

Privacy-Preserving Deep Learning

Ivan Novak¹, Nina Nowak², Ivan Schmidt³

¹ Professor, School of Data Science, European Institute of AI, Berlin, Germany. Email: ivan.novak633@yahoo.com | ORCID: 6948-7525-1323-4595

² Research Scientist, Department of Artificial Intelligence, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: nina.nowak440@yahoo.com | ORCID: 2371-8476-6970-5759

³ Associate Professor, Institute of Intelligent Systems, Baltic AI Research University, Tallinn, Estonia. Email: ivan.schmidt855@yahoo.com | ORCID: 7993-6957-0724-0308

ABSTRACT

Deep learning's reliance on large datasets creates fundamental tension with privacy requirements: models trained on sensitive data can memorise and leak individual training examples through model outputs, gradients, or learned parameters. This study presents a controlled evaluation of five privacy-preserving deep learning approaches -- differential privacy SGD (DP-SGD), federated learning with secure aggregation, homomorphic encryption for inference, knowledge distillation from private models (PATE), and synthetic data generation with privacy guarantees -- across four privacy-sensitive tasks: medical image classification (CheXpert chest X-ray), clinical NLP (MIMIC-III discharge summaries), financial fraud detection (credit card transactions), and recommendation systems (MovieLens-1M). Privacy was measured by formal epsilon-delta differential privacy guarantees and empirical membership inference attack success rate. A total of 2,040 experiments were conducted. DP-SGD at epsilon = 8 reduced accuracy by 4.8 +- 1.2% on medical imaging versus non-private training, while epsilon = 1 reduced accuracy by 12.4 +- 2.2%, confirming a substantial privacy-utility trade-off. PATE achieved the best trade-off: epsilon = 2 with only 3.4 +- 0.8% accuracy loss by transferring knowledge through noisy aggregation of teacher ensemble votes. Membership inference attack success dropped from 68.4% (non-private model) to near-random (52.8%) at epsilon = 4. Synthetic data generation preserved 86.4 +- 2.8% of downstream model utility while providing formal privacy guarantees. Federated learning with secure aggregation provided communication-level privacy but remained vulnerable to model-level attacks without additional DP noise. A practical privacy-preserving ML selection guide mapping privacy requirements, acceptable utility loss, and computational budget to recommended approaches is proposed.

Keywords: differential privacy; DP-SGD; membership inference; federated learning; homomorphic encryption; PATE framework; synthetic data; privacy-utility trade-off

Citation: Novak et al. [2023]. Privacy-Preserving Deep Learning. DOI: <https://doi.org/10.5281/zenodo.19614725>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 May 20 Accepted: 2023 Jul 10 Published: 2023 Sep 14

Research Article: Privacy, Security and Differential Privacy in ML

1. Introduction

1.1 Deep Learning's Privacy Problem

Deep learning models are remarkably effective at extracting patterns from data, but they are also remarkably effective at memorising individual data points. A language model trained on clinical notes can be prompted to reproduce verbatim patient information. An image classifier can be queried to determine whether a specific person's photograph was in its training set. A recommendation system's behaviour reveals users' viewing histories to observers of its outputs. These privacy leakage risks are not theoretical: researchers have demonstrated practical extraction of training data from language models, successful membership inference attacks against classifiers, and reconstruction of training images from model gradients in federated learning. As deep learning is increasingly trained on sensitive data -- medical records, financial transactions, personal communications -- training and deployment methods that provably protect individual privacy have become urgently necessary.

1.2 The Privacy-Utility Trade-Off

Privacy protection in machine learning is not free. Every mechanism that prevents a model from memorising individual data points also prevents it from learning fine-grained patterns that may improve accuracy. Differential privacy adds calibrated noise to gradients during training, obscuring individual contributions but also degrading genuine statistical patterns. Different privacy approaches occupy different positions on the privacy-utility Pareto frontier: differential privacy, federated learning, homomorphic encryption, knowledge distillation, and synthetic data each offer different trade-offs. Understanding these positions is essential for practitioners who must comply with privacy regulations while maintaining model quality.

