Evaluating Classifiers Trained on Differentially Private Synthetic Health Data

Parisa Movahedi
Department of Computing
University of Turku
Turku, Finland
parmov@utu.fi

Valtteri Nieminen
Department of Computing
University of Turku
Turku, Finland
valtteri.a.nieminen@utu.fi

Ileana Montoya Perez
Department of Computing
University of Turku
Turku, Finland
iimope@utu.fi

Tapio Pahikkala
Department of Computing
University of Turku
Turku, Finland
aatapa@utu.fi

Antti Airola
Department of Computing
University of Turku
Turku, Finland
ajairo@utu.fi

***Abstract*—The release of differentially private (DP) synthetic data has been proposed as a solution to sharing sensitive individual-level medical data for statistical analysis and machine learning model development. The approach holds promise to generate realistic data that preserves many of the statistical properties of the original data while giving privacy guarantees that bound the risk of leaking any sensitive information about the individuals in the data. However, evaluating the generalization of machine learning models trained on DP-synthetic data remains an open question. A model selected based on its accuracy on synthetic data does not necessarily generalize well to real-world data, leading to poor results and incorrect insights. In this study, we experimentally compare two different protocols for model evaluation and hyperparameter selection for classifiers trained on DP-synthetic medical data. In the first protocol, we use only synthetic data for model selection and final evaluation of selected model, whereas in the second one, we assume limited DP access to a private real validation and test set held by the data curator. Our results provide novel insights into the practical feasibility and utility of different evaluation protocols for classifiers trained on DP synthetic data based on a comprehensive empirical study.**

I. Introduction

Sharing individual-level medical data is challenging due to privacy concerns and strict regulation like GDPR [1]. These factors limit the availability and usability of medical data for tasks such as statistical analysis and machine learning. Synthetic data generation has been proposed as a method to overcome these difficulties. Unfortunately, it has been repeatedly shown, that synthetic data is not inherently privacy preserving [2], [3]. It is well known that the outputs of machine learning (ML) models [4], can leak information about their training data, and generative models used to create synthetic data are no different in this respect. 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. [5], which quantifies and enforces privacy by injecting statistical noise to results of computations done on sensitive data. It provides a probabilistic guarantee on how much information can be learned about any individual from the computations, that could not be inferred if the individual was not present in the data.

In recent years, many different approaches to generate DP synthetic data have been proposed [6]–[11]. One standard way of measuring the quality of DP-synthetic data and the method its been generated with, is to train what is called a *downstream* model on the synthetic data and to evaluate it against a real dataset. In works conducting such evaluations, the focus has been on the synthetic data generation methods rather than questions regarding the downstream model selection and evaluation [8], [12], [13].

There is no clear consensus on how model selection and evaluation for models trained on DP-synthetic data should be set up between the data curator and the analyst. The question of whether one can or should use only synthetic data for model selection and evaluation is highly non-trivial. This is because enforcing privacy through DP always comes with a cost, and DP synthetic data is by definition, a distorted version of the real data and does not yield the same results as the original would [14] [5], [15]. Synthetic data are an approximation of the real data that may not capture all aspects of that data. Consequently, a setup where the model selection and final testing of a downstream model is done using only on synthetic data may lead to worse results and incorrect insights compared to the case where some real data can be used. This work aims to address the aforementioned gap in the literature concerning downstream model selection and evaluation with DP-synthetic data. Two data access protocols are considered:

1) In protocol A (Syn-Only) the analyst has access only to DP-synthetic data to validate and test the model.
2) In protocol B (Syn-Real) a DP-query is allowed to be performed on private held-out data from the curator's side. This data is never shown to the analyst, but analyst is able to submit the models trained on DP-synthetic data

This is the author's version of the work that has been published in 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS) | 979-8-3503-1224-9/23/ $31.00 ©2023 IEEE | DOI: https://doi.org/10.1109/CBMS58004.2023.00313.

to the curator and receive DP-information about the best performing downstream model and its performance on real private test set.

In this study we conduct comprehensive empirical evaluations on three medical datasets, based on which DP-synthetic is generated to investigate the pros and cons of utilizing each of these protocols in terms of privacy-utility trade-off.

