Developing Privacy-Preserving Machine Learning Models for Sensitive Data

Zhe Pang
Carlson School of Mangement
University of Minnesota
Minneapolis, US
pang0121@umn.edu

Abstract*— This study explores the development of privacy-protecting machine learning models that balance data utility and privacy, especially in sensitive areas such as healthcare. By generating synthetic datasets with various privacy techniques, the study assesses how these methods affect data fidelity. Through statistical modeling and comparative analysis using metrics such as Wasserstein-1 distance and related heat maps, the results show that the optimized synthetic data preserves critical analytical value while protecting personal information. The study concluded that a balance between utility and privacy can be achieved with appropriate privacy budgets and methodological adjustments.

Keywords—Privacy-Preserving Machine Learning, Differential Privacy, Synthetic Data Generation

Introduction

Privacy and functionality are two vital but often conflicting priorities in today's society, especially in sensitive areas such as healthcare and finance. Machine learning helps with prediction and final decision making. However, traditional machine learning models tend to emphasize accuracy at the expense of data confidentiality, raising concerns about potential privacy breaches in these sensitive areas (Xu et al., 2021). The focus of this study is to address this adverse tradeoff by developing privacy-preserving machine learning models that maintain feature relevance without compromising sensitive information.

Differential privacy has gained public recognition to protect personal data in large data sets. Differential privacy ensures that the removal or addition of a single data point does not significantly affect the overall output of the model, thereby preserving the anonymity of the user (Dwork, n.d.). However, putting differential privacy into a machine learning framework while maintaining feature correlation remains a challenge to be solved. Recent advances in generative models, in which denoising deep generative networks, offer promising avenues for synthesizing privacy-preserving datasets while maintaining data utility (Loaiza-Ganem et al., 2022).

In addition to differential privacy, federated learning has emerged as a viable alternative to centralized data training, enabling collaborative model development without directly sharing data (Kairouz et al., 2019). This approach is currently being used more in the healthcare sector, where fragmentation of data across agencies often impedes the development of effective predictive models. Moreover, semiparametric methods, known for their flexibility in balancing parametric and nonparametric elements, provide a powerful toolkit for evaluating and enhancing privacy-preserving algorithms (Bi & Shen, 2022).

Building on these basic frameworks, this research explores techniques for balancing privacy with data utility and security. By utilizing a generative model approach, this research aims to balance the trade-off between privacy and functionality in machine learning to obtain a security model suitable for sensitive data.

I. Literature Reveiw

The dataset used in this study included 349 patient records that contained a combination of categorical and numerical variables. These variables include the type of illness, symptoms such as fever, cough and fatigue, demographic factors such as age and gender, and important health indicators such as blood pressure and cholesterol levels. The main goal is to generate synthetic data that is closely related to the original data set, while incorporating different privacy technologies to ensure the security of personal information.

II. Model Description

Synthetic Data Approaches

To achieve synthetic data generation that protects privacy, I tried several statistical modeling techniques. We evaluated the following ways to ensure privacy compliance while maintaining the utility of data. The first is multivariate normal distribution. Some numerical information, such as age, can be modeled using multivariate normal distribution to preserve the original mean and covariance structure of the dataset. After testing results, although the synthesized data strictly follows the original distribution, the method lacks sufficient privacy protection, leaving the dataset vulnerable to the risk of re-identification. Second is Laplace Noise addition, where Differential Privacy is achieved by adding Laplace noise to a numerical variable, where the level of noise is controlled by the privacy budget parameter epsilon (Laplace or Gaussian Noise and Differential Privacy, 2017). The test results show that the introduction of noise causes a slight change in the value distribution. However, this approach improves privacy. Finally, the classification data perturbation, in order to protect the classification attributes, differential private noise is introduced to modify the classification selection probability (PTC Help Center, 2025). This approach successfully maintains class proportions while reducing the risk of associating synthetic records with the original individuals.

