

Quantifying Membership Privacy Risk Using Shadow Model Membership Inference Attack: Observations from a Breast Cancer Use Case

Wenyi Tang¹, Konstantinos Koukoutegos^{1,2}, Akis Linardos^{1,2},
, and Spyridon Bakas^{1,2,3,4,†}

¹ Division of Computational Pathology, Department of Pathology and Laboratory Medicine, Indiana University School of Medicine, Indianapolis, IN 46202, USA

² Indiana University Melvin and Bren Simon Comprehensive Cancer Center, Indianapolis, IN 46202, USA

³ Departments of Radiology and Imaging Sciences; Neurological Surgery; Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN 46202, USA

⁴ Department of Computer Science, Luddy School of Informatics, Computing, and Engineering, Indiana University, Indianapolis, IN 46202, USA

[†] Corresponding author: spbakas@iu.edu

Abstract. Privacy regulations restrict sharing sensitive patient data across institutions, hindering collaborative medical research. Federated learning (FL) addresses this by keeping data local to the acquiring institution, but Membership Inference Attacks (MIAs) remain feasible even without direct access to training data. Differential privacy (DP) provides formal mathematical guarantees, yet translating its abstract ϵ parameter into concrete risk assessments that clinicians can interpret remains an open challenge. In this study, we quantify membership privacy risk from the attacker’s perspective by simulating a malicious FL participant who uses a fixed auxiliary data pool to attack target models via Shadow Model-based MIA. We instantiate this scenario in the EMBED breast cancer screening dataset, based on which we trained 300 ConvNeXt-Tiny models on binary classification without and with DP ($\epsilon \in \{3, 5, 8, 10\}$) with data level variance. Non-DP models exhibit consistent membership leakage with a mean ROC-AUC of 0.661 across training configurations, with the rare malignant class substantially more vulnerable (malignant-only AUC up to 0.938). DP training collapses malignant-only AUC to 0.500–0.574 across all privacy budgets, with effective protection achieved at $\epsilon = 10$. We further establish that 60 shadow models suffice for stable Likelihood Ratio Attack (LiRA) calibration, equivalent to $\sim 60\times$ the computation of a single target model. Our results provide an empirically grounded visualization of membership privacy risk and demonstrate, based on stable LiRA calibration, that moderate privacy budgets offer meaningful protection against inference attacks in medical imaging FL.

Keywords: Membership Inference Attack · Differential Privacy · Federated Learning · Privacy Quantification · Breast Cancer

1 Introduction

Medical research increasingly depends on large patient cohorts spanning electronic health records, diagnostic images, and laboratory results. These data bring huge benefits in terms of improved diagnosis and treatment, but also raise privacy concerns. Medical images have shown to carry privacy risks beyond structured clinical records. Mammograms, CT, and MRI scans contain biometric information and/or unique anatomical signatures that can re-identify individual patients, e.g., reported facial reconstruction from brain or head and neck scans [16]. Unlike sparse clinical records of obvious and easily recognizable protected information, medical images pack enormous detail across their voxels/pixels, making them harder to protect.

Data protection regulations such as the Health Insurance Portability and Accountability Act (HIPAA [15]) in the United States and the General Data Protection Regulation (GDPR [20]) in Europe, require data to be handled under strict confidentiality. These frameworks restrict sharing across institutions, creating a persistent tension between collaborative research and patient privacy.

Federated learning (**FL**) addresses this problem by enabling collaborative model training without centralized data aggregation [4]. It has already been deployed in real multi-site healthcare settings [13, 6], with FL-as-a-service platforms enabling hospitals to jointly train models on distributed patient data while meeting regulatory requirements. Each participating institution trains a shared model locally and exchanges only model updates, so raw patient images never leave the originating site. This separation of data storage from computation fits the regulatory and ethical constraints of collaborative medical research.

