



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

W. Tang^{1,2}, K. Koukouteagos^{1,2}, A. Linardos^{1,2}, S. Bakas^{1,2,3}

1. Division of Computational Pathology, Department of Pathology and Laboratory Medicine, Indiana University (IU) School of Medicine, Indianapolis, USA
2. IU Melvin and Bren Simon Comprehensive Cancer Center, Indianapolis, USA 3. Luddy School of Informatics, Computing, and Engineering, IU, USA

Background

- Regulations restrict sharing patient images[1][2]; federated learning (FL) keeps data local but still exposes model outputs.
- Membership inference attacks (MIA) ask whether a record was in the training set - a re-identification risk.
- Differential privacy (DP) gives formal guarantees, but ϵ is hard for clinicians to interpret as concrete risk.
- We quantify risk from the attacker side: a malicious FL participant with a fixed auxiliary pool attacks other sites' model with shadow-model LiRA (Likelihood-Ratio Attack).

Materials & Methods

EMBED[3] breast cancer mammography



Benign sample

Malignant sample

345,220 images from 5 hospitals

Class imbalance: 343,818 benign vs **1,402 malignant**

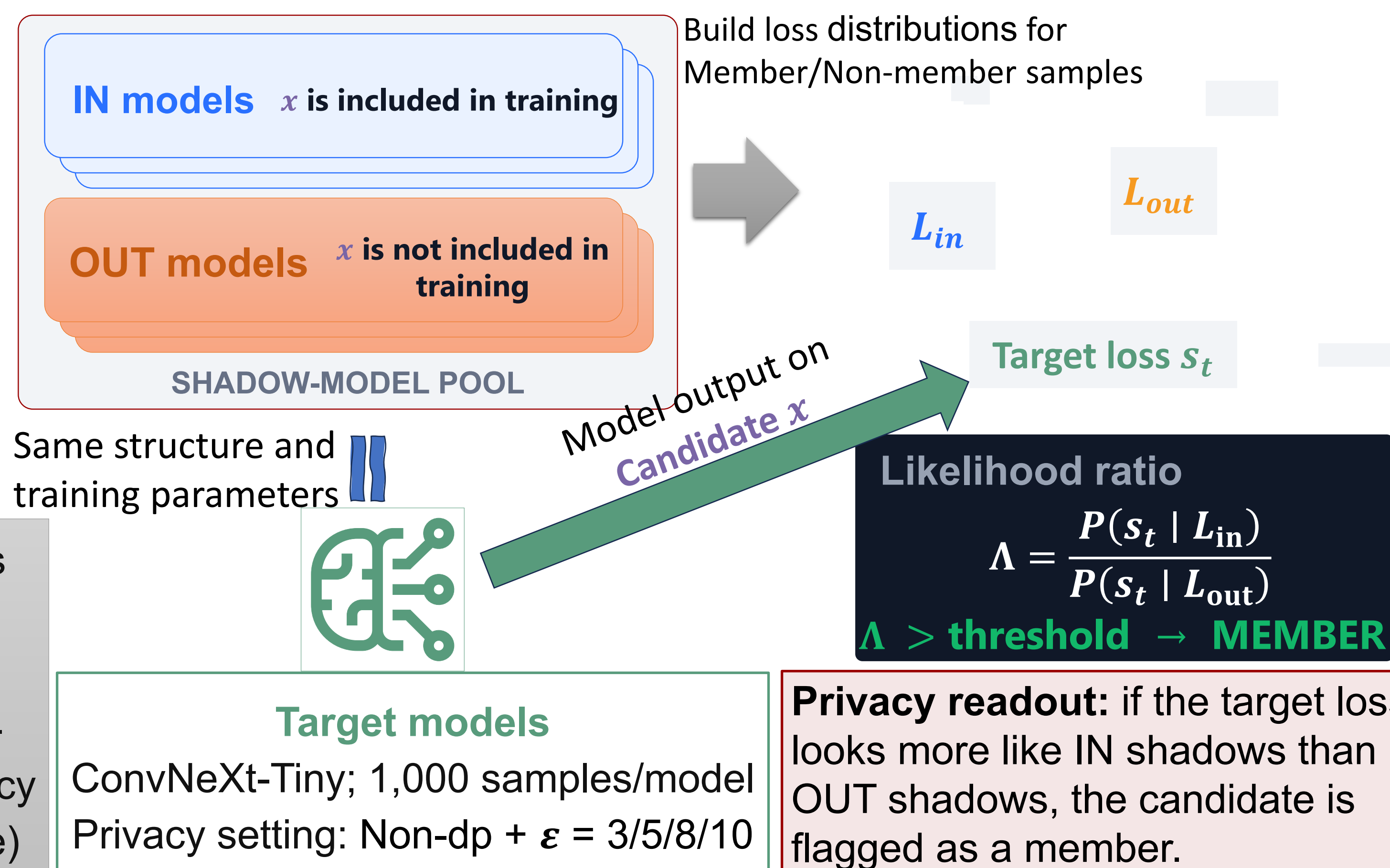
Standard single-view mammograms retained; resized to **256×256 NIFTI**.