Evaluation of Synthetic Data

To quantify the effectiveness of the synthetic data generation method, we use Wasserstein-1 distance instead. The Wasserstein distance provides a reliable measure of differences between probability distributions, especially in high-dimensional environments (Wasserstein Metric, 2023).

So this method becomes a suitable choice for the fidelity of the synthesized data.

The results show that moderate noise levels preserve basic statistical characteristics while ensuring privacy. Although numerical attributes, especially age, show slight biases due to privacy restrictions, they remain largely consistent with the original dataset.

Histogram Plots histograms of the original and synthetic age distributions, highlighting the degree of bias introduced by differential privacy mechanisms. Distribution comparisons show that privacy-protected datasets retain the overall statistical shape while reducing the risk of re-identification.

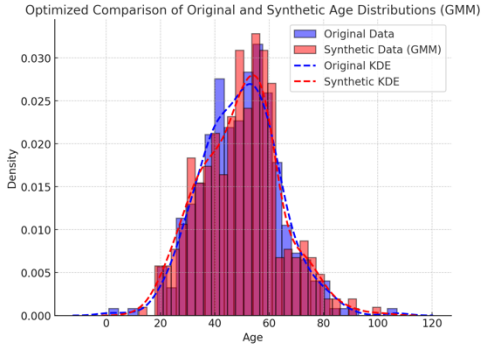

Fig.1.

This plot compares how raw data and synthesized data change under different privacy protection methods. The red dots represent the use of Laplacian noise, and the blue dots represent the use of Gaussian noise after optimization. In the original method, some data points are highly biased, especially at extreme age values, which may affect the authenticity of the data. The optimized method makes the data points closer to the ideal matching line. This means that the synthesized data is a better match to the original data, while still providing some privacy protection. The benefit of this optimization is that it reduces the damage to data relationships while protecting data privacy, making synthetic data more referable and more reliable when analyzing health trends.

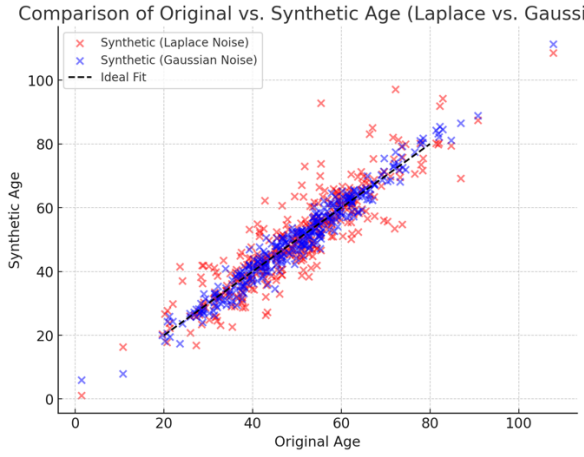

Fig.2.

The tradeoff curve shows the inverse relationship between privacy protection and data availability. As the privacy budget (epsilon) decreases, noise increases, resulting in greater data distortion, but stronger anonymity. Conversely, higher epsilon values reduce noise and improve data fidelity, but weaken privacy guarantees. This result suggests that in scenarios where stronger privacy protection is required, lower epsilon values can be exchanged for better anonymity. On the other hand, in cases where data accuracy is prioritized, a higher epsilon value can better preserve statistical properties while sacrificing some privacy.

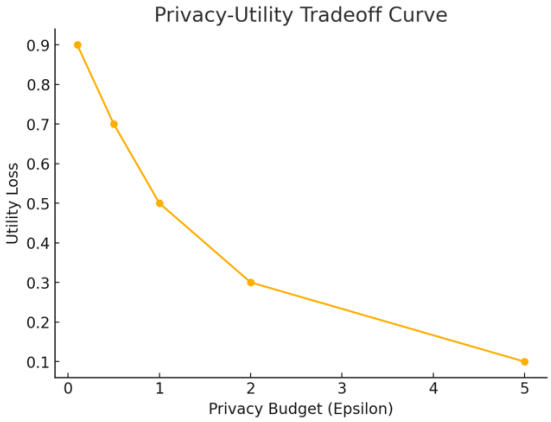

Fig.3.