Membership inference is the problem of determining whether a specific data record was part of a model’s training dataset [10]. Such Membership Inference Attacks (**MIA**) are applicable to FL as they remain feasible even without direct access to the actual training data. Previous work [10, 17] shows that model parameters and prediction outputs can leak detailed membership information, exposing individual patients to re-identification even within FL pipelines. Therefore, privacy leakage through machine learning models can be considered a compliance obligation for healthcare systems, and not only a technical concern.

Differential privacy (**DP**) provides formal mathematical bounds on this leakage by limiting the amount that any single training example can influence the model’s output [7]. DP has become the primary framework for privacy accounting in machine learning [1, 22], and its combination with FL has been studied extensively [19, 8, 4]. However, what remains less well understood is how to translate the formal DP guarantee into risk assessments that medical practitioners/clinicians can interpret and act upon, since the privacy parameter epsilon (ϵ) maps poorly to concrete clinical privacy expectations.

In this study, we quantify membership privacy risk from the attacker’s perspective by simulating a malicious FL participant who holds a fixed auxiliary data pool and uses shadow model-based LiRA to attack target models trained by other participants. We sweep training set sizes, malignant case percentages, and DP privacy budgets to characterize how these factors affect membership

leakage. Our results show that non-DP models are consistently vulnerable to MIA, with minority cases substantially easier to detect than majority cases, and that DP training eliminates the membership signal even at moderate privacy budgets. We further characterize the computational cost, establishing that 30 shadow models per side suffice for stable LiRA calibration, which is about $60\times$ training time comparing to the target model training.

Specifically, our contributions are three-fold: *(i)* We characterize non-DP membership leakage under a range of training set sizes and malignant case percentages, showing that minority classes are disproportionately vulnerable to MIA, *(ii)* We quantify the protective effect of DP training, demonstrating that even moderate privacy budgets ($\epsilon = 10$) collapse the membership signal to near-chance for both majority and minority classes, and *(iii)* We establish a practical shadow model budget per candidate sample for stable LiRA calibration in the case of breast cancer research.

2 Related Work

First introduced by Shokri et al. [17], MIAs determine whether a specific data record was used to train a target model, posing a fundamental privacy threat even in black-box settings. Carlini et al. formalized the Likelihood Ratio Attack (LiRA), which evaluates membership leakage at low false-positive rates and achieves an order-of-magnitude improvement over prior methods [5]. In the context of FL, Gu et al. proposed CS-MIA, which exploits prediction confidence dynamics across FL training rounds to infer membership from both local and global adversary perspectives [9]. Rao et al. provide a structured taxonomy of MIAs and defenses across centralized and FL paradigms [14].

DP provides a formal bound on information leakage: a randomized algorithm is (ϵ, δ) -DP if the probability of any output changes by at most a factor of e^ϵ when a single record is added or removed. Abadi et al. established the practical framework for DP in deep learning through DP stochastic gradient descent (DP-SGD), which clips per-sample gradients and adds calibrated noise to guarantee (ϵ, δ) -DP [1]. Deploying DP in deep learning involves a fundamental utility-privacy trade-off; Tramèr and Boneh showed that DP training requires either large private datasets or access to high-quality public features to match non-private baselines [18]. Bagdasaryan et al. demonstrated that DP can exacerbate existing biases, causing disproportionate accuracy degradation on underrepresented subgroups, which is especially relevant to medical imaging cohorts with naturally skewed class distributions [2]. Fu et al. provide a systematic review of DP in FL, organizing methods by protection granularity across FL scenarios [7].

3 Materials and Methods

3.1 Data

The EMBED dataset is a large-scale real-world multi-site breast screening mammography collection released through the AWS Open Data Program [11]. We

assess the task of binary classification between benign and malignant cases, by mapping BI-RADS assessments with no suspicion of malignancy to benign and those with confirmed or high-probability cancer to malignant. Here, we use data from hospitals with both classes available, comprising 345,220 images, with a severe class imbalance of 343,818 benign to 1,402 malignant. Standard single-view mammograms are retained while implant-displaced, spot, and magnification views are excluded to avoid acquisition artifacts. All DICOM images are resized via bicubic interpolation to 256×256 and converted to NIFTI files.