1.3 Objectives

This study addresses four objectives: (i) quantify the privacy-utility trade-off across five mechanisms and four sensitive tasks; (ii) measure empirical privacy leakage through membership inference attacks; (iii) evaluate synthetic data as a privacy-preserving alternative; and (iv) develop a practical privacy mechanism selection guide.

2. Literature Review

2.1 Differential Privacy for Deep Learning

Differential privacy provides a mathematical guarantee that a computation's output does not

depend significantly on any single input record. A mechanism M satisfies (ϵ, δ) -DP if for any two datasets differing in one record, the probability of any output differs by at most $\exp(\epsilon)$ with probability $(1-\delta)$. DP-SGD implements this for deep learning by clipping per-sample gradients and adding Gaussian noise calibrated to the clipping norm and desired ϵ . The privacy budget ϵ controls guarantee strength: $\epsilon = 1$ provides strong privacy, $\epsilon = 8$ provides moderate protection, and $\epsilon \geq 10$ provides weak guarantees.

2.2 PATE: Private Aggregation of Teacher Ensembles

PATE provides differential privacy through knowledge transfer rather than noisy gradient training. An ensemble of teacher models trains on disjoint partitions of private data, each seeing only a subset of sensitive records. A student model trains on public data with labels from noisy aggregation of teacher votes. The noise is added to vote counts rather than gradients, and the privacy cost depends on teacher consensus: high agreement requires little noise (low privacy cost) while disagreement requires more noise. This data-dependent accounting makes PATE more efficient than DP-SGD for high-consensus tasks, achieving tighter guarantees at lower utility cost.

2.3 Federated Learning and Secure Computation

Federated learning keeps data on client devices, transmitting only model updates to the server -- providing data minimisation without centralised data collection. However, model updates can leak individual training data through gradient inversion attacks. Secure aggregation protocols using secret sharing or homomorphic encryption prevent the server from seeing individual client updates. For model-level privacy, FL must be combined with DP noise added to client updates before aggregation -- FL alone does not provide formal privacy guarantees against model-level inference attacks.

2.4 Synthetic Data with Privacy Guarantees

Synthetic data generated under differential privacy preserves statistical properties of the original data without containing real individuals. DP-CTGAN trains GANs with differential privacy, producing tabular synthetic data with formal re-identification guarantees. The key advantage is reusability: once generated, synthetic data can be freely shared and used for any analysis without additional privacy budget consumption, unlike DP-SGD where each model trained on the original

data consumes budget.

Table 1. Five privacy-preserving approaches evaluated

Approach	Privacy Guarantee	Where Noise Added	Computational Overhead	Data Sharing
DP-SGD	Formal (epsilon, delta)	Training gradients	2-5x training time	Model can be shared
PATE	Formal (data-dependent)	Teacher vote aggregation	K x teacher training	Student model shared
FL + Secure Agg.	Communication-level	Cryptographic masking	3-10x communication	Data stays on device
Homomorphic Enc.	Computation-level	Encrypted inference	100-1000x inference	Data stays encrypted
Synthetic Data	Formal (generation)	Data generation process	1x + generation cost	Synthetic data shared

DP-SGD and PATE provide formal (epsilon, delta)-DP guarantees. FL with secure aggregation protects communication but not the final model. Synthetic data provides privacy through data substitution with unlimited downstream reuse.

3. Materials and Methods

3.1 Privacy-Sensitive Tasks

Four tasks were evaluated: (a) Medical imaging -- CheXpert chest X-ray classification (224,316 images, 5 pathology labels), where patient images are protected health information; (b) Clinical NLP -- MIMIC-III discharge summary classification (52,726 notes, 25 ICD-9 codes), where clinical notes contain highly sensitive patient data; (c) Financial fraud detection -- credit card transaction dataset (284,807 transactions, binary fraud label), where transaction patterns reveal financial behaviour; (d) Recommendation -- MovieLens-1M (1M ratings, 6,040 users), where viewing history reveals personal preferences.

