed by comparing true and predicted classes for the real test set.
  8) The resulting confusion matrix is released to the analyst with a DP guarantee using the Laplace mechanism.

In the outlined approach, the same ϵ value is applicable to both training and testing due to the disjoint nature of the datasets, leveraging the parallel composition principle of differential privacy [33]. This principle allows for the same privacy budget to be used across separate analyses without cumulative privacy loss, given the independence of the training and testing sets. Thus, employing the same ϵ for both phases does not compromise the overall privacy guarantees, efficiently maintaining privacy without additional expenditure.

The protocols provide performance evaluation of the model trained on synthetic data. As a concrete example, we consider the calculation of a confusion matrix, as it can be used as basis for calculating multiple different performance measures. In a binary classification problems, the matrix consists of counts of true positive, true negative, false positive, and false negative predictions. In Protocol A, these counts are calculated from the DP-wise generated synthetic test set. For Protocol B, where real test data is used to calculate the confusion matrix, one needs to apply DP to ensure that privacy of individuals in the test set will not be compromised when releasing the matrix. Sharing the matrix directly could violate privacy regulations constituting an information breach. In some cases, an attacker could use a confusion matrix to infer information about individuals if the dataset is small and specific enough or highly unbalanced. For example, if there's a small number of individuals with a rare condition such as a disease, the confusion matrix might provide clues about whether a particular individual is in that rare group or not. Therefore, by incorporating DP to the generation of a confusion matrix, the privacy of individuals' data can be protected. The process of generating a DP confusion matrix involves introducing carefully calibrated noise to the counts in the confusion matrix. Noise is added to each component of the confusion matrix using Laplace mechanism [8] with the ϵ for classifier evaluation. As a result, a DP version of metrics such as accuracy, sensitivity, specificity, precision, recall, and F1-score, among others, can be derived from these four outcomes [34]. Note that we assume that the classifier makes predictions independently. This bounds the sensitivity of the classification process so that its DP can be enforced. The evaluation metric does not need to be limited to the confusion matrix. Any metric that can be computed in a differentially private manner, such as the area under the ROC curve (AUC) or average precision (AP), may be utilized [34].

B. ASSESSING THE PROTOCOLS: MEAN ERROR AND STANDARD DEVIATION

We propose measuring the goodness of the evaluation protocols with a series of n repeated experiments by considering both the systematic and random components of the error. For the ith repetition, let $r_i = p_i - y_i$ denote the (signed) evaluation error between the true classification performance y_i and its prediction p_i made by the protocol. Then, let

$$\mu = \frac{1}{n}\sum_{i=1}^{n} r_i \tag{2}$$

denote the mean error (ME) of the protocol over the n repetitions. Significantly non-zero ME indicates that the predictions made by the protocol differ systematically from the true classification performances, that is, the protocol tends to be either optimistic (positive ME) or pessimistic (negative ME) in its evaluations. The systematic error can be considered as the "bias" of the protocol in the sense that it indicates how much it deviates from zero.

In addition to the above quantified systematic error, we also assess the random error component via the standard deviation:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(r_i - \mu)^2} \tag{3}$$

Together, the systematic and random components form the root mean square error (RMSE) $\sqrt{\mu^2 + \sigma^2}$ of the protocol. All these three quantities, ME, STD and RMSE, are intuitive in the sense that they have the same unit as classification performance measure under consideration.

IV. EXPERIMENTAL SETUP

A. GENERATION METHODS

Marginal-based methods privately select a set of relevant marginals from the real data, add noise to them, and generate synthetic data from these noisy marginals. To ensure differential privacy, noise is added using well-known mechanisms such as the Laplace, Gaussian, and Exponential mechanisms [10], [35]. All five marginal-based generative models, selected for this study have demonstrated good performance with tabular data [10] and are listed as follow:

- **Privbayes**: Introduced by Zhang et al. [20], this method develops a probabilistic representation of the underlying population from which the initial dataset is drawn. Initially, it selects a root node at random, allocating half of the privacy budget to employ the exponential mechanism for identifying optimal child nodes that maximize mutual information with their parent nodes. Once the graph structure is established, the remaining privacy budget is dedicated to measure the essential marginals (low-dimensional conditional distributions) using the Laplace mechanism to adjust the parameters of the Bayesian network. Finally, Privbayes generates synthetic data from the constructed network and the noisy marginals. Privbayes satisfies (ϵ)-DP.
- **MST**: This generative method unfolds in three primary phases. Initially, MST chooses a set of high-quality, low-dimensional marginals from real data, allocating one-third of the privacy budget to this marginal selection process. The selection begins by examining all one-way marginals and identifying pairs of attributes (two-way marginals) that create a maximum spanning tree within the correlation graph of the underlying data. In the next step, the method privately measures the marginals using a Gaussian noise mechanism, consuming two-thirds of the privacy budget. Finally, MST employs a probabilistic graphical model called Private-PGM (introduced in [12]) to estimate the true data distribution based on the selected noisy marginals. It's important to note that MST ensures (ϵ, δ)-differential privacy.
- **MWEM-PGM**: MWEM-PGM is a scalable instantiation of the multiplicative weights exponential mechanism, as originally introduced by Hardt et al. [36]. In its iterative process, MWEM-PGM employs exponential mechanism to identify a marginal query that has not been accurately approximated. Following the selection, the mechanism applies Gaussian noise to measure the marginal, and then it leverages the PGM approximation engine [12] to derive a new estimation of the data distribution. This process involves learning a concise graphical model representation of the data distribution, effectively capturing the noisy measurements while adhering to differential privacy (DP) constraints. The derived distribution serves as the basis for generating synthetic tabular data. It's noteworthy that MWEM-PGM is specifically designed to ensure (ϵ, δ)-differential privacy.
- **AIM**: An Adaptive and Iterative Mechanism for Differentially Private Synthetic Data [10]. AIM is an improved mechanism derived from MWEM-PGM and likewise satisfies (ϵ, δ)-DP. AIM follows the approach of select-measure-generate framework. It incorporates several unique features that enable it to iteratively select the most relevant measurements (Marginals), considering both their significance to the chosen set of workloads, given by the user, to be preserved and their ability to approximate the original data accurately. In the generation phase, Private-PGM [12] is utilized to integrate noisy sets of measurements into a unified generative model.
- **PrivMRF**: privacy-preserving Markov Random Field [30], The fundamental concept behind PrivMRF involves choosing an appropriate set of marginals to construct a Markov random field (MRF). This MRF captures the inter-dependencies among the attributes present in the input data, and subsequently utilizes it to synthesize new data incorporating DP. PrivMRF is designed for (ϵ, δ)-DP.