3.2 Classification Model

For the classification, we use a ConvNeXt-Tiny model initialized with ImageNet-pretrained weights for two-class output. The backbone and classification head are optimized jointly with Adam, using learning rates of 1×10^{-4} and 2×10^{-4} for the backbone and head, respectively. Cross-entropy loss with label smoothing of 0.1 and class weighting is applied over 30 epochs, with balanced accuracy used for model checkpoint selection. Training uses random horizontal flipping, rotation, and affine augmentation with a batch size of 128. We use the Opacus library [21] to implement DP training of the target model. Privacy accounting follows the PRV accountant with Poisson sampling, per-sample gradient clipping at a maximum norm of 212.8, and a target $\delta = 1 \times 10^{-5}$. We train DP variants at target ϵ values of 3, 5, 8, and 10. In our setting, both the target and all shadow models share the identical ConvNeXt-Tiny architecture and training procedure; they only differ in the data they are trained on. We do not apply any remedy to compensate for DP-induced performance degradation, such as oversampling, SMOTE or synthetic data generation, [12], since model performance is not the main focus of this work.

3.3 Shadow Model-based MIA

Membership inference is inherently model- and sample-specific, making it hard to standardize privacy risk in a portable way. Shadow model-based attacks provide a general solution: given any auxiliary dataset drawn from a similar distribution, they extract membership signals via surrogate training and produce standard binary classification metrics, such as the true positive rate at fixed false-positive rates (TPR@FPR) and ROC-AUC. This makes them well-suited for quantifying privacy risk in FL scenarios, where a malicious participant naturally holds a local dataset and can query models trained by other participants.

The MIA seeks to determine whether a given candidate sample x was included in the training set of a target model M_t . The attacker has black-box query access to M_t and holds a fixed auxiliary data pool D_a drawn from a similar distribution to the target model’s training data. The attacker knows neither the target training set D_t nor the training procedure of M_t , only the model’s output.

To simulate how M_t behaves on members vs non-members, the attacker trains shadow models on random fixed-size subsets of D_a . All shadow models use the same architecture and training procedure as M_t . For a given candidate sample

x , a shadow model is an IN model if its training subset contains x , and an OUT model otherwise. By training many IN and OUT models, the attacker builds two empirical distributions of model behavior: one for samples the model has seen during training and one for those it has not.

We adopt the Likelihood Ratio Attack (LiRA) as presented in [5]. For each candidate x , the attacker queries all IN and OUT shadow models and records their loss values. A Gaussian is fitted to each set, yielding two parametric distributions over the loss: $\mathcal{L}_{\text{in}}$ for IN models and $\mathcal{L}_{\text{out}}$ for OUT models.

To infer the membership of x , the attacker queries M_t to obtain the target-model loss s_t , then computes the likelihood ratio $\Lambda = P(s_t | \mathcal{L}_{\text{in}}) / P(s_t | \mathcal{L}_{\text{out}})$. If Λ exceeds a chosen threshold, x is predicted as a member. The threshold is calibrated on a held-out set of known non-members to control the FPR. The overall attack performance is evaluated using ROC-AUC and the TPR at a fixed FPR, with the threshold calibrated on a held-out set of known non-members to control the FPR at a target operating point.

3.4 Experimental Design

For the auxiliary data pool D_a we use the data from the EMBED site 5, containing approximately 53,000 mammograms. The target model M_t is trained on data from the remaining sites (1, 2, 6). Each shadow model is trained on a fixed-size subset of 1,000 samples randomly drawn from D_a (with a fixed malignant case percentage around 5%), and per candidate sample we train 64 IN and 64 OUT shadow models to estimate the loss distributions. This setup maps directly to a FL scenario, where a malicious participant with data from one site (the auxiliary pool) queries a model collaboratively trained by other sites (the target model).