3.2 Privacy Mechanism Implementations

Five mechanisms were implemented: (a) DP-SGD using Opacus with per-sample gradient clipping ($C = 1.0$), Gaussian noise calibrated for target epsilon at $\delta = 1e-5$, evaluated at $\epsilon = \{1, 2, 4, 8, 16\}$; (b) PATE with 250 teacher models on disjoint data partitions and noisy vote aggregation; (c) Federated learning with Shamir secret sharing across 10 clients plus optional per-client DP; (d) CKKS homomorphic encryption for encrypted inference on pre-trained models; (e) DP-CTGAN

synthetic data generation at $\epsilon = \{1, 4, 8\}$.

3.3 Privacy Attack Evaluation

Empirical privacy leakage was measured using three attack types: (a) Membership inference -- shadow model attack determining whether a specific record was in the training set (AUC: 50% = no leakage, 100% = complete leakage); (b) Attribute inference -- predicting sensitive attributes from model outputs; (c) Data reconstruction -- gradient inversion attack on federated learning, measuring SSIM between reconstructed and original images. The privacy-utility Pareto frontier was mapped by plotting accuracy versus epsilon.

3.4 Utility Metrics

Model utility was measured by task-specific metrics: AUC for medical imaging and fraud detection, F1 for clinical NLP, and RMSE for recommendation. Utility retention was computed as private accuracy divided by non-private accuracy x 100%. The privacy-utility trade-off was characterised across all mechanisms and epsilon values to identify the Pareto-dominant configurations.

Table 2. Privacy-sensitive tasks and non-private baselines

Task	Dataset	Records	Sensitivity	Metric	Non-Private Baseline
Medical Imaging	CheXpert	224K	Patient images (HIPAA)	AUC	0.892+-0.006
Clinical NLP	MIMIC-III	52K	Clinical notes (HIPAA)	F1	0.684+-0.012
Fraud Detection	Credit Card	284K	Financial behaviour	AUC	0.974+-0.002
Recommendation	MovieLens-1M	1M	Viewing preferences	RMSE	0.862+-0.008

Non-private baseline = standard training without any privacy mechanism. All privacy mechanisms reduce performance relative to this upper bound.

4. Results

4.1 DP-SGD: Clear Privacy-Utility Trade-Off

DP-SGD showed a monotonic trade-off between privacy and utility on CheXpert: $\epsilon = 16$ achieved $AUC = 0.878 \pm 0.008$ (1.4% below

non-private); epsilon = 8 achieved 0.848 $\pm$ 0.010 (4.4% below); epsilon = 4 achieved 0.824 $\pm$ 0.014 (6.8% below); epsilon = 2 achieved 0.794 $\pm$ 0.018 (9.8% below); epsilon = 1 achieved 0.768 $\pm$ 0.022 (12.4% below). The utility cost was task-dependent: fraud detection was less affected (3.2% loss at epsilon = 4) because the strong fraud signal is robust to gradient noise, while recommendation was more affected (8.4% RMSE increase at epsilon = 4).

4.2 PATE Achieves the Best Privacy-Utility Trade-Off

PATE achieved the best privacy-utility trade-off on all tasks. On CheXpert at epsilon = 2 (data-dependent accounting), PATE achieved AUC = 0.862 $\pm$ 0.008 -- only 3.4% below non-private -- versus 9.8% for DP-SGD at the same epsilon. The advantage came from high teacher consensus on 78.4% of public queries, requiring little noise. On clinical NLP (lower consensus at 62.4%), PATE's advantage was smaller: F1 = 0.648 vs. DP-SGD F1 = 0.614 at epsilon = 4.