II. Methods

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:

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

where ϵ is a upper bound on the privacy loss given by the user, which affects the amount of noise added to the algorithm and δ is a small constant representing some extremely unlikely event, where the DP constraints do not hold [16]. (ϵ)-DP is a special case of (ϵ, δ)-DP where $\delta = 0$. A smaller value of ϵ means a stronger guarantee. Although a single acceptable value for the privacy budget ϵ can not be given as it depends on the context, in the literature, values of $\epsilon \leq 1$ have been considered to provide strong protection [17], and depending on the type of data and task, values of $\epsilon \leq 10$ have been seen to still result in meaningful guarantees [18].

B. DP synthetic data generation methods

The objective of DP-synthetic data generation is to produce synthetic data that resembles the original data and provides similar results in analyses, while providing privacy guarantees. This is achieved by generating synthetic records via a DP model trained on the original data. The utility of the synthetic data for a given task relies on the ability of the model to accurately capture the statistical properties and interrelationships of features present in the original data. The three marginal-based generative models, selected for this study have demonstrated good performance with tabular data [13], [19] and are listed as follow:

- **Privbayes**: Proposed by Zhang et al. [8] builds a probabilistic model of the underlying population from which the original data was sampled. Half of the privacy budget ϵ is used to learn a Bayesian network structure that captures the dependencies in real data. The remaining budget is used to measure the necessary marginals for learning the Bayesian network parameters. Finally, Privbayes generates synthetic data from the constructed network and the noisy marginals. Privbayes satisfies (ϵ)-DP.
- **MST**: This generative method consists of three steps. First, MST selects a set of high quality low-dimensional marginals from the real data, where $1/3$ of the privacy budget (ϵ) is devoted towards marginal selection. Secondly, it measures the marginals privately via a noise mechanism spending $2/3$ of the privacy budget. In the last step, MST uses a probabilistic graphical model, Private-PGM [20] to estimate the real data distribution from the selected noisy marginals. MST satisfies (ϵ, δ)-DP.
- **MWEM-PGM**: is a scalable instantiation of the multiplicative weights exponential mechanism introduced by Hardt et al. [21]. MWEM-PGM uses the PGM approximation engine to learn a compact graphical model representation of data distribution while satisfying DP [22]. MWEM-PGM is designed for (ϵ, δ)-DP.

C. Data access protocols

Two different data access protocols have been defined for the data analyst to validate a desired downstream model using either real or synthetic data.

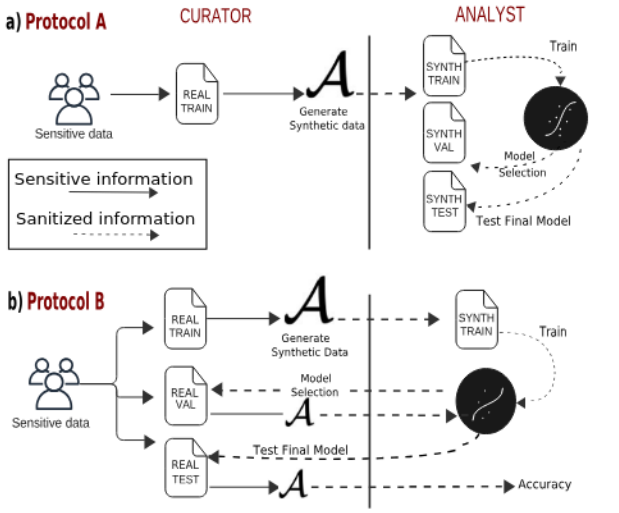

Fig. 1. Protocol A (top): the data analyst has access only to the DP-synthetic data to train, validate and test the model. Protocol B (bottom): The data analyst trains the model with DP-synthetic data, but the validation and test are done on the curator site.

- **Protocol A: Syn-Only**. As shown in Figure 1 (a), a generative model is trained on the train set in the curator site spending all the pre-specified privacy budget and DP-synthetic data is sampled from the trained generative model to be released to the analyst. The data analyst splits the DP-synthetic data for training, validating and testing a downstream classifier.
- **Protocol B: Syn-Real**. In this setting (depicted in Figure 1(b)) the data curator splits the sensitive real data to train, validation and test sets. A DP-generative model is trained on the training data partition, while consuming a portion of the specified privacy budget, and DP-synthetic data is sampled and released to the analyst to train a downstream classifier. Model validation for the trained classifier is done on the curator side, and the curator uses the validation set from private real data and a part of the remaining privacy budget to validate the trained classifier. Selection of classifiers' hyperparameters are

done in a DP-manner using the exponential mechanism [23]. The best DP-hyperparameter set is released back to the data analyst as depicted in Figure 1. Lastly, the analyst sends the final trained model back to the curator to be tested on the private test set. The accuracy of the final model is released to the analyst DP-wise using the Laplace mechanism [5] utilizing the remaining privacy budget.