We evaluate the privacy risk of M_t using shadow model-based MIAs under different D_a configurations. The experiments are designed to quantify how the size of D_a and its similarity to the training data of M_t affect the success of MIAs, thereby providing insights into membership privacy risks in practical settings.

4 Results and Observations

We train 60 non-DP ConvNeXt-Tiny target models spanning five training set sizes (1,000, 3,000, 5,000, 8,000, 10,000) and four malignant case percentages (5%, 10%, 15%, 20%), with three replicates per cell. Each target is attacked via LiRA with 64 IN and 64 OUT shadow models, evaluating 500 samples per target (250 members, 250 non-members).

Rare Class Samples are More Vulnerable to MIA. Fig.1a shows the cell-wise mean ROC-AUC across all 20 size-ratio combinations. LiRA recovers consistent membership signal across the full grid, with an overall mean AUC of 0.661 and cell means ranging from 0.608 (8,000 samples, 20% malignant) to 0.729 (3,000 samples, 5% malignant). AUC decreases as the training set grows, confirming the dilution effect, while higher malignant case percentages further suppress attack success. This suggests that larger training sets reduce per-sample

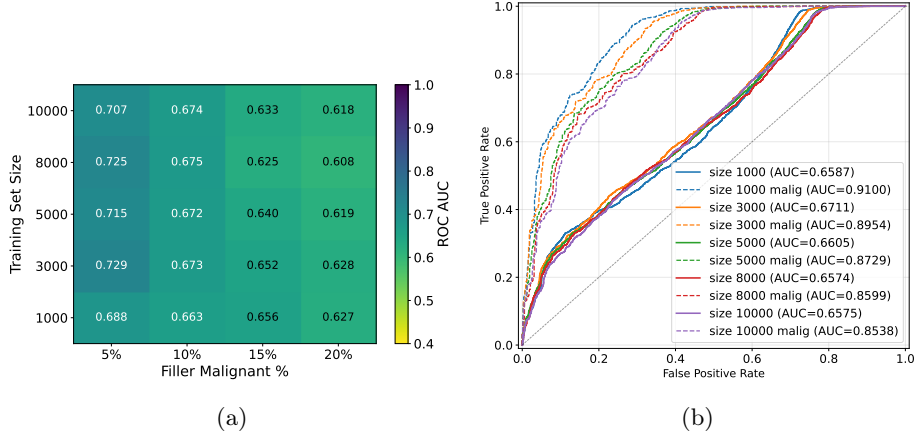


Fig. 1: Non-DP membership leakage across training configurations. (a) Mean LiRA ROC-AUC across training set sizes and malignant case percentages. (b) Overall versus malignant-only ROC curves.

memorization and that harder class compositions make individual membership harder to isolate.

Fig. 1b overlays the overall and malignant-only ROC curves, revealing a sharp class-conditional disparity. Malignant-only AUC reaches 0.938 at the easiest setting (1,000 samples, 5% malignant) and stays above 0.807 even under the hardest configuration (8,000 samples, 20% malignant). At a 5% FPR, the TPR on malignant samples is 46.9%, more than double the overall TPR of 20.0%. This indicates that models memorize rare class examples more easily than frequent ones, making minority cases substantially easier targets for MIA.

DP Significantly Reduces Membership Leakage To quantify the protective effect of DP, we repeat the sweep with DP-trained ConvNeXt-Tiny models at $\epsilon \in \{3, 5, 8, 10\}$, fixing the malignant case percentage at 10%. We train 75 target models (5 regimes $\times$ 5 training sizes $\times$ 3 replicates), each attacked via LiRA with 64 IN and 64 OUT shadow models, evaluating 500 samples per target.

Fig. 2 shows the malignant-only AUC per training size for each privacy regime. Non-DP malignant-only AUC ranges from 0.866 to 0.927, confirming that rare-class membership is highly detectable without protection. Under DP, the malignant-only AUC collapses to near-chance across all regimes, ranging from 0.500 to 0.574. The gap between non-DP and even the weakest DP setting ($\epsilon = 10$) is dramatic, and further tightening ϵ produces no meaningful additional reduction. This demonstrates that gradient clipping and noise addition eliminate the membership signal that LiRA exploits, providing effective protection at moderate privacy budgets regardless of training set size.