TABLE 1. Datasets, number of records, feature types and fraction of the positive classes.

Dataset	Records	Categorical	Numeric	positive(%)
IMPROD	500	5	4	48%
Diabetes	768	1	8	35%
Liver	583	2	9	71%
Thyroid	3772	16	6	8%
Cardio	70000	7	5	50%

B. DATASETS

We consider five datasets from the medical domain with varying size and dimensionality summarized in Table 1.

- **Prostate Cancer Dataset (IMPROD)** [37]. The data originates from two clinical trials, NCT01864135 (IMPROD) and NCT02241122 (MULTI-IMPROD), with approvals from the Institutional Review Board. Written informed consent was obtained from all participants. The dataset comprises information on 500 prostate cancer patients, including clinical variables, blood biomarkers, MRI features, and a binary label denoting the patient's condition. It is divided into two groups: 242 patients categorized as high-risk and 258 as benign/low-risk.
- **Diabetes Dataset (Diabetes)** [38] is a well-known benchmark dataset, widely used in data analysis tasks, features a variety of health-related attributes such as blood pressure, BMI (Body Mass Index), and insulin levels. It also includes a binary label to represent the patient's health status,indicating whether the patient has diabetes. The dataset is comprised of 268 patients diagnosed with diabetes and 500 individuals who do not have diabetes.
- **Cardiovascular Disease Dataset (Cardio)** is publicly available at [39]. The dataset encompasses a range of examination metrics like glucose and cholesterol levels, alongside subjective factors such as alcohol consumption and physical activity. It features a binary label that denotes whether or not an individual has cardiovascular disease. Within this dataset, 34,979 patients are identified with cardiovascular disease, whereas 35,021 patients are recorded without the condition.
- **Indian Liver Patient Records (Liver)** is publicly available at [40]. The dataset contains various attributes

or features related to liver health, such as age, gender, total bilirubin levels, direct bilirubin levels, alkaline phosphatase levels, alanine aminotransferase levels, aspartate aminotransferase levels, total proteins, albumin, and more. These attributes are measured from the patients' medical tests and clinical examinations. The class label in this dataset indicates whether a patient has liver disease (416 with liver disease) or not (167 with a healthy liver).

- **Thyroid Disease dataset (Thyroid)** is publicly available at [41]. This dataset is a highly unbalanced dataset with 291 normal thyroid function and 3481 patients with thyroid issues (hypothyroid), including various attributes related to thyroid function, such as hormone levels (e.g., T3, T4, TSH), patient demographics, and additional medical indicators.

C. EXPERIMENTS

To empirically compare the classifier evaluation protocols, a series of experiments were conducted. These experiments involved generating DP-synthetic data from five real tabular medical datasets.

In the experiments, we compared the accuracy and F1-scores provided by the evaluation protocols A or B to the true classification performance calculated on a held-out private test set. Note that this true performance would not be directly accessible to the analyst in either of the protocols. For the DP privacy level we considered the range ϵ : [0.01, 0.1, 1, 3, 5, 7, 9, 10, 15, 20, 50] and $\delta = 1e^{-5}$.

The generative models used in this study only accept categorical features, therefore, all the continuous valued features in each dataset have been discretized into number of bins based on literature and the curator's knowledge. The open source implementation for Privbayes can be found from [42]. The implementations of AIM, MWEM-PGM and MST are those depicted in [43] and PrivMRF [30]. All parameters of the DP-synthetic data generation methods were left to their default values. A detailed list of parameters for each generative method can be found in the implementation details section of the supplementary materials (Table 1S).

We conducted 100 repetitions of the experiments for each protocol and generative method, except for PrivMRF with the Cardio and Thyroid datasets where the experiments were only performed 10 times due to PrivMRF being computationally intensive.

In each repetition the private data was divided randomly into (new) train (80%), and test (20%) sets. In the subsequent step, a differentially private (DP) generative model was fitted to the training set, generating DP-synthetic data. For protocol A, the size of synthetic data sampled matched the combined size of the actual training and testing samples, while for protocol B, the sampled synthetic data equaled the size of the real training set. Ultimately, a classifier was trained using the synthetic training set and evaluated based on either protocol A or B.

To establish an upper bound for accuracy and F1-score achievable with the given datasets, we also present classification metrics for a classifier trained using the original real training data, and tested on the real test data. This would correspond to the case where no privacy is enforced, and the analyst has unrestricted access to the real private dataset.