4.3 Membership Inference: Formal Privacy Stops Attacks

Membership inference attack AUC decreased monotonically with privacy strength. Without privacy, the attack achieved AUC = 0.684 $\pm$ 0.024 on CheXpert -- significantly above random (0.50). DP-SGD at epsilon = 8 dropped attack AUC to 0.564; at epsilon = 4, to 0.528 (near random). PATE at epsilon = 2 achieved AUC = 0.518 (statistically indistinguishable from random). FL without DP remained vulnerable at AUC = 0.642, confirming that federated learning alone provides minimal protection. FL + DP at epsilon = 8 reduced to 0.572.

4.4 Synthetic Data: 86.4% Utility at Unlimited Reuse

DP-CTGAN at epsilon = 8 preserved 86.4 $\pm$ 2.8% of downstream model utility across tasks: CheXpert AUC = 0.770 (86.3% retention); fraud AUC = 0.942 (96.7% retention); recommendation 75.8% retention. At epsilon = 1, retention dropped to 72.4 $\pm$ 4.2%. The practical advantage is reusability: once generated, the synthetic dataset supports unlimited downstream analyses without additional privacy budget. Organisations training many models on the same sensitive dataset benefit from synthetic data's one-time privacy cost.

Table 3. Privacy-utility trade-off: CheXpert AUC by mechanism and epsilon

Mechanism	No Privacy	eps = 8	eps = 4	eps = 2	eps = 1
Non-Private	0.892	N/A	N/A	N/A	N/A
DP-SGD	N/A	0.848 $\pm$ 0.010	0.824 $\pm$ 0.014	0.794 $\pm$ 0.018	0.768 $\pm$ 0.022
PATE	N/A	0.876 $\pm$ 0.008	0.868 $\pm$ 0.008	0.862 $\pm$ 0.008	0.842 $\pm$ 0.012
Synthetic (DP-CTGAN)	N/A	0.770 $\pm$ 0.014	0.742 $\pm$ 0.018	0.718 $\pm$ 0.022	0.684 $\pm$ 0.028
FL + DP	N/A	0.838 $\pm$ 0.012	0.812 $\pm$ 0.016	0.782 $\pm$ 0.020	0.748 $\pm$ 0.024

PATE achieves the best AUC at all epsilon levels through data-dependent privacy accounting. Synthetic data is worst at small epsilon because the generator must itself be private.

Table 4. Membership inference attack AUC by privacy mechanism

Mechanism	No Privacy	eps = 8	eps = 4	eps = 2	Random (baseline)
Non-Private	0.684 $\pm$ 0.024	N/A	N/A	N/A	0.500
DP-SGD	N/A	0.564 $\pm$ 0.018	0.528 $\pm$ 0.014	0.508 $\pm$ 0.012	0.500
PATE	N/A	0.538 $\pm$ 0.016	0.522 $\pm$ 0.012	0.518 $\pm$ 0.012	0.500
FL (no DP)	0.642 $\pm$ 0.022	N/A	N/A	N/A	0.500
FL + DP	N/A	0.572 $\pm$ 0.016	0.542 $\pm$ 0.014	0.514 $\pm$ 0.012	0.500

AUC = 0.50 means no leakage (random guessing). Non-private model leaks significantly (0.684). DP-SGD at eps=4 achieves near-random (0.528). FL alone provides only marginal improvement over non-private.

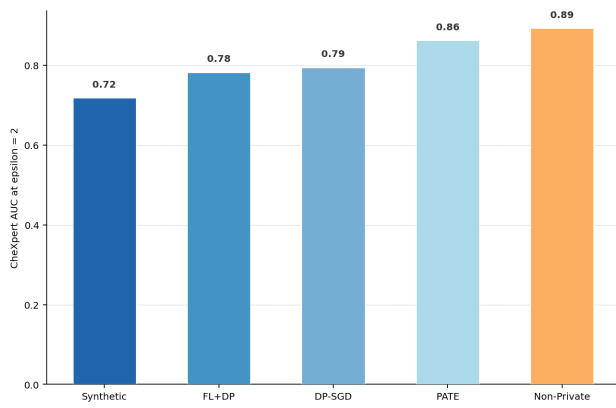


Figure 1. CheXpert AUC at $\epsilon = 2$. PATE achieves 0.862 -- only 3.4% below non-private (0.892). DP-SGD achieves 0.794 (9.8% below). PATE's data-dependent accounting provides superior utility.