III. Experiments

A series of experiments, where DP-synthetic data was generated based on real tabular medical data, were conducted to empirically compare the data access protocols for downstream model selection and evaluation.

A. Datasets

We consider three datasets from the medical domain (see Table I).

1. The Prostate Cancer Dataset (IMPROD) [24] originates from two clinical trials, NCT01864135 (IMPROD) and NCT02241122 (MULTI-IMPROD), both approved by the Institutional Review Board and all patients provided written informed consent. This dataset consists of 500 prostate cancer patients, including their clinical variables, blood biomarkers, MRI features, and a binary label indicating the patient's condition (242 in the high-risk group and 258 in the benign/low-risk group).

2. The Diabetes Dataset (Diabetes) [25] is a commonly used benchmark dataset in machine learning. It includes various measured health features such as blood pressure, BMI and insulin levels In addition, a binary label indicates the patient's condition. The dataset contains 268 diabetic patients and 500 non-diabetics.

3. Kaggle Cardiovascular Disease Dataset (Cardio) is publicly available at [26]. The dataset contains examination features such as glucose and cholesterol, subjective features like alcohol intake and physical activity and a binary label indicating the presence or absence of cardiovascular disease. In this study 10000 samples were used which have been chosen randomly using a stratified selection to preserve the proportion of the labels. The number of patients having cardiovascular disease is 4995 while 5005 patients do not have the disease.

TABLE I
Data Sets, 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%
Cardio	10000	7	5	50%

B. Experimental setting

Model selection and final model evaluation is done following either protocol A or B. Also, the true classification accuracy of the final model is calculated with a held-out private test set. The values of ϵ utilized are reported in Table II. In protocol A, the whole privacy budget (ϵ) is used to generate the synthetic data, whereas in protocol B, privacy budget is divided between data generation ($1/2\,\epsilon$), model selection ($1/4\,\epsilon$) and testing ($1/4\,\epsilon$). For MST and MWEM-PGM, the value of δ was set to $1e^{-5}$. The generative models used in this study only accept categorical features, therefore, all the continuous valued features in each data set have been discretized into number of bins based on literature and the curator's knowledge. The open source implementation for Privbayes used can be found from [27]. The implementations of MWEM-PGM and MST are those depicted in [7]. All parameters of the DP synthetic data generation methods were left to their default values.

For each of the protocols we have repeated the experiments 100 times. The test median accuracy of these repetitions are reported for the final chosen classifier.

In each round of iterations the private data is divided randomly into (new) train (60%), validation (20%) and test (20%) sets. Secondly, a DP-generative model is trained on the train set and DP-synthetic data is sampled. In protocol A for each synthetic train, validation and test sets, 1000 data points are sampled from the trained DP-generator. For protocol B, 1000 data points are sampled for synthetic train set.

A number of classification models corresponding to different hyperparameter choices are trained on the released DP-synthetic data. The final chosen classifier is selected either based on having highest accuracy on synthetic validation data (protocol A) or a DP-query about which model has highest accuracy on real validation data (protocol B).

As a point of comparison representing an upper bound on the accuracy achievable with these data sets, we also report the classification accuracy of models trained using the original private real training data.

C. Downstream classifier

In the experiments, we tested the model selection and evaluation protocols for a widely used non-linear classification method, regularized least-squares (RLS) [28]. For RLS with RBF kernel we select kernel width γ and regularization parameter λ using similar grid as suggested by [29]. Major advantage of using RLS is efficiency of tuning the regularization parameter. The range of tested hyper-parameters is reported in Table II. The RLS implementation is from the RLScore library [30].

IV. Results

Figure 2 represents a single repetition of the experiment for protocol B, visualizing the classifier's hyperparameter search result on Cardio data, with $\epsilon = 9.0$ and MWEM-PGM generative model. The heatmap visualizes the parameter grid search for the RLS classifier, plotting the validation set accuracy for classifiers trained with each possible combination of the hyperparameters. The model is chosen among the tested hyperparameter combinations using exponential noise mechanism [16], The amount of noise used for the selection depends on the amount of epsilon reserved for model selection.

