

---

# Reconciling privacy and accuracy in AI for medical imaging

---

In the format provided by the  
authors and unedited

# Reconciling Privacy and Accuracy in AI for Medical Imaging - Supplementary Material

## A Threat models

Here, we provide a more concrete explanation of how the risks of re-identification through a data reconstruction attack are moderated by Differential Privacy (DP). Concretely, we avail ourselves of the framework of reconstruction robustness (ReRo), which will allow us to formulate an upper bound on the success rate of data reconstruction attacks against Artificial Intelligence (AI) models trained with DP under the specific threat models discussed below.

ReRo was introduced by [1]. It is a guarantee pertaining to an algorithm which processes sensitive data, e.g. an AI model trained with DP. Intuitively, if at most a proportion  $0 \leq \gamma \leq 1$  of the total samples used to train the model can be successfully reconstructed by an adversary with a reconstruction error lower than  $\eta \geq 0$ , then the model satisfies  $(\eta, \gamma)$ -ReRo. Recent works have proven that all models trained with DP automatically satisfy ReRo and that for certain settings, it is possible to directly quantify the upper bound for  $\gamma$  [1, 2, 3]. In other words, DP automatically provides strong and quantifiable protection against data reconstruction attacks.

We study the ReRo guarantees of models trained with DP under three distinct sets of assumptions about the capabilities of the adversary, i.e. three distinct threat models:

1. The **worst-case** threat model: This corresponds to the adversary usually considered in DP, namely one who has unbounded computational abilities, can deeply manipulate the model's (hyper-)parameters and has access to the target image itself, which they can use to attack the model. Evidently, this threat model is not realistic (as an adversary who has access to the target point would not need to attack the model), but is used to provide guarantees when "all bets are off", i.e. in a so-called *privacy auditing* scenario when one is interested in the absolute worst-case behaviour of a system.
2. The **relaxed** threat model [4]: This threat model is still quite pessimistic, as it still assumes unbounded computational ability and access to model (hyper-)parameters. However, this adversary only has restricted access to the dataset, notably, they cannot use the target image itself to attack the model. Although it renders this threat model more appropriate for scenarios where the dataset can be safely assumed to be kept secure, e.g. in a hospital's database, it still makes assumptions, which are not encountered in any practical scenario. Most importantly, the adversary has a black-box reconstruction algorithm, which yields either a perfect reconstruction or fails, and the only decision the adversary has to make is whether the reconstruction was indeed the target data. The term *relaxed* stems from security research, where a *relaxation* denotes a weakening of a security assumption (as can be seen in e.g. [1, 2, 5, 6, 7, 8]).
3. The **realistic** threat model: The final threat model considers an adversary with unbounded computational ability and the power to manipulate model (hyper-)parameters but only very limited access to information about the dataset. For example, the adversary can know the dimensions of the images to be reconstructed but not any of their contents. We note that even this threat model is relatively pessimistic, as it assumes an active adversary who is trusted and therefore the actions are not reviewed by other participants. Such adversaries could manipulate the model to their advantage in order to reconstruct training data. We term this threat model as realistic as it stems from federated learning research, although in many cases, this could be detected simply by inspecting the model architecture. Nonetheless, we use this threat model as it is conceivable that such adversaries can exist in, e.g. federated learning settings in untrustworthy consortia,

which are common in real-world settings. Protection against this threat model generalizes to all weaker threat models, such as black-box attacks after training and adversaries lacking the ability to manipulate the deep learning model or its hyperparameters.

Of note, it is possible to provide theoretical bounds on the reconstruction attack success rate in both the worst-case and the relaxed threat models using the techniques presented in [2, 3]. For the realistic threat model, we assess the attack success rate empirically. A concise overview of the aforementioned threat models is provided in Table 1.

In summary, while conservative (i.e. worst-case or relaxed) threat models are important tools in security research because they allow one to derive closed-form bounds on the attack success rate of very powerful adversaries, such threat models are in most reasonable scenarios too pessimistic.