In the experiments, we tested the evaluation protocols using both widely used linear and non-linear classification methods. The scikit-learn library's implementation of the Random Forest (RF), linear Support Vector Machine (SVM), and K-nearest neighbour algorithms were employed [44]. We used the classifier's default settings, as optimal tuning of the classifier hyperparameters is not required for evaluating the protocols. For SVM classifier regularization parameter was set to C = 1.0, for random forest classifier the number of estimators was 100 and the number of neighbours for KNN was k = 5.

V. RESULTS

Figures 2 and 3 illustrate the classifier evaluation results for protocol A (Syn-Only) and protocol B (Syn-Real), for synthetic data derived from IMPROD, Diabetes, Thyroid and Cardio datasets, utilizing the SVM classifier. First column of each dataset subplot depicts protocol A and second column presents protocol B. Row (a) represents mean value of the classification metric (i.e, F1-score or accuracy) for different epsilon values over 100 repetitions obtained from the synthetic test dataset in protocol A or from the DP confusion matrix calculated from real test data in protocol B. Row (b) represents the mean of the classification metric for both protocols obtained from the real test set without DP. As a baseline comparison, we also report the mean metric of the classifiers trained on the original real training data, and tested on the real test data presented with dashed line in Row (b). Finally, Row (c) presents the mean error (ME) between the estimated and the real test classification metric averaged over 100 repetitions, as well as the standard deviations for the mean errors.

The F1-score results of the downstream classifier are illustrated in Figure 2. For IMPROD dataset, with strict privacy guarantees ($\epsilon \leq 1$), mean errors between the estimated and real test F1-scores in both protocol A and B exhibit significant deviation from zero as presented in Figure 2(C). Further looking at the MEs, protocol A has noticeably higher standard deviation of the errors for most of the generative methods even for $\epsilon \geq 5$. For both evaluation protocols, estimated and real test F1-score values gradually approach the real data setting (dashed line, Mean $F1 = 0.81$) as ϵ increases (see Figure 2 a) and b)). In protocol B, for all synthesizers, from $\epsilon \geq 3$ the mean error approaches zero which indicates that the estimated F1-score are similar to the real test F1-score. For protocol A, it's notable that among all the synthesizers, MST exhibits a consistently high negative error across nearly all $\epsilon \geq 1$ whereas, MWEM-PGM has high positive error for ($\epsilon \leq 1$).