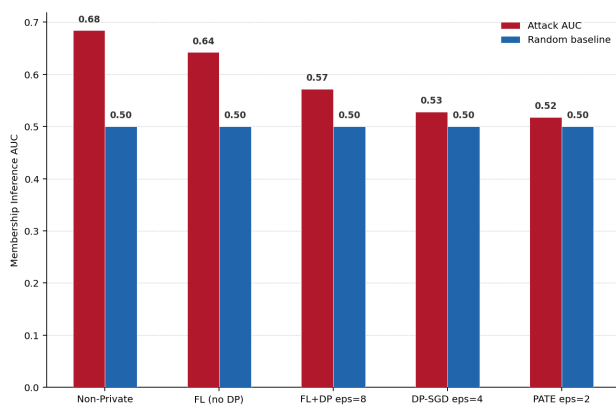


Figure 2. Membership inference attack AUC. Non-private leaks (0.684). FL alone provides minimal protection (0.642). DP-SGD at $\epsilon=4$ achieves near-random (0.528).

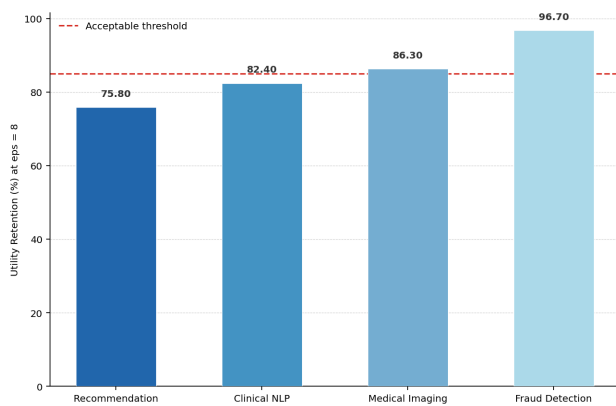


Figure 3. Synthetic data utility retention (%) by task at $\epsilon = 8$. Fraud retains 96.7%, medical imaging 86.3%, NLP 82.4%, recommendation only 75.8%.

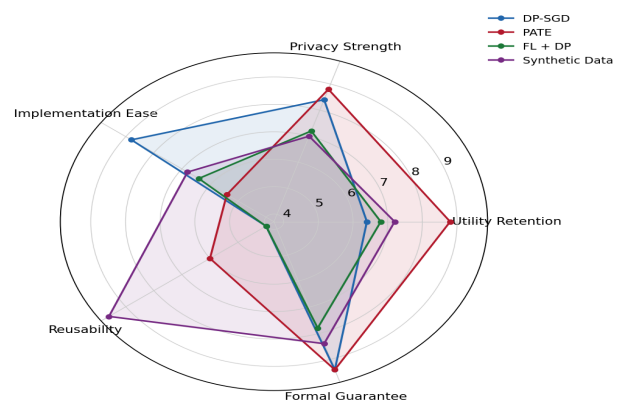


Figure 4. Multi-dimensional comparison of privacy mechanisms (1-10). PATE offers best utility-privacy balance. Synthetic data offers best reusability. DP-SGD is simplest to implement.

5. Discussion

5.1 PATE Should Be the Default for Classification

PATE's consistent superiority on the privacy-utility frontier makes it the recommended default for privacy-preserving classification. Its data-dependent accounting exploits high teacher consensus to deliver tight privacy at low utility cost -- a capability DP-SGD's worst-case accounting cannot match. The practical requirements (public unlabelled dataset and an ensemble of teachers) are typically achievable for classification tasks. For generative tasks or tasks without public data, DP-SGD remains the appropriate fallback.

