EUROPEAN MASTER IN OFFICIAL STATISTICS
EMOS

Abstract
EMOS Master thesis competition 2025

'Privacy-Enhancing Technologies for Synthetic Data Creation with Deep Generative Models'

Author: **Alessio Crisafulli Carpani, University of Bologna**

Keywords: Privacy-Enhancing Technologies (PETs), Differential Privacy, Generative AI

1. INTRODUCTION

In light of the recent technological advancements, our society has evolved into a prolific source of data, which is then gathered, processed, and subjected to analysis, effectively converting our society, economy, and physical environment into expansive reservoirs of data, resembling what could be termed as "data fountains" (Ricciato 2019).

The utilization of data, particularly datasets containing micro-level, individual-specific information, have drawn significant attention in the realm of data mining research. In today's world, numerous real-world systems heavily depend on machine learning (ML) models to carry out a diverse range of tasks, including uncovering novel data patterns and facilitating recommendation systems. However, a significant challenge arises, as many of these ML algorithms have an insatiable demand for data, often necessitating the inclusion of personal sensitive information, in spite of the fact that these systems are vulnerable to privacy breaches (Shokri 2017). Thus, the organizations responsible for these technologies must strike a delicate balance between complying with GDPR and EUDPR and minimizing the risks associated with data loss, theft, or misuse, and cater to the needs of the "modeler", whose aim is to optimize such systems.

On the other side, National Statistical Institutes' (NSIs), alongside other relevant institutions, have the critical responsibility of providing reliable, pertinent, timely, and high-quality data to support evidence-based decision-making. Nevertheless, to respond effectively to emerging issues, NSIs often require supplementary data from secondary sources, including administrative or private sector data. This scenario calls for a coordinated international response, necessitating timely access to new data sources and potentially sensitive data shared among multiple partners, some of whom may be in different countries. However, due to legitimate privacy concerns, unrestricted access to all data cannot be granted to these partners.

Entities that lack access to extensive data-gathering resources, including researchers, small businesses, and ordinary individuals, face difficulties in accumulating sufficient data for training specific types of models. In such cases, generating synthetic data offers a more accessible alternative to acquiring original data, combining two aspects: usefulness for the statistical analysis (data

augmentation) and the preservation of confidentiality. There are numerous scenarios September 2024 in which companies employ synthetic data to make information available for processing, especially in a post-GDPR world when regulations or privacy concerns impose restrictions on accessing the original data.

Within this context, generative models have emerged to create synthetic samples across various domains. Ideally, these models should prevent the exposure of individual-specific information from the training data. Unfortunately, recent literature has shown that this assumption is not consistently met, particularly with Generative Adversarial Networks (GANs), which lacks robust privacy guarantees. Nevertheless, there is a critical need to strike a balance between our responsibilities as data stewards and the importance to advance data mining research. In this regard, Privacy-Enhancing Technologies (PETs) can help mitigate these challenges by imposing privacy constraints on models or more generally in algorithms, enabling their use and sharing without compromising the confidentiality of the training data.

2. OBJECTIVE

One promising approach in this domain involves modelling the data-generating distribution by training a generative model on the sensitive data, introducing the mechanism of Differential Privacy, a mathematical foundation for quantifying and achieving privacy in data analysis. This privacy-preserving model is then shared along with its private parameters, allowing anyone to generate a synthetic dataset that closely mirrors the original training data without compromising the robust protection of privacy.

This research is dedicated to exploring the latest techniques in the field of Privacy-Enhancing Technologies, in particular of Differential Privacy, by injecting these constraints into generative neural networks to create differentially private synthetic datasets and investigating the trade-off between data utility and privacy preservation through state-of-the-art programming libraries.

Beyond this, applications of the differential privacy mechanism were also studied with other supervised learning algorithms, to demonstrate that the contribution of the individuals' data is also masked out within these models, thus preventing data-leakage, and leading to comparable performance with the non-privatised models.