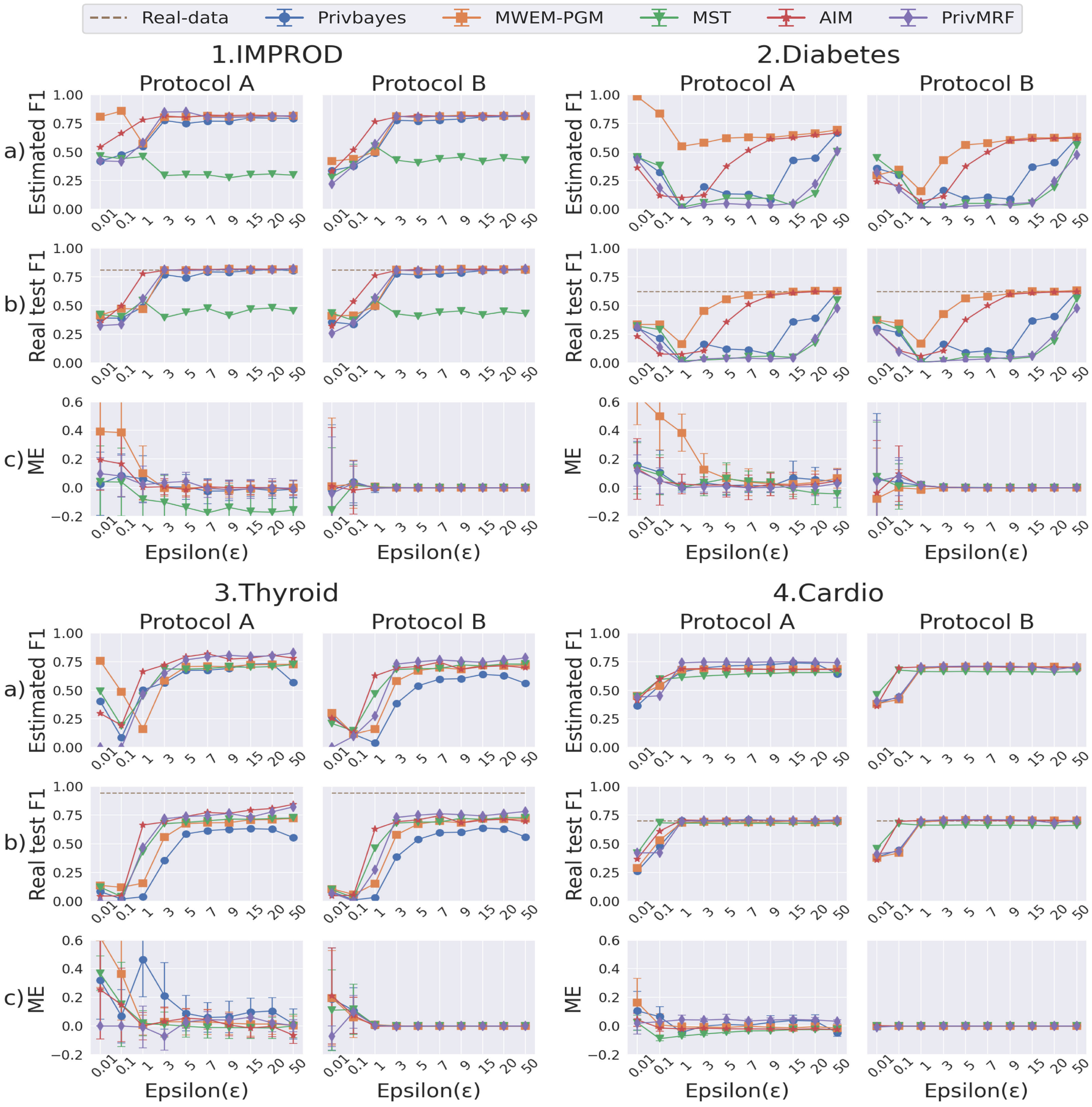

FIGURE 2. F1-score results for SVM classifier on IMPROD, Diabetes, Thyroid and Cardio datasets, with Privbayes, MWEM-PGM, MST, AIM and PrivMRF as DP-generative models.

Concerning Diabetes dataset, F1-score results show that protocol A leads to substantially higher deviation of ME compared to protocol B for all values of epsilon, with a few exceptions when ($\epsilon \leq 1$). ME approaches zero in protocol B where $\epsilon \geq 5$. Synthesizers AIM and MWEM-PGM have the best performance with F1-score for Diabetes dataset, approaching test F1-score for classifiers trained with real data (dashed line in column b, $F1 = 0.62$) when $\epsilon \geq 7$. The MWEM-PGM synthesizer for protocol A produces overly optimistic estimates of F1-score values when compared to the model trained with real data (mean $acc = 0.77$). Consequently, this leads to positive bias in MEs.

In the case of the Thyroid dataset, protocol A exhibits significantly higher errors between estimated and real test F1-scores across for all values of epsilon. For both evaluation protocols where epsilon values are small ($\epsilon \leq 1$) the standard deviation of the errors are high. Estimated and true test F1-scores gradually rise by increasing epsilon although the values do not reach the real data setting F1-score (dashed line, $F1 = 0.94$).

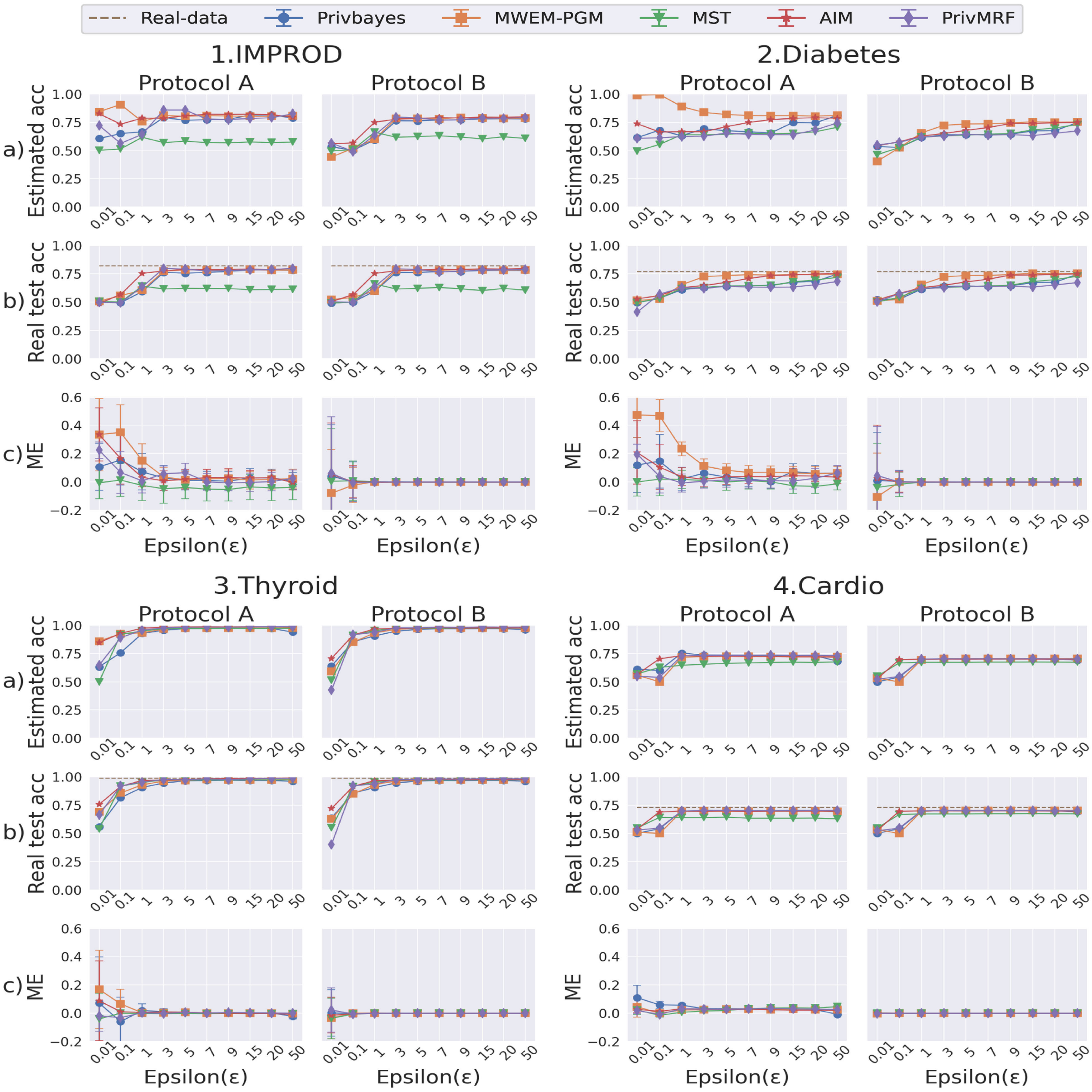

FIGURE 3. Accuracy results for SVM classifier on IMPROD, Diabetes, Thyroid and Cardio datasets, with Privbayes, MWEM-PGM, MST, AIM and PrivMRF as DP-generative models.