Shadow Model Budgeting. LiRA needs shadow model loss distributions to estimate IN and OUT membership hypotheses, but training and storing shadow

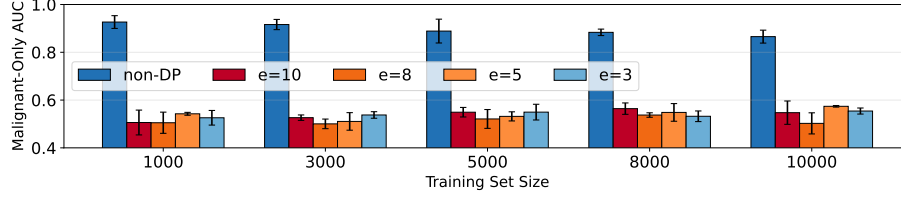


Fig. 2: Malignant-only LiRA ROC-AUC grouped by training set size across privacy regimes. Non-DP bars dominate while all DP-regime bars collapse to near-chance. Error bars show ± 1 standard deviation over 3 replicates.

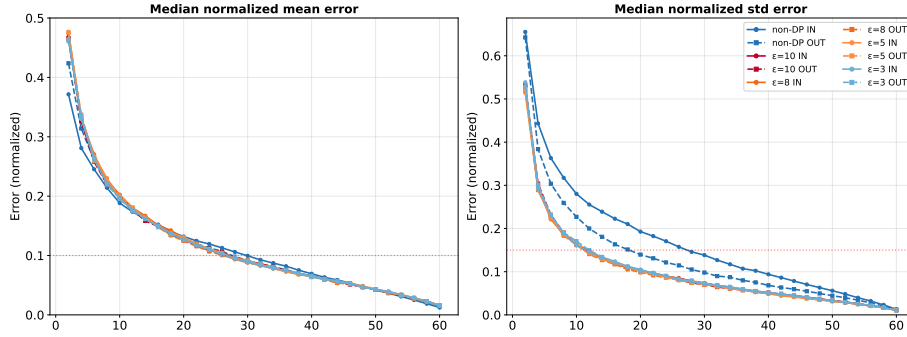


Fig. 3: Median normalized movement of Gaussian μ and σ estimates as a function of shadow count. Dashed lines track OUT distributions; solid lines track IN distributions. Horizontal dotted lines mark the stability threshold: a budget passes when the median normalized error of the mean falls below 0.10 and the standard deviation falls below 0.15, i.e., both estimates are within 10–15% of one reference- σ from their convergent values.

models is expensive. We therefore ask: *how many shadow models are needed before the fitted loss distributions stop changing as more shadows are added?*

For each privacy pool (non-DP and DP at $\varepsilon = 3, 5, 8, 10$), we fit Gaussian μ and σ estimates at every integer budget from 2 to 64 shadows per side, using five deterministic seed selections. The budget-64 fit serves as the empirical reference, and errors for each smaller budget are normalized by the reference σ . All comparisons are privacy-matched (non-DP targets use non-DP shadows, each DP ε uses the corresponding DP shadow pool) and use fixed targets only.

Fig.3 shows the parameter stability curves across privacy pools. The IN distribution stabilizes by approximately 28 shadows, while the OUT distribution stabilizes slightly earlier. The non-DP pool drives the 30-shadow recommendation because its shadow losses exhibit the largest variance; DP-trained models produce more concentrated loss distributions and therefore reach stability at slightly lower budgets, even though this does not necessarily imply better attack

performance due to the involvement of DP. These results indicate that approximately 30 IN and 30 OUT shadow models per candidate sample suffice for stable LiRA calibration, equivalent to roughly $60\times$ the computation of training a single target model. In practice, the cost can be amortized when targeting multiple candidate samples, since a shadow model trained for one sample also covers neighboring samples in the auxiliary pool. Further efficiency gains are achievable through shared shadow model pools or mixture-of-experts architectures that reuse sub-networks across shadow models [3].