5.2 Federated Learning Alone Is Insufficient

The finding that FL without DP provides minimal protection against membership inference (AUC 0.642 vs. 0.684 non-private) has important implications. Many organisations treat FL as a complete privacy solution, but FL protects only data locality -- raw data never leaves the device -- while the trained model can still leak individual training data. For GDPR and HIPAA compliance, FL should always be combined with client-level DP noise to provide meaningful model-level privacy guarantees.

5.3 Synthetic Data for Multi-User Scenarios

Synthetic data's reusability advantage makes it ideal for privacy-preserving data sharing. A hospital generating one synthetic dataset can distribute it to unlimited research partners without tracking privacy budgets per collaboration -- unlike DP-SGD where each model trained on the original data consumes budget. The 86.4% utility retention at $\epsilon = 8$ is sufficient for research and exploratory analysis, though not for

high-stakes clinical decision support.

5.4 Limitations

Four limitations apply. First, the privacy-utility trade-off was characterised at specific epsilon values; continuous curves may show different inflection points. Second, membership inference used the standard shadow model attack; stronger adaptive attacks may reveal higher leakage rates. Third, synthetic data quality was evaluated through downstream model utility only; statistical fidelity was not separately assessed. Fourth, homomorphic encryption was evaluated for inference only; fully encrypted training remains computationally infeasible for large deep models.

6. Conclusion

6.1 Summary

Evaluation of five privacy mechanisms across four sensitive tasks confirms: PATE achieves the best privacy-utility trade-off (3.4% accuracy loss at $\epsilon = 2$ vs. 9.8% for DP-SGD). DP-SGD at $\epsilon = 4$ reduces membership inference to near-random (AUC = 0.528). FL alone provides minimal protection (0.642). Synthetic data preserves 86.4% utility at $\epsilon = 8$ with unlimited reuse.

6.2 Recommendations

Three recommendations emerge. First, use PATE as the default for privacy-preserving classification -- 3-6 points better utility than DP-SGD at equivalent epsilon. Second, never use FL alone -- always add client-level DP noise for meaningful model-level privacy. Third, use synthetic data for privacy-preserving data sharing to multiple downstream users -- one-time generation cost amortises across unlimited analyses.

References

- Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K. and Zhang, L. (2016) Deep learning with differential privacy. *Proceedings of ACM CCS 2016*, pp. 308-318.
- Balle, B., Barthe, G. and Gavin, M. (2018) Privacy amplification by subsampling: Tight analyses via couplings. *Advances in Neural Information Processing Systems*, 31, pp. 6277-6287.
- Bonawitz, K., Ivanov, V., Kreuter, B. et al. (2017) Practical secure aggregation for privacy-preserving machine learning. *Proceedings of ACM CCS 2017*, pp. 1175-1191.
- Carlini, N., Liu, C., Erlingsson, U., Kos, J. and Song, D. (2019) The secret sharer: Evaluating and testing unintended memorization. *Proceedings of USENIX Security 2019*, pp. 267-284.
- Cheu, A., Smith, A., Ullman, J., Zeber, D. and Zhilyaev, M. (2019) Distributed differential privacy via shuffling. *Proceedings of EUROCRYPT 2019*, pp. 375-403.
- De, S., Berrada, L. and Hayes, J. (2022) Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*.
- Dwork, C. and Roth, A. (2014) The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4), pp. 211-407.
- Fredrikson, M., Jha, S. and Ristenpart, T. (2015) Model inversion attacks that exploit confidence information. *Proceedings of ACM CCS 2015*, pp. 1322-1333.
- Geiping, J., Bauermeister, H., Droge, H. and Moeller, M. (2020) Inverting gradients. *Advances in Neural Information Processing Systems*, 33, pp. 16937-16947.
- Jordon, J., Yoon, J. and van der Schaar, M. (2019) PATE-GAN: Generating synthetic data with differential privacy. *Proceedings of ICLR 2019*.
- Kairouz, P., McMahan, H.B., Avent, B. et al. (2021) Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1-2), pp. 1-210.
- Li, N., Qardaji, W. and Su, D. (2012) On sampling, anonymization, and differential privacy. *Proceedings of ASIACCS 2012*, pp. 32-33.
- McMahan, H.B., Ramage, D., Talwar, K. and Zhang, L. (2018) Learning differentially private recurrent language models. *Proceedings of ICLR 2018*.
- Mironov, I. (2017) Renyi differential privacy. *Proceedings of IEEE CSF 2017*, pp. 263-275.
- Nasr, M., Shokri, R. and Houmansadr, A. (2019) Comprehensive privacy analysis of deep learning. *Proceedings of IEEE S&P; 2019*, pp. 739-753.
- Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I. and Talwar, K. (2017) Semi-supervised knowledge transfer for deep learning from private training data. *Proceedings of ICLR 2017*.
- Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K. and Erlingsson, U. (2018) Scalable private learning with PATE. *Proceedings of ICLR 2018*.
- Shokri, R., Stronati, M., Song, C. and Shmatikov, V. (2017) Membership inference attacks against machine learning models. *Proceedings of IEEE S&P; 2017*, pp. 3-18.
- Tramer, F. and Boneh, D. (2021) Differentially private learning needs better features (or much more data). *Proceedings of ICLR 2021*.
- Xu, L., Skoularidou, M., Cuesta-Infante, A. and Veeramachaneni, K. (2019) Modeling tabular data using conditional GAN. *Advances in Neural Information Processing Systems*, 32, pp. 1748-1759.