For Cardio, estimated F1-scores for the $\epsilon \leq 1$ have smaller standard deviations of the ME for both of the evaluation protocols compared to the other datasets. Protocol B exhibits very low level of error and standard deviation of the errors even for considerably small epsilons ($\epsilon \leq 0.1$). For protocol A, there remains a persistent positive or negative bias of the ME for some methods (i.e, PrivMRF or MST) even for high values of epsilon. Out of all synthesizer AIM has the best performance where even for protocol A ME approaches zero when ($\epsilon \leq 0.1$).

Looking at the accuracy results in Figure 3, for IMPROD and cardio datasets the results are akin to the once's obtained with the F1-score metric. Looking at the IMPROD results, with small epsilons ($\epsilon \leq 1$) estimated accuracies in both protocols show substantial variations, despite protocol B having smaller MEs. For Cardio dataset, for all synthesizers and almost all epsilons MEs and standard deviation of the errors are quite low with exception for Privbayes method with $\epsilon \leq 0.1$ in Protocol A which has slightly higher ME compared to other synthesizers. In both protocols, Estimated

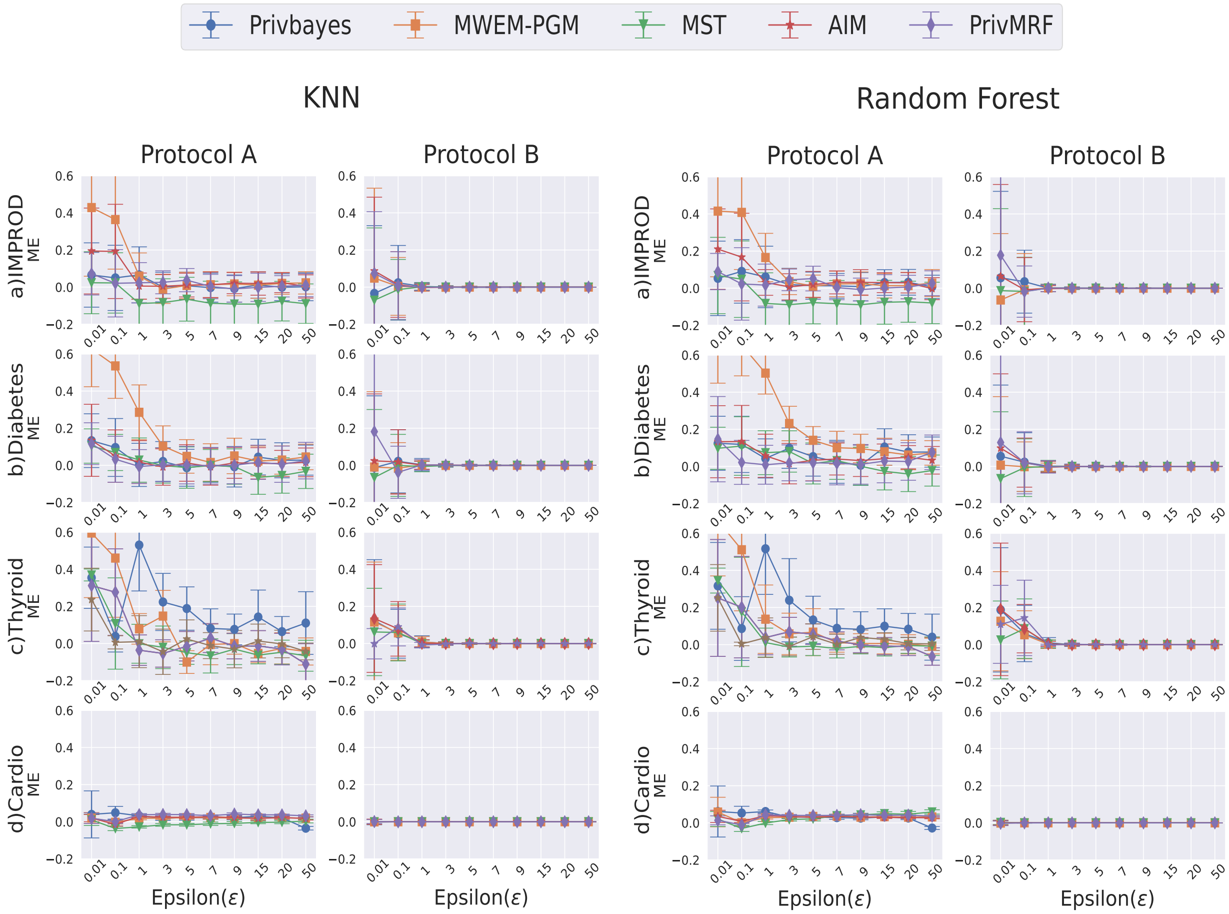

FIGURE 4. Mean errors (ME) results for IMPROD, Diabetes, Thyroid and Cardio using KNN and Random forest classifiers.

and real test accuracy values approach the real data setting (dashed line) with higher values of epsilon.