We further verify these findings at the raw distribution level using Wasserstein distance and Kolmogorov-Smirnov statistics between pooled shadow-loss distributions. Additional results, including OUT-only budget extensions to 128 shadows and raw empirical stability analyses, are provided in Appendix A.

5 Discussion

In this study, we quantified membership privacy risk in a FL setting by simulating a malicious participant who uses a fixed auxiliary data pool to attack a target model via LiRA. We instantiated this scenario with the EMBED breast cancer screening dataset, using data from one hospital site as the auxiliary pool and the remaining sites as the target training data.

We observe non-DP targets being highly vulnerable to MIA. The mean ROC-AUC across 60 non-DP targets was 0.661, with malignant-only AUC reaching 0.938. Lower leakage was observed for the DP-trained models. Across all ϵ values (3, 5, 8, 10), the malignant-only AUC dropped to 0.500–0.574, with individual DP targets showing no effective ability to determine membership status.

In terms of the attack’s computational cost, we established that 30 shadow models per side suffice for stable LiRA distributions and provided a closed-form formula to estimate the number of equivalent trainings a given attack query.

Our findings are limited to the ConvNeXt-Tiny model structure, EMBED data, LiRA, and a single-site auxiliary pool with a fixed malignant percentage. Further experimentation is needed to assess if the 30/30 shadow budget generalizes across architectures, training sizes, attack methods, and site distributions.

Future work should also evaluate if site-level distributional differences and rarer malignant class compositions alter the protective effect of DP, as well as how DP performance remedies (such as oversampling or SMOTE [12]) affect membership leakage.

Acknowledgments. This work was partially supported by the NIH/NCI/ITCR under award number U24CA279629. The content is solely the authors’ responsibility and does not represent the official views of the NIH.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. pp. 308–318 (2016)
2. Bagdasaryan, E., Poursaeed, O., Shmatikov, V.: Differential Privacy Has Disparate Impact on Model Accuracy. In: Advances in Neural Information Processing Systems. vol. 32 (2019)
3. Bai, L., Ye, Q., Zhang, X., Zhang, S., Liang, Z., Xu, J., Hu, H.: Toward efficient inference attacks: Shadow model sharing via mixture-of-experts. Advances in Neural Information Processing Systems **38**, 138757–138785 (2026)
4. Bakas, S., Li, X., Shah, P., Roth, H.R.: Federated learning in healthcare: From research to real-world deployment. Annual Review of Biomedical Engineering **28** (2026)
5. Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., Tramèr, F.: Membership Inference Attacks From First Principles. In: 2022 IEEE Symposium on Security and Privacy (SP). pp. 1897–1914 (2022). <https://doi.org/10.1109/SP46214.2022.9833649>
6. Dayan, I., Roth, H.R., Zhong, A., Harouni, A., Gentili, A., Abidin, A.Z., Liu, A., Costa, A.B., Wood, B.J., Tsai, C.S., et al.: Federated learning for predicting clinical outcomes in patients with covid-19. Nature medicine **27**(10), 1735–1743 (2021)
7. Fu, J., Hong, Y., Ling, X., Wang, L., Ran, X., Sun, Z., Wang, W.H., Chen, Z., Cao, Y.: Differentially Private Federated Learning: A Systematic Review. ACM Computing Surveys **58**(11), 271:1–271:38 (2026). <https://doi.org/10.1145/3801079>
8. Fu, J., Hong, Y., Ling, X., Wang, L., Ran, X., Sun, Z., Wang, W.H., Chen, Z., Cao, Y.: Differentially private federated learning: A systematic review. ACM Computing Surveys **58**(11), 1–38 (2026)
9. Gu, Y., Bai, Y., Xu, S.: CS-MIA: Membership Inference Attack Based on Prediction Confidence Series in Federated Learning. Journal of Information Security and Applications **67**, 103201 (2022). <https://doi.org/10.1016/j.jisa.2022.103201>
10. Hu, H., Salicic, Z., Sun, L., Dobbie, G., Yu, P.S., Zhang, X.: Membership inference attacks on machine learning: A survey. ACM Computing Surveys (CSUR) **54**(11s), 1–37 (2022)
11. Jeong, J.J., Vey, B.L., Bhimireddy, A., Kim, T., Santos, T., Correa, R., Dutt, R., Mosunjac, M., Oprea-Ilie, G., Smith, G., et al.: The emory breast imaging dataset (embed): A racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images. Radiology: Artificial Intelligence **5**(1), e220047 (2023)
12. Johnson, J.M., Khoshgoftaar, T.M.: Survey on deep learning with class imbalance. Journal of big data **6**(1), 27 (2019)
13. Pati, S., Baid, U., Edwards, B., Sheller, M., Wang, S.H., Reina, G.A., Foley, P., Gruzdev, A., Karkada, D., Davatzikos, C., et al.: Federated learning enables big data for rare cancer boundary detection. Nature communications **13**(1), 7346 (2022)
14. Rao, B., Zhang, J., Wu, D., Zhu, C., Sun, X., Chen, B.: Privacy Inference Attack and Defense in Centralized and Federated Learning: A Comprehensive Survey. IEEE Transactions on Artificial Intelligence **6**(2), 333–353 (2025). <https://doi.org/10.1109/TAI.2024.3363670>
15. Rights (OCR), O.f.C.: HIPAA for Individuals, <https://www.hhs.gov/hipaa/for-individuals/index.html>