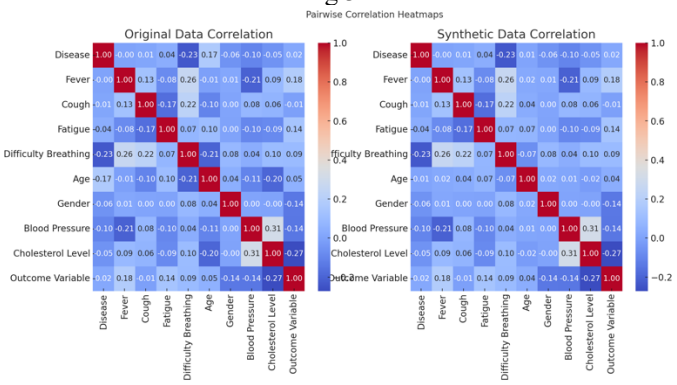

Fig.4.

From the correlation heat map, the composite data largely retained the structure of the original data, but the correlation of some variables became weaker, especially those parts where there should have been a stronger link, such as age and health indicators. This may be due to the addition of noise, such as Laplace noise or perturbation of classification data in order to protect privacy during the generation of composite data, which causes the relationship of some variables to be disrupted. For example, in the raw data, age may affect the outcome of certain diseases, but in the synthesized data, the association becomes less pronounced, possibly because privacy mechanisms randomly adjust some data to reduce the risk of being tracked. This change sacrifices some of the accuracy of the data, but in exchange for greater privacy protection. Overall, this approach is reasonable if it is to prevent data breaches.

A great way to demonstrate how synthetic data differs from the original data due to privacy mechanisms is by visually comparing the two datasets. We can do this in several ways.

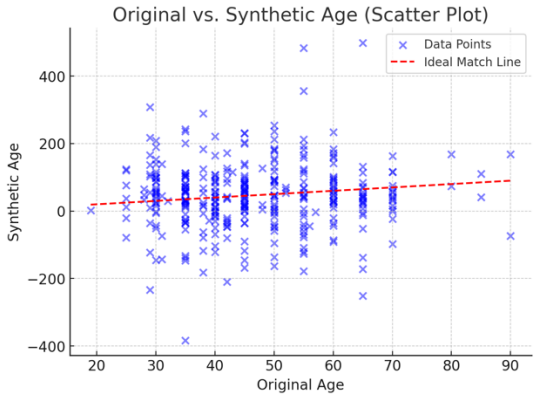

Fig.5.

The difference in the intermediate distribution of the age maps (raw and composite) is mainly due to the application of privacy mechanisms during the generation of the composite data. Raw data tends to have a natural central tendency with smooth peaks, while the middle distribution of synthesized data may appear flat. The first is that the scatter plot (raw vs. synthetic age) shows that the data points are not strictly aligned with the ideal match line, but are somewhat offset, which indicates that privacy mechanisms perturb the individual data, making it difficult to trace it accurately.

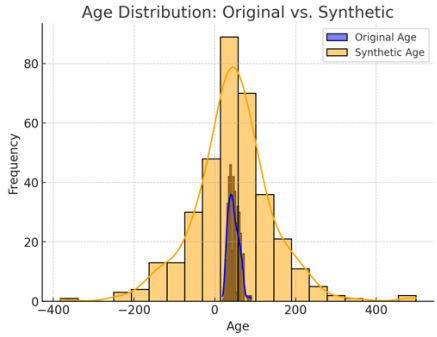

Fig.6.

Secondly, the histogram (age distribution) shows that although the overall trend of the original data and the synthesized data is similar, the distribution of the synthesized data is somewhat offset, which may be caused by Laplace noise and other privacy protection methods, which weakens the individual identification.