3. METHODS

In this work, the two primary approaches focus on incorporating differential privacy directly into the training process of Generative Adversarial Networks, which offer a distinct advantage by introducing noise within the latent space, rather than directly altering the data as with other output privacy techniques. These approaches allow us to ensure privacy while minimizing the overall loss of information.

At the core of the GAN framework lies the concept of adversarial training, wherein the generator aims to produce samples that are indistinguishable from real data and the discriminator strives to differentiate between genuine and generated samples, leading to data samples that accurately capture the characteristics of a desired target distribution. Differential privacy can be seamlessly integrated into the discriminator, introducing gaussian noise into the stochastic gradient descent (SGD) algorithm (Martin Abadi 2016). Similarly, this can be applied to Conditional Tabular GANs

(CTGANs), a more robust version of the neural network that models the conditional probability distribution among the rows of tabular data, hence more able to mitigate the effect of heterogeneity, imbalance, or more generally highly sparse vectors. Another method to guarantee the privacy of the training data is to transfer the knowledge from an ensemble of "teacher" models to a "student" model during the learning process. This is achieved through the Private Aggregation of Teacher Ensembles (PATE) mechanism, which replaces the standard architecture of the GANs' discriminator (Jordon 2022).

Therefore, various generative neural networks, including DP-GANs, DP-CTGANs, PATE-GANs, and PATE-CTGANs, were implemented with different privacy budget parameters to study the trade-off between data privacy and utility. The experiments were conducted on two open-source datasets, reflecting this real-world scenario in which this sensitive type of data is made publicly available and could also be used as input for different privacy attacks. The quality of the datasets generated by these differentially private generative models is assessed by training a set of binary classification models, including Logistic Regression, Gaussian Naive Bayes, Random Forest, AdaBoost, Bagging of Decision Trees and Gradient Boosting, which have been evaluated using the accuracy, the area under the receiver operating characteristics curve (AUC), the recall and the F1 score. The same evaluation was also considered using the repeated k-fold cross validation, on three different training testing settings:

1. Setting TRTR: the models are trained on the real training set and assessed on the real testing set, the standard setting for benchmarking purposes.

2. Setting TSTR: the models are trained on the synthetic training set and assessed on the real testing set, to determine how well the synthetic data is able to capture the relationship between the variables.

3.Setting TSTS: the models are trained on the synthetic training set and assessed on the synthetic testing set, to evaluate the consistency of relative performance with TRTR.

To evaluate the utility of the synthetic data, also the synthetic ranking agreement (SRA) and the propensity score mean-squared error (pMSE) were computed. Whereas for the privacy risk assessment, multiple membership inference attacks (MIA) were conducted at each privacy budget level.

Moreover, to demonstrate the efficacy of differential privacy (DP) and to evaluate the deviation in performance in supervised learning algorithms, when privacy constraints are applied, two models - Logistic Regression and Gaussian Naive Bayes - were trained, both involving the inclusion and exclusion of differential privacy

4. RESULTS

From the experimental results, it is possible to notice that the classification scores for synthetically generated datasets (TSTS) are close to those of non-private datasets (TRTR). In particular, the DP-CTGANs or PATEGANs were found to be more robust to overfitting and therefore more able to capture the whole variability of source data, compared to the DP-GANs which often led to a generalization of the training data.

		PATEGAN							
		Baseline Scores				Cross Validation Scores			
Setting	ε	Accuracy	AUC	F1	Recall	Accuracy	AUC	F1	Recall
TRTR		82.18%	79.76%	71.44%	75.28%	77.79%	88.50%	61.48%	64.68%
TSTR		83.15%	65.90%	34.36%	32.56%	66.23%	50.23%	10.20%	18.91%
TSTS	0.1	84.52%	71.83%	53.48%	49.73%	76.91%	70.19%	27.29%	22.05%
	10	85.62%	75.06%	91.13%	95.44%	76.99%	72.72%	85.70%	91.15%
	100	77.53%	73.56%	61.97%	65.96%	74.70%	76.39%	43.25%	39.73%

Table 1. Evaluation Results for PATE-GAN

When we consider the privacy budget of a model, we can observe that greater values of this parameter (ε) imply fewer privacy constraints being imposed, which in turn, suggests a more transparent generation of data. Consequently, it was observed that increasing this parameter, led to an improvement in the model's performance on synthetic data utility metrics.