16. Schwarz, C.G., Kremers, W.K., Therneau, T.M., Sharp, R.R., Gunter, J.L., Vemuri, P., Arani, A., Spychalla, A.J., Kantarci, K., Knopman, D.S., et al.: Identification of anonymous mri research participants with face-recognition software. *New England Journal of Medicine* **381**(17), 1684–1686 (2019)
17. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE symposium on security and privacy (SP). pp. 3–18. IEEE (2017)
18. Tramèr, F., Boneh, D.: Differentially Private Learning Needs Better Features (or Much More Data). arXiv preprint arXiv:2011.11660 (2021). <https://doi.org/10.48550/arXiv.2011.11660>
19. Truex, S., Liu, L., Chow, K.H., Gursay, M.E., Wei, W.: Ldp-fed: Federated learning with local differential privacy. In: Proceedings of the third ACM international workshop on edge systems, analytics and networking. pp. 61–66 (2020)
20. Voigt, P., Von dem Bussche, A.: The eu general data protection regulation (gdpr). A practical guide, 1st ed., Cham: Springer International Publishing **10**(3152676), 10–5555 (2017)
21. Yousefpour, A., Shilov, I., Sablayrolles, A., Testuggine, D., Prasad, K., Malek, M., Nguyen, J., Ghosh, S., Bharadwaj, A., Zhao, J., Cormode, G., Mironov, I.: Opacus: User-friendly differential privacy library in PyTorch. arXiv preprint arXiv:2109.12298 (2021)
22. Ziller, A., Usynin, D., Braren, R., Makowski, M., Rueckert, D., Kaissis, G.: Medical imaging deep learning with differential privacy. *Scientific Reports* **11**(1), 13524 (2021)

A Detailed Shadow Model Budgeting Results

We present the remaining analyses from the stable baseline privacy sweep that could not be included in the main text due to space constraints. These results provide additional evidence for the recommended shadow model budgets.

The Gaussian parameter analysis in the main text asks whether the fitted μ and σ estimates stabilize. The empirical analysis asks whether the raw shadow-loss distributions themselves stabilize in shape. For each target, side, seed, and budget, we pool the selected shadow losses across target samples and compare the resulting empirical distribution with the matching pooled reference distribution. We use normalized Wasserstein distance and the Kolmogorov-Smirnov statistic as comparison metrics.