TABLE II
LIST OF HYPERPARAMETERS SELECTED FOR THE EXPERIMENTS.

Hyperparameters	Ranges
Epsilon (ϵ)	[0.01, 0.1, 1.0, 3.0, 5.0, 7.0, 9.0, 15.0, 20.0, 50.0]
delta (δ)	$1e^{-5}$
Kernel width γ (RLScore)	$2^{-15}, ..., 2^{15}$
Regularization parameter λ (RLScore)	$2^{-15}, ..., 2^{15}$

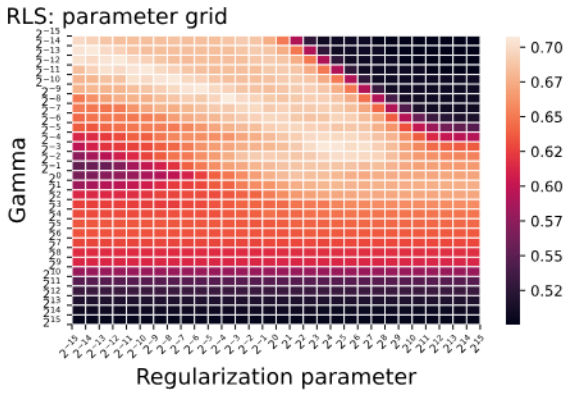

Fig. 2. Parameter grid of one run of the RLS on Cardio data. Classifiers trained on synthetic data (MWEM-PGM with $\epsilon = 9.0.0$) with different hyperparameter combinations are evaluated on real validation data. In this particular case, hyperparameter set [0.0019, 0.063] (acc=0.71) would yield highest accuracy for RLS. However, if too small ϵ is used for model selection, protocol B may make suboptimal choices for hyperparameters.

If there is no DP ($\epsilon \to \infty$) the combination of hyperparameters that results in the best validation accuracy is selected as the optimal set of hyperparameters. The lower the value of ϵ the more random noise is injected to selection, and when ϵ approaches zero the hyperparameters are chosen from a uniform distribution over the parameter grid.

Figure 3 represents the results for the prostate cancer dataset (IMPROD) where the first column depicts protocol A (Syn-Only) and second column presents protocol B (Syn-Real). Figure 3 (a) represents median accuracies over 100 runs obtained from the synthetic test data set in protocol A and the DP-wise accuracy of the private test set in protocol B. Figure 3 (b) represents the median accuracies for both protocols obtained from the real private test set without DP. Finally, Figure 3 (c) presents the mean absolute error (MAE) between the estimated and the real test accuracy averaged over all the runs of each protocol. Estimated DP-accuracies in protocol B with small epsilons ($\epsilon \leq 1$) show high levels of variability. Both estimated and real test accuracy values gradually approach the real train data setting (median $acc = 0.82$) as ϵ increases. Looking at the MAEs, protocol A seems to be optimistically biased even for $\epsilon \geq 5$, whereas in protocol B for $\epsilon \geq 5$ the error approaches zero which indicates that the estimated DP-accuracies are similar to the true accuracies. Out of all the data synthesizers, the MST method appears to yield the smallest accuracies.