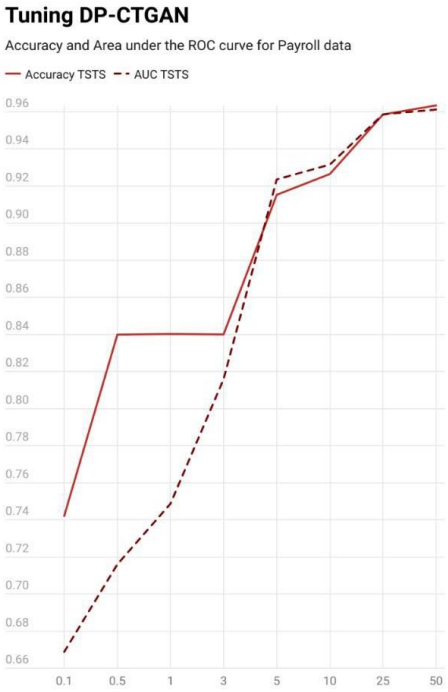

Figure 1. INCREASING THE PRIVACY BUDGET, THE MODEL PERFORMANCE INCREASES

However, from a privacy risk perspective, this implicates a higher vulnerability to privacy attacks. This was evidenced by the results of membership inference attacks performed on the synthetic datasets, which reported significantly lower success rates on those generated with a low privacy budget (Figure 2).

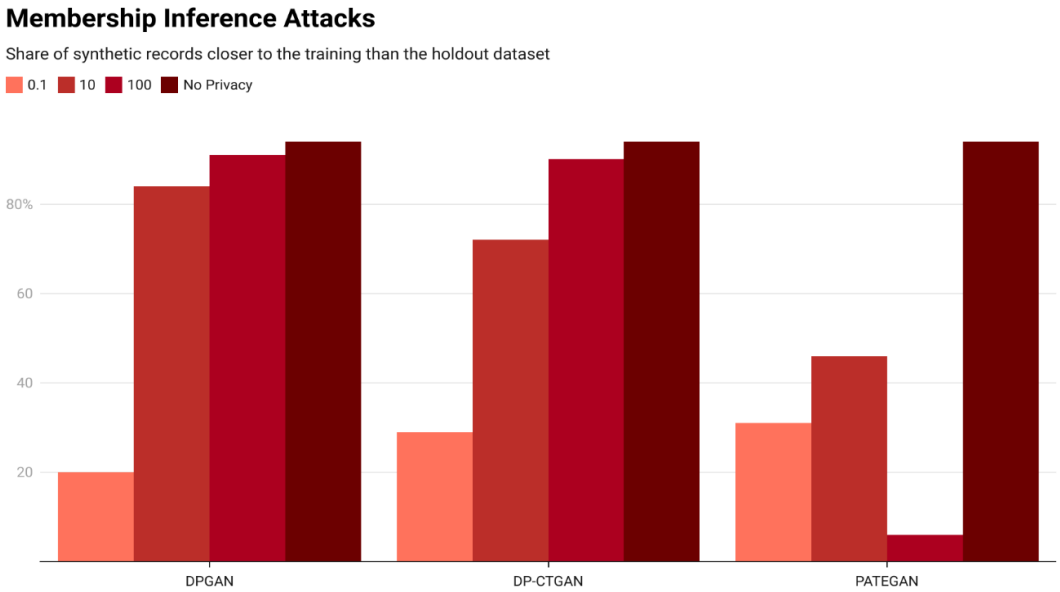

Figure 2. Privacy Attacks Results (MIA)

Furthermore, the experiments conducted with differentially private classification models reported good performances. As expected, the DP scores were found slightly lower but still very competitive with the non-private counterparts, as their averages ranged between 60%-70%. Particularly, the score of the DP-version of the logistic regression decreased in accuracy and AUC of just 6 and 4 basis points respectively.