For Diabetes and Thyroid, mean errors and deviation of the errors are smaller in comparison to the F1-score metric. Looking closely at the ME for both protocols it is clear that protocol B results in lower ME and deviation of the errors. For Diabetes dataset, we note that for PrivMRF, MST and Privbayes, even for higher values of ϵ, the mean accuracies for both protocols do not surpass those of the majority classifier ($acc = 0.65$).

Figure 4 present the summary of the mean errors between estimated and F1-scores for both evaluation protocols utilizing Random forest and KNN classifiers. For each classifier, the first and second columns represents MEs for protocol A and B respectively and each row depicts the datasets. Error bars indicate the standard deviation of the difference between estimated and true classification performance in each evaluation protocol under different epsilon values. Comparing the MEs and the standard deviation of the errors between These two classifiers and the SVM classifier it is evident that the trends in all three classifier are similar given each dataset and synthesizer. The detailed results for Liver dataset can be found in the supplementary materials.

VI. DISCUSSION

There is a growing interest in utilizing DP-synthetic health data for purposes such as training machine learning models or statistical inference. However, it is crucial to ensure the validity of conclusions drawn from DP-synthetic data. We introduced two protocols that enable a data analyst to evaluate the performance of a classifier trained on DP-synthetic data. Further, we conducted an empirical investigation measuring the evaluation error of these protocols. The same protocols could also be applied to various model selection tasks such as feature selection, hyperparameter tuning, or choosing the best classification algorithm among alternatives. DP model selection is treated in more detail in [23].

Considering all five medical datasets tested in this study, at high levels of enforced privacy ($\epsilon \leq 1$) both evaluation protocols tend to be quite unreliable with large standard deviation of the error especially in case of protocol A, where also the mean error (ME) is often high. Increased privacy level correlates with a decrease in the quality of the generated synthetic data, consequently leading to lower F1-score and accuracy values on both real and DP-synthetic test datasets. Generally a minimum privacy budget of $\epsilon = 1$ was required to ensure reliable performance for both the trained classifier and the evaluation protocols.

When privacy level is less strict ($3 \leq \epsilon \leq 9$) for protocol B, the F1-scores and accuracies have near zero MEs, accompanied by low standard deviation of the errors. Mean errors for protocol A also exhibit a distinct reduction compared to $\epsilon \leq 1$, particularly for classification accuracy, but there are still numerous cases with relatively high ME and standard deviation (e.g. MST with IMPROD and Diabetes Figure 3).

When $\epsilon \geq 15$, the synthetic data generated exhibits a close resemblance to the characteristics of real data. This is observable in both protocols, where the resulting mean errors and, specially, the standard deviations approach zero for classification accuracy. However, for F1-scores the mean error and standard deviation still remain clearly higher for protocol A than for protocol B.

Based on the MEs, it is noticeable that with protocol B, the estimated F1-scores for the balanced datasets such as IMPROD and Cardio closely align with the actual F1-scores obtained from the real held-out test set when $\epsilon \geq 3$. However, for Diabetes and Thyroid, where the proportion of positive classes is low, a higher privacy budget is necessary to achieve scores comparable to the real test results.

A distinct reduction in variability, measured by the standard deviation of errors, is evident in cases where the sample size is sufficiently large, as observed in the Cardio dataset. For both F1-scores and accuracy analysis, the standard deviations across all epsilon values are notably smaller in comparison to other four datasets.

Among the synthesizers used in this study, AIM, MWEM-PGM and privMRF performed better in terms of downstream classifier, which is in line with results reported in other studies based on DP-synthetic tabular data generation [10], [15], [22]. In many cases with protocol A results for MWEM-PGM showed substantial overoptimistic bias. It appears that AIM in comparison to other generative methods is better able to approximate underlying distribution of the real data whilst consuming less privacy budget as claimed in [10].

The ME results were quite similar for all the considered classifiers illustrated in Figures 4 for F1 score and in supplementary materials Figure 3 for accuracy. Based on our experimental results, the performance of the two evaluation protocols is not significantly influenced by the classification method employed.

VII. CONCLUSION

In summary, protocol B tends to provide more reliable classifier performance evaluations than protocol A. This is especially important in medical data analysis, where for example taking into use a diagnostic classifier based on severely overoptimistic performance evaluation could have significant real world consequences. However, these advantages come with the price of requiring the data curator to implement this additional protocol that allows testing trained classifiers on real test data. Further, when multiple users access the same private data, the privacy budget must be incrementally increased with each evaluation conducted using Syn-Real (protocol B). In contrast, Protocol A does not entail any additional privacy costs once the synthetic data has been released.

A practical challenge for the curator who produces the synthetic data to the analyst is the required computational infrastructure. Especially deep learning based synthetic data generation methods can require specialize GPU hardware to be efficient, which may not be readily available in typical sensitive data storage environments. Further, with protocol B, the curator needs to validate whether the classifier sent by the data analyst is trustworthy, so that it can be used behind the privacy barrier. Obviously, the curator can not accept and run arbitrary code. Instead, one needs to standardize the types of models that can be used, such as the coefficients of a linear model.