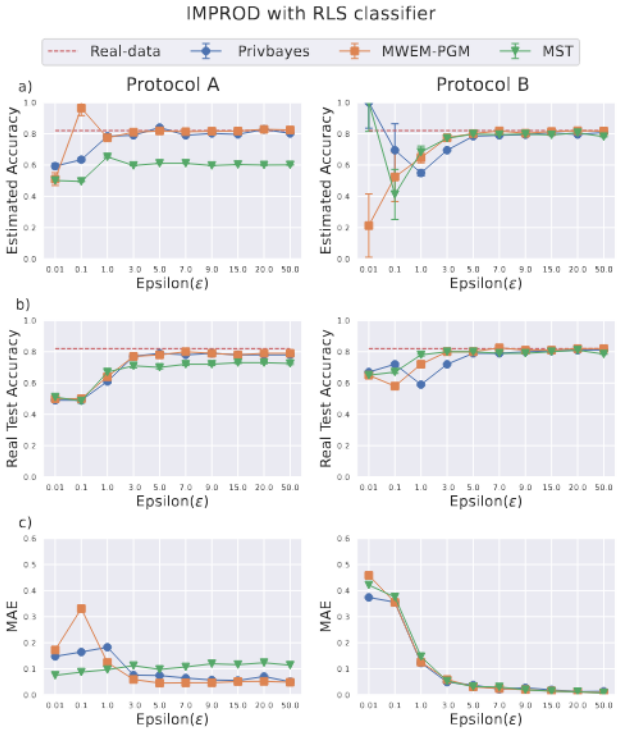

Fig. 3. IMPROD, classification accuracy obtained from RLS classifier in protocol A and B with three DP-generative models. The dashed line represents the accuracy of model trained on real data. a) Median estimated accuracies. b) Median accuracies obtained from real test data without DP. c) Mean absolute error between estimated and true accuracies.

Figure 4 presents the results of Diabetes dataset. The findings are similar to the ones obtained from IMPROD data set, although with Diabetes data set the MWEM-PGM synthesizer for protocol A has overoptimistic estimated DP-accuracies compared to the model trained with real data showing with the dashed line and the median $acc = 0.75$ which results in higher values of the MAE. It should be noted that for MST and Privbayes with all ($\epsilon \leq 50$) the median accuracies for both protocols are no better than the majority classifier ($acc = 0.65$)

Figure 5 presents the results of RLS classification for Cardio dataset where the median accuracy of the model trained with real data is $acc = 0.72$ (dashed line). The results are akin to those depicted in Figures 4, 3 except for the estimated DP-accuracies of smaller epsilons which show less variability compared to the other two data sets, and therefore smaller values of MAE, resulting in less estimation bias.

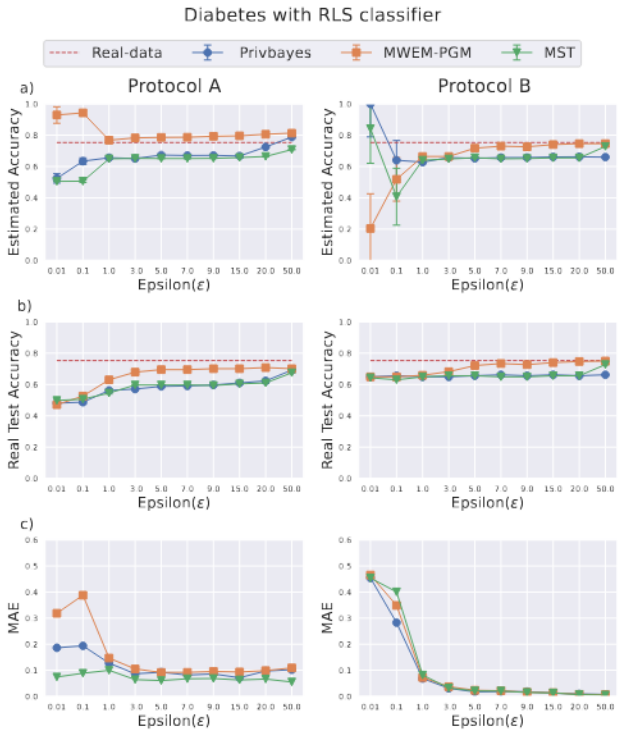

Fig. 4. Diabates, classification accuracy obtained from RLS classifier in protocol A and B with three DP-generative models. The dashed line represents the accuracy of model trained on real data. a) Median estimated DP accuracies. b) Median accuracies obtained from real test data without DP. c) Mean absolute error between estimated and true accuracies.