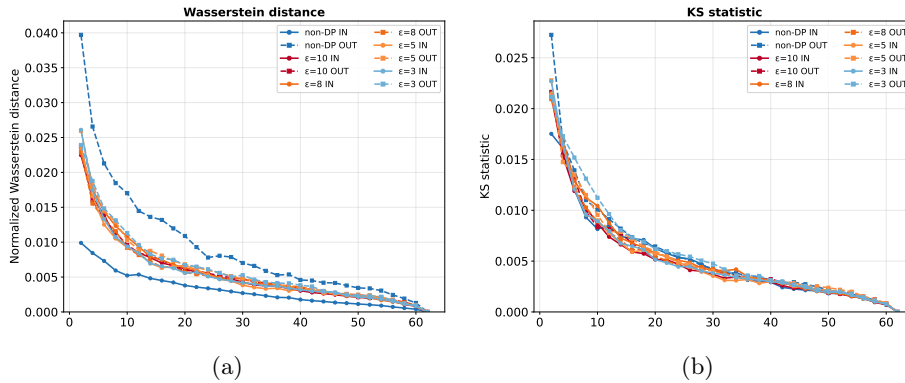


Fig. 4: Empirical shadow-loss distribution stability across shadow budgets. (a) Normalized 1D Wasserstein distance. (b) Kolmogorov-Smirnov statistic.

Fig. 4 shows that the empirical shadow-loss distributions stabilize in a range consistent with the Gaussian parameter results, typically between 25 and 35 shadows depending on the privacy pool, which is consistent with the Gaussian parameter findings in the main text.

We further verify these findings by measuring LiRA attack effectiveness directly as a function of shadow budget.

Fig. 5 shows that AUC converges from approximately 0.58 at budget 2 to 0.66 at the reference budget 62. Diminishing returns set in around budget 20–30; beyond 30, each additional pair of shadows adds less than 0.001 to AUC, confirming that the 30-shadow recommendation from the main text captures the point at which additional computation yields negligible improvement in attack performance.

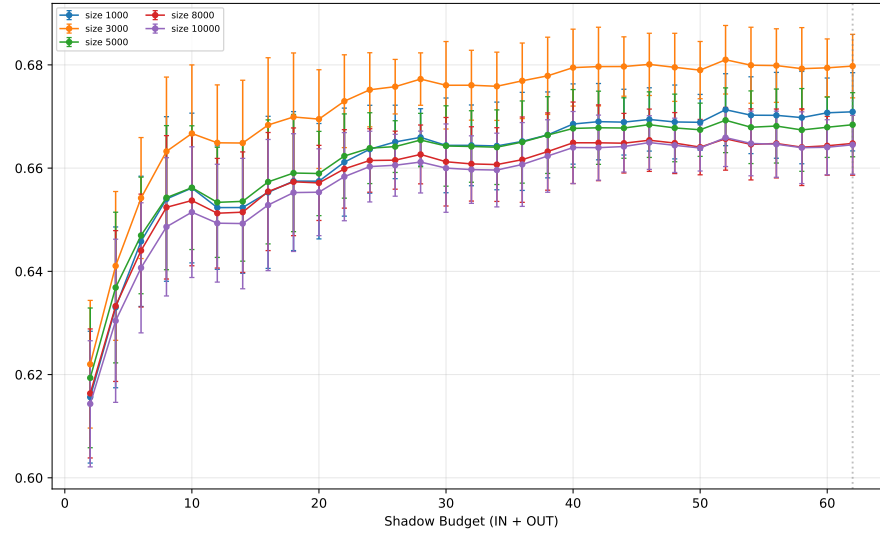


Fig. 5: LiRA ROC-AUC as a function of $\text{IN} + \text{OUT}$ shadow model count. Each line represents one training set size, with error bars showing ± 1 standard deviation across 10 seeds. The dashed vertical line marks the reference budget of 62.