Impact of using synthetic dataset for model training on users' privacy.

***Abstract*—The abstract**

I. Introduction

A. Research questions

What is the impact of using synthetic data instead of real data on users' privacy ?

II. Background

A. Classification task

B. Machine learning

In classification tasks, a machine learning model is a function that maps features of a data record to its label. Is function has an architecture which describes the strucure of the internal computing and parameters. For instance with mono dimensialon data, the linear model is $f(x) = ax+b$ where x is the feature and a and b are the parameters. Training a machine learning model means using an optimization algorithm that will find optimal parameters to best achieve the calssification.

C. Synthetic datas

A generator is a function that takes as input a real dataset and outputs a synthetic dataset. This definition is general enough so that the identity function is a generator. Even though synthetic datasets are supposedely different than real world datasets.

D. Membership inference attack

This attack infer if a data record has been used in the training of a machin learning model. This attack is effectivaly made by leveragin shadow models: models that imitates the behaviour of the target [7].

Differential privacy is a probabilistic definition that bound mambership inference attack's success. In practice, thoes guaranties are achieved thrgout gradient clipping and additive noise in the training algorithme [1].

E. Attribut inference attack

This attack infer sensitive attributes of a data record.

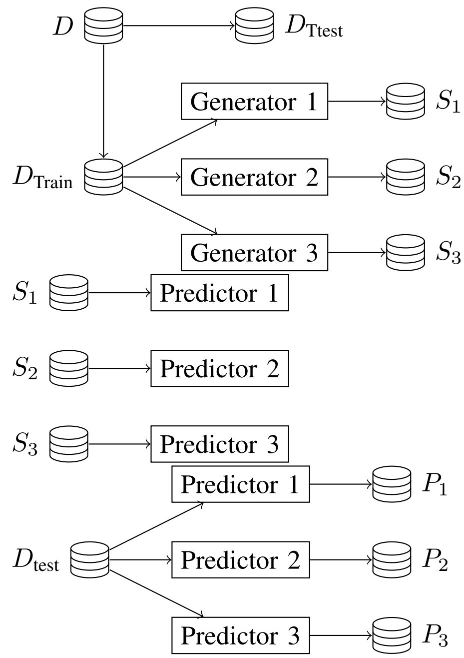
III. Methodology

A. Datasets

1) US census: Retiring adult allow to control which states of the US are used [3].

2) UTKFace: This images dataset is composed of 20,000 pictures of faces [10].

B. Data pipeline



C. Generator training

1) Auto encoder:

2) Variational auto encoder:

D. Predictor training

1) Fully connected neural network:

2) Convolutional neural network:

E. Attack training

1) Membership inference attack:

2) Attribute inference attack:

IV. Results

A. Utility

Using synthetic datatset degrades the utility of the predictor.

B. Membership inference attack

Using synthetic dataset slightly degrades the success of membership inference attack.

C. Attribute inference attack

Using synthetic dataset does not have an impact on the success of attrivute inference atttack.

V. Related work

Privacy and synthetic datasets [2].

Datasynthesizer: privacy preserving synthetic datasets [6].

Towards improving privacy of synthetic datasets [5].

User-Driven Synthetic Dataset Generation with Quantifiable Differential Privacy [9].

Synthetic data-A privacy mirage [8].

Hide-and-seek privacy challenge: Synthetic data generation vs. patient re-identification [4].

VI. Conclusion

Even though synthetic dataset are promising regarding users' data protection, in itself it does not bring guaranties that membership status nor sensitive attributes are protected.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] Steven M Bellovin, Preetam K Dutta, and Nathan Reitinger. Privacy and synthetic datasets. *Stan. Tech. L. Rev.*, 22:1, 2019.

[3] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[4] James Jordon, Daniel Jarrett, Evgeny Saveliev, Jinsung Yoon, Paul Elbers, Patrick Thoral, Ari Ercole, Cheng Zhang, Danielle Belgrave, and Mihaela van der Schaar. Hide-and-seek privacy challenge: Synthetic data generation vs. patient re-identification. In *NeurIPS 2020 Competition and Demonstration Track*, pages 206–215. PMLR, 2021.

[5] Aditya Kuppa, Lamine Aouad, and Nhien-An Le-Khac. Towards improving privacy of synthetic datasets. In *Annual Privacy Forum*, pages 106–119. Springer, 2021.

[6] Haoyue Ping, Julia Stoyanovich, and Bill Howe. Datasynthesizer: Privacy-preserving synthetic datasets. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, pages 1–5, 2017.

[7] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[8] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data-a privacy mirage. *arXiv preprint arXiv:2011.07018*, 2020.

[9] Bo-Chen Tai, Yao-Tung Tsou, Szu-Chuang Li, Yennun Huang, Pei-Yuan Tsai, and Yu-Cheng Tsai. User-driven synthetic dataset generation with quantifiable differential privacy. *IEEE Transactions on Services Computing*, 2023.

[10] Song-Yang Zhang, Zhifei and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.