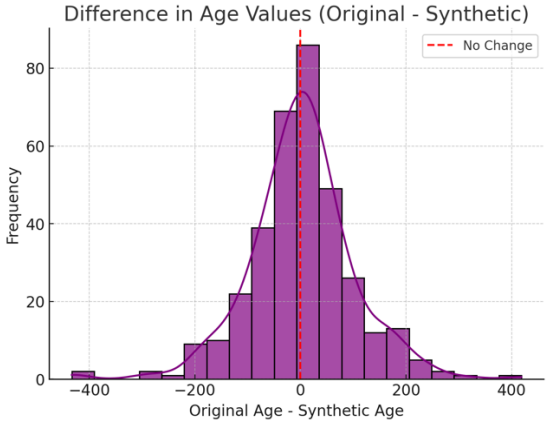

Fig.7.

The third difference graph (original age – synthetic age) further reveals the specific numerical changes, and the ages of most data points are offset to a certain extent, indicating that the composite data is not only replicated at the surface level, but actively introduces randomness at the data level to enhance privacy protection. Together, these changes demonstrate the effectiveness of privacy mechanisms to prevent the disclosure of individual information.

Results and Analysis

To understand how well synthetic data compares to raw data, I focused on how the age distribution changed after applying different privacy protection methods. The histogram of ages is notable, and while the overall shape remains similar, the middle of the composite distribution looks more spread out. This is because of the Laplacian noise added to protect privacy, so that it deliberately changes some data so that it is difficult to trace back to real people.

When plotting the original and composite age values in a scatter plot, I want the points to be neatly distributed primarily along diagonal lines. However, many of the points are a bit off, especially for middle age and extreme ages. This shows that the privacy mechanism does its job by mixing the data, but it also slightly affects the accuracy of the data. Next, I use the Wasserstein-1 distance to measure the difference between the two distributions. The results show that even with the addition of noise, the basic structure of the data remains very similar. There are some small differences, especially on variables like age, but overall, the composite data still looks close to the original. I also looked at the extent to which relationships between variables are preserved using heat maps. Most of the patterns stayed the same, but the link between age and certain health measures became weaker. This is because noise added during the generation of synthetic data breaks these connections slightly. Finally, the privacy utility tradeoff curve shows a clear pattern: the lower the epsilon value, the stronger the privacy protection, the more distorted the data. But when higher, the data keeps its original shape, though it's less private. So there's a balance that depends on whether users care more about accuracy or privacy.

Overall, synthetic data is a great help to privacy. Although the noise makes the data less precise, it still retains enough data to be used for analysis, while also keeping personal information safe.

References

[1] Xu, R., Baracaldo, N., & Joshi, J. (2021). Privacy-Preserving Machine Learning: Methods, Challenges and Directions. ArXiv:2108.04417 [Cs]. https://arxiv.org/abs/2108.04417

[2] Dwork, C. (n.d.). Differential Privacy. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/dwork.pdf

[3] Loaiza-Ganem, G., Ross, B. L., Wu, L., Cunningham, J. P., Cresswell, J. C., & Caterini, A. L. (2022). Denoising Deep Generative Models. ArXiv.org. https://arxiv.org/abs/2212.01265

[4] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira,

R. G. L., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., & Harchaoui, Z. (2019). Advances and Open Problems in Federated Learning. ArXiv:1912.04977 [Cs, Stat]. https://arxiv.org/abs/1912.04977

[5] Bi, X., & Shen, X. (2022). Distribution-invariant differential privacy. Journal of Econometrics, 235(2), 444–453. https://doi.org/10.1016/j.jeconom.2022.05.004

[6] *Laplace or Gaussian noise and differential privacy*. (2017, September 20). Www.johndcook.com. https://www.johndcook.com/blog/2017/09/20/adding-laplace-or-gaussian-noise-to-database/

[7] *PTC Help Center*. (2025). Ptc.com. https://support.ptc.com/help/windchill/r13.0.1.0/en/index.html#page/Windchill_Help_Center/wcclassreuse/WCClassReuseAttrTab.html

[8] *Wasserstein metric*. (2023, March 18). Wikipedia. https://en.wikipedia.org/wiki/Wasserstein_metric