Our study has several limitations to be addressed in future work. One limitation is that all marginal based methods used in this study require the input data to be discrete values. How the continuous data is transformed into discrete values can have significant impact on the quality of the generated DP-synthetic data [10], [20]. Secondly, we have fixed the hyperparameters of each DP-synthesization method to be the default ones. Experimenting with more varied hyperparameter values could lead to improved performance for some of the methods, though on the other hand additional privacy budget would need to be allocated for model selection for the generator. Furthermore, the impact of the size of the generated synthetic data has on the downstream classifier results could be further investigated.

ACKNOWLEDGMENT

The authors would like to thank Ivan Jambor and collaborators for their contribution in providing the prostate cancer datasets. This work has received funding from Business Finland and European Union's Horizon Europe research and innovation programme. Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them.

REFERENCES

[1] C. Kuner, L. A. Bygrave, C. Docksey, L. Drechsler, and L. Tosoni, "The EU general data protection regulation: A commentary/update of selected articles," May 2021.

[2] K. El Emam, S. Rodgers, and B. Malin, "Anonymising and sharing individual patient data," *BMJ*, vol. 350, no. 1, pp. h1139–h1142, Mar. 2015.

[3] M. Giuffrè and D. L. Shung, "Harnessing the power of synthetic data in healthcare: Innovation, application, and privacy," *npj Digit. Med.*, vol. 6, no. 1, p. 186, Oct. 2023.

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

[5] T. Stadler, B. Oprisanu, and C. Troncoso, "Synthetic data–anonymisation groundhog day," in *Proc. 31st USENIX Secur. Symp.*, 2022, pp. 1451–1468.

[6] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 3–18.

[7] D. Chen, N. Yu, Y. Zhang, and M. Fritz, "GAN-leaks: A taxonomy of membership inference attacks against generative models," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2020, pp. 343–362.

[8] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. Theory Cryptogr. Conf.*, 2006, pp. 265–284.

[9] K. Donhauser, J. Abad, N. Hulkund, and F. Yang, "Privacy-preserving data release leveraging optimal transport and particle gradient descent," 2024, *arXiv:2401.17823*.

[10] R. McKenna, B. Mullins, D. Sheldon, and G. Miklau, "AIM: An adaptive and iterative mechanism for differentially private synthetic data," 2022, *arXiv:2201.12677*.

[11] R. Castellon, A. Gopal, B. Bloniarz, and D. Rosenberg, "DP-TBART: A transformer-based autoregressive model for differentially private tabular data generation," 2023, *arXiv:2307.10430*.

[12] R. McKenna, G. Miklau, and D. Sheldon, "Winning the NIST contest: A scalable and general approach to differentially private synthetic data," *J. Privacy Confidentiality*, vol. 11, no. 3, pp. 1–20, Dec. 2021.

[13] H. Ping, J. Stoyanovich, and B. Howe, "DataSynthesizer: Privacy-preserving synthetic datasets," in *Proc. 29th Int. Conf. Sci. Stat. Database Manage.*, Jun. 2017, pp. 1–5.

[14] C. Dwork, N. Kohli, and D. Mulligan, "Differential privacy in practice: Expose your epsilons!" *J. Privacy Confidentiality*, vol. 9, no. 2, pp. 1–29, Oct. 2019.

[15] M. Pereira, M. Kshirsagar, S. Mukherjee, R. Dodhia, J. Lavista Ferres, and R. de Sousa, "Assessment of differentially private synthetic data for utility and fairness in end-to-end machine learning pipelines for tabular data," 2023, *arXiv:2310.19250*.

[16] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov, "Differential privacy has disparate impact on model accuracy," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–19.

[17] J. Ullman and S. Vadhan, "PCPs and the hardness of generating synthetic data," *J. Cryptol.*, vol. 33, no. 4, pp. 2078–2112, Oct. 2020.

[18] M. Boedihardjo, T. Strohmer, and R. Vershynin, "Covariance's loss is privacy's gain: Computationally efficient, private and accurate synthetic data," *Found. Comput. Math.*, vol. 1, pp. 1–48, Jun. 2022.

[19] O. Giles, K. Hosseini, G. Mingas, O. Strickson, L. Bowler, C. Rangel Smith, H. Wilde, J. Ning Lim, B. Mateen, K. Amarasinghe, R. Ghani, A. Heppenstall, N. Lomax, N. Malleson, M. O'Reilly, and S. Vollmerteke, "Faking feature importance: A cautionary tale on the use of differentially-private synthetic data," 2022, *arXiv:2203.01363*.

[20] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao, "PrivBayes: Private data release via Bayesian networks," *ACM Trans. Database Syst.*, vol. 42, pp. 1–26, Oct. 2017.

[21] A. Torfi, E. A. Fox, and C. K. Reddy, "Differentially private synthetic medical data generation using convolutional GANs," *Inf. Sci.*, vol. 586, pp. 485–500, Mar. 2022.

[22] Y. Tao, R. McKenna, M. Hay, A. Machanavajjhala, and G. Miklau, "Benchmarking differentially private synthetic data generation algorithms," 2021, *arXiv:2112.09238*.

[23] P. Movahedi, V. Nieminen, I. M. Perez, T. Pahikkala, and A. Airola, "Evaluating classifiers trained on differentially private synthetic health data," in *Proc. IEEE 36th Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2023, pp. 748–753.

[24] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations Trends Theor. Comput. Sci.*, vol. 9, nos. 3–4, pp. 211–407, 2014.

[25] C. Dwork, "A firm foundation for private data analysis," *Commun. ACM*, vol. 54, no. 1, pp. 86–95, Jan. 2011.