Site split: site 5 = attacker auxiliary pool; sites 1, 2, 6 = target training data.

GOAL for MIA – A binary classification: Candidate $x \in$ Target Model Training Set?

Shadow Model LiRA:

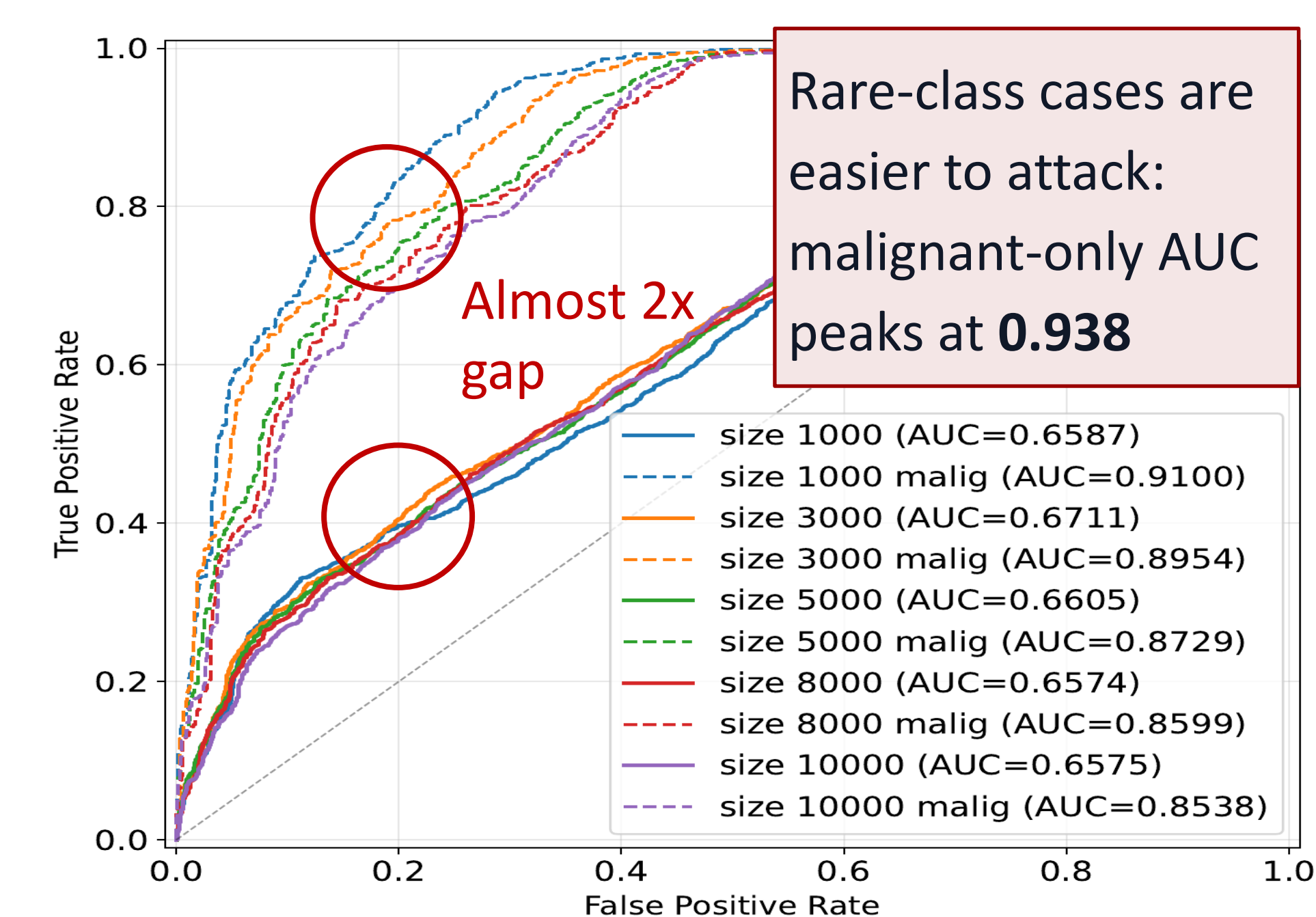
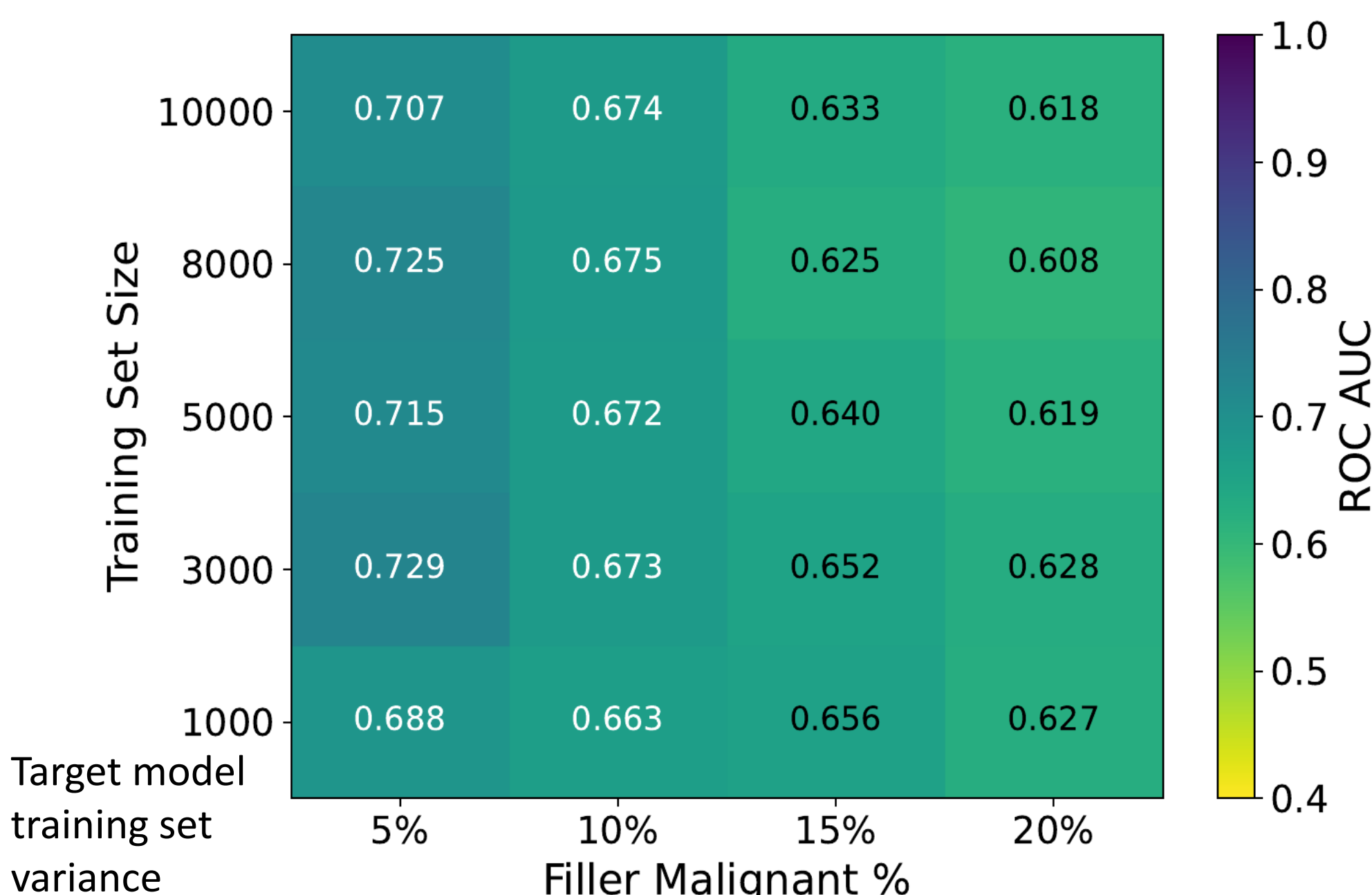
- For each candidate imag, train shadow models that either include it (INx64) or exclude it (OUTx64).
- Query these models to estimate the expected loss distributions for members and non-members.
- Compare the target model's loss to both distributions.