However, training generative adversarial networks, especially when integrating the PATE framework, comes with several challenges. First, this includes the necessity of working with very powerful systems to meet the intense computational costs required to correctly train these networks. Second, GANs are unstable during training and the large bias produced by the critic in the gradient of the generator, when mixed with the imposed gradient noise by the differential privacy, can increase training instabilities. Future work could focus on implementing DP within other generative models and compare the results, by trying over different datasets with higher complexity and assessing them through more types of privacy attacks.

5. CONTRIBUTION

This thesis, which is the outcome of a research project in the field of Privacy- Enhancing Technologies for Official Statistics (PET4OS) conducted with the Italian National Institute of Statistics (ISTAT), revealed the efficiency and feasibility of applying differential privacy for multiple purposes. This study has the potential to:

- Enhance the understanding of the latest techniques for generating synthetic data while respecting the principles of differential privacy and furnish guidance to researchers, organisations, and policymakers on the practical application of differential privacy, also in supervised learning tasks.

- Provide insights about the trade-off between data utility and privacy preservation, specifically in the context of generative models and how to investigate it.

• Contribute to the development of best practices for leveraging synthetic data in data-driven tasks while adhering to stringent privacy regulations.

The synthesizer approach's main advantage is that the resulting dataset can be shared and used for analytical purposes any number of times without increasing the risks associated with privacy loss. Another advantage is that the synthesizer allows producing any arbitrary amount of data derived from the original dataset's distribution, a promising approach for data augmentation to improve a model performance. The python modules developed for this study can be adopted by NSIs to generate new privatised data, investigate the parameters and metrics, share the models and allowing external organisations to make inference on the synthetic dataset. For instance, numerous potential use cases could be identified between the European Statistical System (ESS) and the European System of Central Banks (ESCB). Furthermore, this can serve as starting point for further investigations or can be extended with more generative models or functionalities.

Another example, in a post-GDPR world, the processing of customer data involves stringent compliance and governance requirements for companies. In this scenario, the data curator initially encodes private data into a generative model, subsequently this model is shared with an analyst, who can use it to create data that resembles the original dataset. This provides organisations or companies with greater flexibility and freedom to process data in a secure manner.

6. REFERENCES

[1] Daryna Oliynyk, Rudolf Mayer, and Andreas Rauber. 2023. "I know what you trained last summer: A survey on stealing machine learning models and defences." doi:10.1145/3595292.

[2] European Commission. 2020. "White Paper on Artificial Intelligence: a European approach to excellence and trust." https://commission.europa.eu/publications/white-paper-artificial-intelligence-european- approach-excellence-and-trust_en#related-links.

[3] Jordon, James and Yoon, Jinsung and Schaar, Mihaela van der. 2022. "PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees." Edited by International Conference on Learning Representations. https://openreview.net/forum?id=S1zk9iRqF7.

[4] Joshua Snoke, Gillian M Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. 2018. "General and specific utility measures for synthetic data." Journal of the Royal Statistical Society 181.

[5] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar,. 2016. "Deep

[6] learning with differential privacy." Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (ACM). doi:10.1145/2976749.2978318.

[7] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar. 2018. "Scalable private learning with pate." https://arxiv.org/abs/1802.

[8] OECD. 2023. "Emerging privacy-enhancing technologies." doi:https://doi.org/10.1787/bf121be4-en.

[9] Ricciato, Fabio, et al. 2019. "Trusted smart statistics: Motivations and principles." Statistical Journal of the IAOS

[10] (IOS Press) 35 (4): 589-603. https://cros.ec.europa.eu/system/files/2023-12/sji190584.pdf.

[11] Shokri, Shmatikov, Marco Stronati and Congzheng Song and Vitaly. 2017. “Membership Inference Attacks against Machine Learning Models.” (arXiv). https://arxiv.org/abs/1610.05820.

[12] United Nations Committee of Experts on Big Data and Data Science for Official Statistics. 2023. “United Nations Guide on Privacy-Enhancing Technologies for Official Statistics.” https://unstats.un.org/bigdata/task- teams/privacy/guide/index.cshtml.