[26] C. Arnold and M. Neunhoeffer, "Really useful synthetic data - a framework to evaluate the quality of differentially private synthetic data," 2020, *arXiv:2004.07740*.

[27] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2016, pp. 308–318.

[28] R. McKenna, G. Miklau, and D. Sheldon. (2021). *Private-PGM*. [Online]. Available: https://github.com/ryan112358/private-pgm

[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 11–111.

[30] K. Cai, X. Lei, J. Wei, and X. Xiao, "Data synthesis via differentially private Markov random fields," *Proc. VLDB Endowment*, vol. 14, no. 11, pp. 2190–2202, Jul. 2021.

[31] K. Banachewicz, L. Massaron, and A. Goldbloom, *The Kaggle Book: Data Analysis and Machine Learning for Competitive Data Science*. Birmingham, U.K.: Packt Publishing Ltd, 2022.

[32] P. Meyer and J. Saez-Rodriguez, "Advances in systems biology modeling: 10 years of crowdsourcing DREAM challenges," *Cell Syst.*, vol. 12, no. 6, pp. 636–653, Jun. 2021.

[33] F. D. McSherry, "Privacy integrated queries: An extensible platform for privacy-preserving data analysis," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Jun. 2009, pp. 19–30.

[34] K. Boyd, E. Lantz, and D. Page, "Differential privacy for classifier evaluation," in *Proc. 8th ACM Workshop Artif. Intell. Secur.*, Oct. 2015, pp. 15–23.

[35] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *Proc. 48th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, Oct. 2007, pp. 94–103.

[36] M. Hardt, K. Ligett, and F. McSherry, "A simple and practical algorithm for differentially private data release," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1–26.

[37] I. Jambor et al., "Validation of IMPROD biparametric MRI in men with clinically suspected prostate cancer: A prospective multi-institutional trial," *PLOS Med.*, vol. 16, no. 6, Jun. 2019, Art. no. e1002813.

[38] J. W. Smith, J. E. Everhart, W. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in *Proc. Annu. Symp. Comput. Appl. Med. Care*, 1988, p. 261.

[39] S. Ulianova. (2023). *Kaggle Cardiovascular Disease Dataset*. [Online]. Available: https://www.kaggle.com/sulianova

[40] B. Ramana and N. Venkateswarlu, "ILPD (Indian Liver Patient Dataset)," UCI Mach. Learn. Repository, 2012, doi: 10.24432/C5D02C.

[41] R. Quinlan, "Thyroid disease," UCI Mach. Learn. Repository, 1987, doi: 10.24432/C5D010.

[42] D. Zhang, R. McKenna, I. Kotsogiannis, M. Hay, A. Machanavajjhala, and G. Miklau, "EKTELO: A framework for defining differentially-private computations," in *Proc. Int. Conf. Manage. Data*, May 2018, pp. 115–130.

[43] R. McKenna, D. Sheldon, and G. Miklau, "Graphical-model based estimation and inference for differential privacy," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 4435–4444.

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

PARISA MOVAHEDI (Member, IEEE) received the M.Sc. (Tech.) degree in information technology (computer science) from the Department of Information Technology and Communication Systems, University of Turku, Finland, where she is currently pursuing the Ph.D. degree with the Department of Computing. Her current research interests include algorithms and applications of machine learning and data analysis include privacy-preserving machine learning in medical field.

VALTTERI NIEMINEN received the M.Sc. (Tech.) degree in data analytics (computer science) from the Department of Computing, University of Turku, Finland, where he is currently pursuing the Ph.D. degree. He is also employed as a Data-Analyst with Helsinki University Hospital, Finland, in the iCAN Digital Precision Cancer Medicine Project. His research interests include privacy-preserving machine learning, precision medicine, and the use of real-world data in medicine.

ILEANA MONTOYA PEREZ received the M.Sc. degree in information technology (computer science) from the University of Turku, Finland, where she is currently pursuing the Ph.D. degree with the Department of Computing. Over the past years, she has been researching and working on the technical aspects of privacy preservation in medical and health data. Her research interests include data analysis and machine learning methods addressing limited data availability.

HIBA DAAFANE is currently pursuing the M.Sc. (Tech.) degree in data analytics (information and communication technology) with the Department of Computing, University of Turku, Finland. Her research interests include privacy-preserving techniques in machine learning and the applications of artificial intelligence in understanding neurological patterns and behavioral analysis.

DISHANT SUKHWAL is currently pursuing the M.Sc. (Tech.) degree in data analytics (computer science) with the Department of Computing, University of Turku, Finland. He is currently employed as a Software Engineer Trainee with Trimble Inc. His research interests include synthetic data generation, machine learning, natural language processing, and evaluation of machine learning methods.

TAPIO PAHIKKALA received the Ph.D. degree from the University of Turku, Finland, in 2008. He is currently a Professor in computer science with the Department of Computing, University of Turku. His research interests include the theory and algorithmics of machine learning, data analysis, and artificial intelligence, their performance evaluation especially with resampling methods and their applications in various different fields.

ANTTI AIROLA received the Ph.D. degree from the University of Turku, Finland, in 2011. He is currently an Associate Professor with the Department of Computing, University of Turku. His research interests include machine learning in health data analytics, emphasizing privacy preservation, evaluation of medical AI models, and healthcare AI applications. His work contributes to the development of methods for secure and effective use of AI in healthcare, addressing key challenges in patient data privacy, and the reliability of AI technologies in medical settings.