## B Setup

### B.1 Dataset Description

Here we outline our rationale for choosing specific datasets for our experiments. We identified four characteristics of medical imaging datasets, which reoccur frequently: (1) Datasets are often **small** compared to non-medical datasets. For example, most medical AI algorithms, which are currently approved by the US Food and Drug Administration (FDA), are trained on less than 1 000 data samples [9]. (2) Diagnoses occur with very different frequencies, leading to often **imbalanced** datasets skewed toward more common diagnoses. In segmentation tasks, this may happen due to different spatial extensions of objects. (3) While natural images are all captured with standard cameras as RGB images, medical images are from **multi-modal** imaging devices such as computed tomography (CT), magnetic resonance imaging (MRI), or ultrasound.

In this study, we aim to give a broad discussion of settings in medical AI. Hence, we have chosen three datasets, which encompass the above-discussed scenarios (c.f. Table 2).

1. The RadImageNet dataset [10] contains over 1.3 million 2D images with CT, MRI, and ultrasound scans representing three imaging modalities with 165 classification targets, which are highly imbalanced.
2. The HAM10000 dataset [11] is a collection of 10 000 skin lesion RGB images spread across seven categories. We intentionally amplified the class imbalance to a strong but not untypical 80 : 20 class ratio by merging classes based on the need for immediate treatment (see Section 4.1).
3. Lastly, we use the MSD Liver dataset [12, 13], a demanding image-to-image task involving just 131 CT scans annotated at voxel level. Given the small number of available training samples as well as a segmentation task (i.e., per-pixel classification) with tumours only encompassing a tiny fraction of each scan, it represents a very challenging medically relevant task.

To the best of our knowledge, no prior work shows the performance of AI models trained under formal privacy guarantees on such a comprehensive and large dataset as RadImageNet or a 3D image-to-image task as MSD Liver represents.

### B.2 Metrics

To measure the performance of the models on classification tasks, we use Matthews’ Correlation Coefficient (MCC) [14]. Opposed to more frequently used metrics such as accuracy or  $F_1$ -score, it incorporates the entire confusion matrix and, by that, is extremely robust against any class imbalance [15]. It is also better interpretable as for random predictions it is 0 and for perfect predictions 1, whereas the accuracy depends on the class distribution. For the segmentation task, we measure the class-wise Dice score of the 3D volumes and report the average over all volumes for the liver and tumour, as they are the targets of interest in our task. A perfect prediction yields a 100% Dice score.

## References

- [1] Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1138–1156. IEEE, 2022.
- [2] Jamie Hayes, Saeed Mahloujifar, and Borja Balle. Bounding training data reconstruction in dp-sgd. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [3] Georgios Kaissis, Jamie Hayes, Alexander Ziller, and Daniel Rueckert. Bounding data reconstruction attacks with the hypothesis testing interpretation of differential privacy. *Theory and Practice of Differential Privacy*, 2023.
- [4] Georgios Kaissis, Alexander Ziller, Stefan Kolek, Anneliese Riess, and Daniel Rueckert. Optimal privacy guarantees for a relaxed threat model: Addressing sub-optimal adversaries in differentially private machine learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [5] Christina Brzuska, Marc Fischlin, Nigel P Smart, Bogdan Warinschi, and Stephen C Williams. Less is more: Relaxed yet composable security notions for key exchange. *International Journal of Information Security*, 12:267–297, 2013.
- [6] Boaz Barak, Ran Canetti, Jesper Buus Nielsen, and Rafael Pass. Universally composable protocols with relaxed set-up assumptions. In *45th Annual IEEE Symposium on Foundations of Computer Science*, pages 186–195. IEEE, 2004.
- [7] Ran Canetti, Hugo Krawczyk, and Jesper B Nielsen. Relaxing chosen-ciphertext security. In *Advances in Cryptology-CRYPTO 2003: 23rd Annual International Cryptology Conference, Santa Barbara, California, USA, August 17-21, 2003. Proceedings 23*, pages 565–582. Springer, 2003.
- [8] Peng Li and Steve Zdancewic. Downgrading policies and relaxed noninterference. In *Proceedings of the 32nd ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, pages 158–170, 2005.
- [9] U.S. Food and Drug Administration. Artificial intelligence and machine learning (ai/ml)-enabled medical devices. <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices>.
- [10] Xueyan Mei, Zelong Liu, Philip M. Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E. Link, Thomas Yang, Ying Wang, Hayit Greenspan, Timothy Deyer, Zahi A. Fayad, and Yang Yang. Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 0(ja):e210315, 0.
- [11] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.
- [12] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.
- [13] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.
- [14] Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.
- [15] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21:1–13, 2020.