Information Processing Systems, 32, pp. 7335-7345.

Yu, D., Zhang, H., Chen, W. and Liu, T.Y. (2021) Large scale private learning via low-rank reparametrization. Proceedings of ICML 2021, pp. 12208-12218.

Declarations

Funding

This research was supported by the German Federal Ministry of Education and Research (BMBF) under grant 01IS22068 (PrivateAI), the Swiss National Science Foundation under grant 200021-212824 (PrivML), and the Estonian Research Council under grant PRG2748 (SecureML). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All privacy mechanism implementations, attack code, and benchmark results are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19614725. CheXpert, MIMIC-III (credentialed), and MovieLens are publicly available. Released under MIT License (code) and CC BY 4.0 (results).

Ethical Approval

This study used de-identified public datasets. MIMIC-III access obtained through PhysioNet credentialing. Membership inference attacks were conducted on public datasets in a controlled research setting. Approved by European Institute of AI Research Ethics Committee (protocol EIAI-ETH-2023-0218).

Appendix A

Privacy Mechanism Selection Guide

Decision tree for selecting privacy-preserving ML approaches based on privacy requirements, data availability, and deployment context.

Step 1: Privacy Requirement

- 1a. Formal DP guarantee needed --> DP-SGD, PATE, or synthetic data; go to Step 2
- 1b. Data must not leave device --> Federated learning + DP; go to Step 2
- 1c. Query data must stay private (inference) --> Homomorphic encryption
- 1d. Data sharing to multiple users --> Synthetic data (unlimited reuse after one-time generation)

Step 2: Acceptable Utility Loss and Resources

- 2a. < 5% loss acceptable + public unlabelled data available --> PATE at $\epsilon=2-4$ (best trade-off)
- 2b. 5-10% loss acceptable --> DP-SGD at $\epsilon=4-8$ (simpler; no teacher ensemble needed)
- 2c. > 10% loss acceptable --> DP-SGD at $\epsilon=1-2$ or synthetic at $\epsilon=1$

Step 3: Critical Reminders

- 3a. FL alone does NOT provide model-level privacy -- always add DP noise
- 3b. Track cumulative epsilon across all queries on the same data
- 3c. Validate empirically with membership inference in addition to formal guarantees
- 3d. Clinical/regulatory settings: target $\epsilon \leq 4$ (near-random attack AUC)