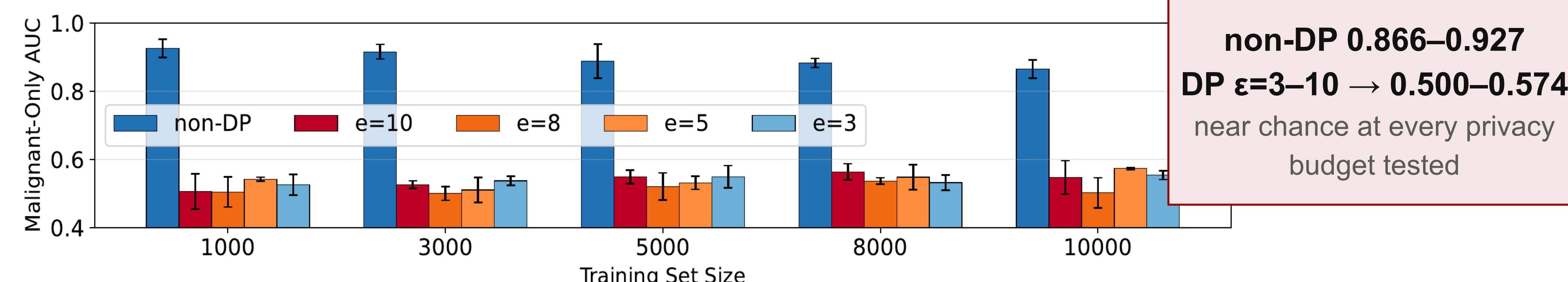
Observations

A. Without DP, membership leaks

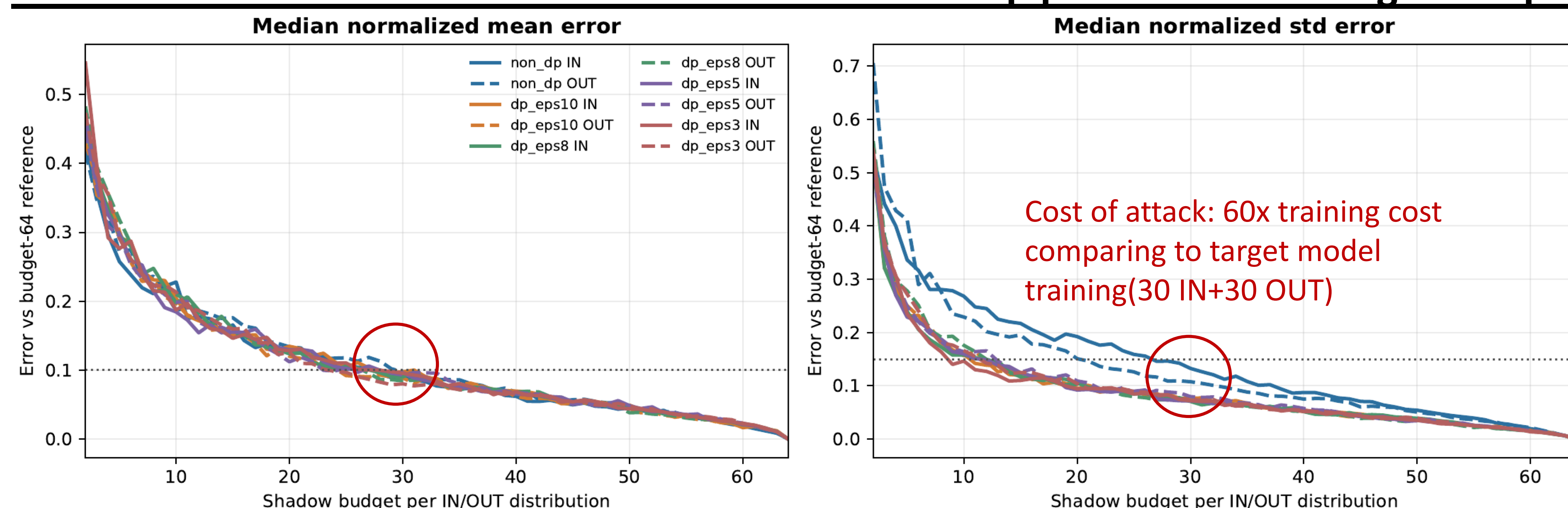
Unprotected models leak consistently: mean attack's AUC **0.661**



B. DP mitigates membership leakage (Even with moderate ϵ settings)



C. Number of shadow models for a stable membership prediction of 1 target sample



30 IN + 30 OUT
per candidate $\approx 60 \times$ one target training
Can be amortizable through shared shadow pools

- LiRA fits an IN and an OUT loss distribution per candidate; too few shadows make the fit unstable.
- Gaussian mean / std errors flatten within $\sim 25-35$ shadows; beyond ~ 30 pairs, each extra pair adds under 0.001 AUC.
- Non-DP pools drive the recommendation because their shadow losses have the largest variance.

Summary & Future Outlook

- Unprotected FL models leak membership consistently; rare malignant cases are disproportionately exposed.
- Moderate DP ($\epsilon = 10$) collapses malignant-only AUC to 0.500-0.574 — near guessing across sizes.
- Stable LiRA calibration needs about 30 IN + 30 OUT shadows per candidate ($\sim 60 \times$ target training).
- Scope: Model, EMBED, LiRA, single-site auxiliary pool; generality across architectures/sites remains future work.

References

[1] Rights, O.f.C.: HIPAA for Individuals, <https://www.hhs.gov/hipaa/for-individuals/index.html>
[2] 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)
[3] Jeong, J.J., 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)
[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] Pati, S., et al.: Federated learning enables big data for rare cancer boundary detection. Nature communications, (2022)
[6] 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. pp. 1897–1914 (2022).
[8] 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. (2016)

Scan me - Paper answers back!



Poster • Full Paper • Ask AI