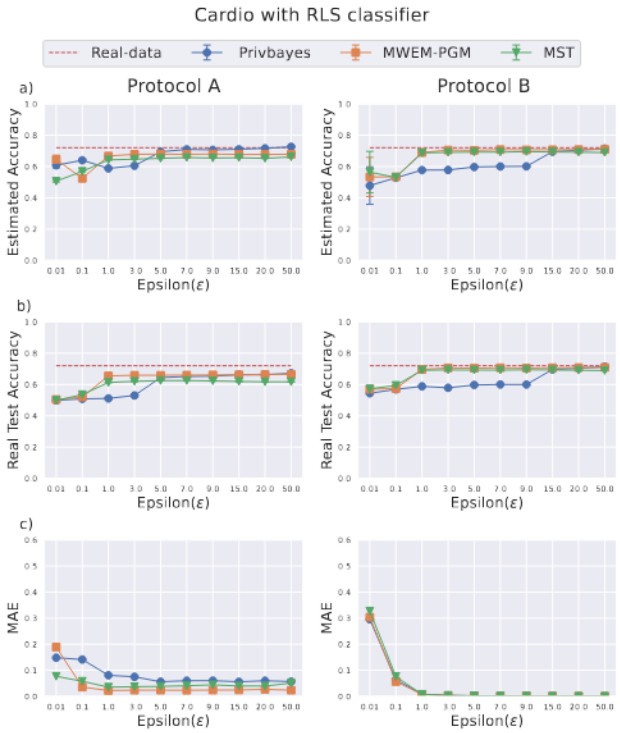

Fig. 5. Cardio, classification accuracy obtained from RLS classifier in protocol A and B with three DP-generative models. The dashed line represents the accuracy of model trained on real data. a) Median estimated DP accuracies. b) Median accuracies obtained from real test data without DP. c) Mean absolute error between estimated and true accuracies.

V. Discussion and Conclusion

There is an increased interest in using DP-synthetic health data to train machine learning models, test hypotheses, and develop predictive algorithms. However, it's important to ensure, that conclusions drawn from DP-synthetic data are valid. In this study we presented an empirical investigation of downstream model validation and hyperparameter selection using DP-synthetic data. We presented two access protocols in which a data analyst can perform model selection and performance evaluation of the final chosen classifier. In this study we have considered tuning the hyperparameters of a single classifier but same protocols are applicable on various model selection tasks such as feature selection or choosing among alternative classification methods.

Our results indicate that when the value of epsilon is sufficiently high (e.g. $\epsilon \geq 5$) for both protocols A and B, the selected final models yield true test accuracies approaching to those obtained when only real private data is used for training, validating and testing the downstream model without DP (dashed lines row (b) in : Figures 3, 4, 5).

Additionally, looking at the MAEs, with protocol B the estimated accuracies are close to the real test accuracies obtained with the real held-out test set when $\epsilon \geq 5$. In contrast, when protocol A is used, there seems to be rather a noticeable bias in the accuracy values when compared to the real test accuracies which is particularly visible with the smaller sized data sets, IMPROD and Diabetes (Figures 3, 4 depicted in row (c) where the mean absolute error of estimated and real test accuracies are shown. For instance, in the Diabetes data set when $\epsilon = 5$ the MAE values is approximately 0.1 indicating a 10% units difference in the accuracy. It is worth noting that even slight differences in the estimated and real test accuracy values can lead to misleading and incorrect conclusions, especially in the medical domain.

Looking at all three tested data sets, when the privacy budget is sufficiently large (e.g. $\epsilon \geq 5$) protocol B allowed selection of good hyperparameter values and the accuracy estimates were close to those obtained without DP noise added. However, it is important to mention that if there were multiple users accessing the same private data, then every time a model is selected or a final evaluation is done based on Syn-Real (Protocol B), the privacy budget has to be cumulatively increased. In contrast, protocol A incurs no further privacy costs after synthetic data release.

When the privacy level was high ($\epsilon \leq 1$), the quality of generated data decreased, resulting in low accuracy and high variability in both protocol A and B. It should be noted that typically at least $\epsilon = 1$ privacy budget was needed for reliable hyperparameter selection.

Among the synthesizers used in this study, MWEM-PGM

performed the best in terms of downstream classifier accuracy, which is in line with results reported in other studies based on DP-synthetic tabular data generation [13]. It should be noted that MWEM-PGM resulted in overoptimistic accuracy values compared to the accuracy obtained when trained only on the private data (dashed line in Figure 4). It appears that in comparison to MST and Privbayes, MWEM-PGM is able to approximate underlying distribution of the real data whilst consuming less privacy budget.

Finally, the results highlight the need for caution when using only synthetic data for evaluating the accuracy of a trained classifier. When the ϵ values used for data generation is small, specially for the datasets with fewer samples, the data generator tends to generate DP-synthetic data with unbalanced binary class labels which hinders the accuracy. While synthetic validation data was often useful for selecting classifier hyperparameters, the actual accuracies obtained on synthetic test data were sometimes very high even for classifiers that would achieve a random level of performance on real data. Thus, it appears that even when it is possible to do both classifier training and model selection using only synthetic data, some real test data may still be required if one needs an unbiased and reliable estimate about the prediction accuracy of the classifier.

Our study has several limitations to be addressed in future work. One limitation was the assumption of discrete input data. How well the continuous data is transformed into discrete values has a significant impact on the quality of the generated DP-synthetic data [6], [8]. Secondly, the impact that the number of sampled synthetic data has on the results should be further investigated. Finally, a broader selection of classifiers should be tested and evaluated with both proposed protocols.

VI. Acknowledgements

This work has been conducted as part of the PRIVASA project funded by Business Finland (grant number 37428/31/2020).

VII. References

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," *Update of Selected Articles (May 4, 2021)*, 2021.

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

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

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

[5] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography conference*, pp. 265–284, Springer, 2006.

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

[7] R. McKenna, D. Sheldon, and G. Miklau, "Graphical-model based estimation and inference for differential privacy," in *International Conference on Machine Learning*, pp. 4435–4444, PMLR, 2019.

[8] 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, oct 2017.

[9] A.-S. Charest, "How can we analyze differentially-private synthetic datasets?," *Journal of Privacy and Confidentiality*, vol. 2, no. 2, 2011.

[10] J. Snoke and A. Slavković, "pMSE mechanism: differentially private synthetic data with maximal distributional similarity," in *International conference on privacy in statistical databases*, pp. 138–159, Springer, 2018.

[11] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou, "Differentially private generative adversarial network," *arXiv preprint arXiv:1802.06739*, 2018.

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

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

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

[15] M. Boedihardjo, T. Strohmer, and R. Vershynin, "Covariance's loss is privacy's gain: Computationally efficient, private and accurate synthetic data," *Foundations of Computational Mathematics*, pp. 1–48, 2022.

[16] C. Dwork, A. Roth, *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[17] C. Arnold and M. Neunhoeffer, "Really useful synthetic data–A framework to evaluate the quality of differentially private synthetic data," *arXiv preprint arXiv:2004.07740*, 2020.

[18] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

[19] K. Cai, X. Lei, J. Wei, and X. Xiao, "Data synthesis via differentially private markov random fields," *Proceedings of the VLDB Endowment*, vol. 14, no. 11, pp. 2190–2202, 2021.

[20] R. McKenna, G. Miklau, and D. Sheldon, "Winning the NIST Contest: A scalable and general approach to differentially private synthetic data," *Journal of Privacy and Confidentiality*, vol. 11, dec 2021.

[21] M. Hardt, K. Ligett, and F. McSherry, "A simple and practical algorithm for differentially private data release," *Advances in neural information processing systems*, vol. 25, 2012.

[22] R. McKenna, G. Miklau, and D. Sheldon, "Private-PGM," 2021.

[23] N. Holohan, S. Braghin, P. Mac Aonghusa, and K. Levacher, "Diffprivlib: the IBM differential privacy library," *arXiv preprint arXiv:1907.02444*, 2019.

[24] I. Jambor, J. Verho, O. Ettala, J. Knaapila, P. Taimen, K. T. Syvänen, A. Kiviniemi, E. Kähkönen, I. M. Perez, M. Seppänen, *et al.*, "Validation of IMPROD biparametric MRI in men with clinically suspected prostate cancer: a prospective multi-institutional trial," *PLoS medicine*, vol. 16, no. 6, p. e1002813, 2019.

[25] 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 *Proceedings of the annual symposium on computer application in medical care*, p. 261, American Medical Informatics Association, 1988.

[26] "Kaggle cardiovascular disease dataset." Available online at https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset, Accessed on 23.02.2023.

[27] D. Zhang, R. McKenna, I. Kotsogiannis, M. Hay, A. Machanavajjhala, and G. Miklau, "Ektelo: A framework for defining differentially-private computations," in *Proceedings of the 2018 International Conference on Management of Data*, pp. 115–130, 2018.

[28] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Advances in computational mathematics*, vol. 13, pp. 1–50, 2000.

[29] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," 2003.

[30] T. Pahikkala and A. Airola, "RLScore: regularized least-squares learners," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 7803–7807, 2016.