Privacy in Practice: Private COVID-19 Detection in X-Ray Images

Lucas Lange[1][a], Maja Schneider[1][b], Peter Christen[2][c] and Erhard Rahm[1][d]
[1]*Leipzig University & ScaDS.AI Dresden/Leipzig, Leipzig, Germany*
[2]*The Australian National University, Canberra, Australia*
{lange, mschneider, rahm}@informatik.uni-leipzig.de, peter.christen@anu.edu.au

Keywords: Privacy-Preserving Machine Learning, Differential Privacy, Membership Inference Attack, Practical Privacy, COVID-19 Detection, Differentially-Private Stochastic Gradient Descent.

Abstract: Machine learning (ML) can help fight pandemics like COVID-19 by enabling rapid screening of large volumes of images. To perform data analysis while maintaining patient privacy, we create ML models that satisfy Differential Privacy (DP). Previous works exploring private COVID-19 models are in part based on small datasets, provide weaker or unclear privacy guarantees, and do not investigate practical privacy. We suggest improvements to address these open gaps. We account for inherent class imbalances and evaluate the utility-privacy trade-off more extensively and over stricter privacy budgets. Our evaluation is supported by empirically estimating practical privacy through black-box Membership Inference Attacks (MIAs). The introduced DP should help limit leakage threats posed by MIAs, and our practical analysis is the first to test this hypothesis on the COVID-19 classification task. Our results indicate that needed privacy levels might differ based on the task-dependent practical threat from MIAs. The results further suggest that with increasing DP guarantees, empirical privacy leakage only improves marginally, and DP therefore appears to have a limited impact on practical MIA defense. Our findings identify possibilities for better utility-privacy trade-offs, and we believe that empirical attack-specific privacy estimation can play a vital role in tuning for practical privacy.

1 INTRODUCTION

The COVID-19 pandemic pushed health systems worldwide to their limits, showing that rapid detection of infections is vital to prevent uncontrollable spreading of the virus. Detecting COVID-19 in patients can be achieved using a RT-PCR test[1]. Although they are more reliable in terms of sensitivity than rapid antigen tests, results can take hours to arrive, and even if displaying negative, the virus could have already left the throat and manifested itself in the lungs, rendering it undetectable for either test (Albert et al., 2021).

In hospitals, chest X-rays can mitigate these drawbacks by enabling a fast and reliable diagnosis. Figure 1 shows chest X-ray scans of healthy (top) and COVID-19 (bottom) patients in direct comparison. Even though patchy consolidations are recognizable in the COVID-19 scans, such X-rays remain challenging to interpret. Specialists, however, are able to identify the severity of a case early on and can take measures without waiting for lab results.

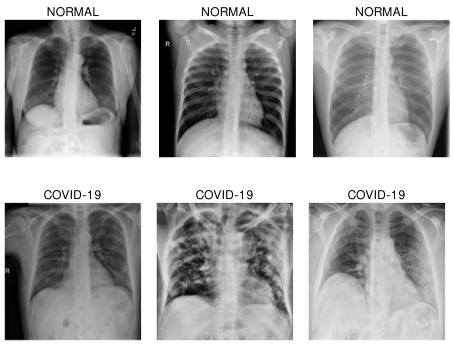


Figure 1: Chest X-ray images of different patients extracted from the COVID-19 Radiography Database (Chowdhury et al., 2020; Rahman et al., 2021). COVID-19 positive scans are characterized by patchy consolidations of the lungs.

Machine Learning (ML) techniques can effectively assist medical professionals in an initial screening by quickly classifying large numbers of images. However, the amount of data needed for training such classifiers poses problems due to clinical data privacy

[a] https://orcid.org/0000-0002-6745-0845
[b] https://orcid.org/0000-0001-5936-1415
[c] https://orcid.org/0000-0003-3435-2015
[d] https://orcid.org/0000-0002-2665-1114

[1]Reverse Transcription Polymerase Chain Reaction (RT-PCR) testing is broadly used for COVID-19 diagnosis.

Table 1: Existing solutions from related work next to our private model at $\varepsilon = 1$ (Müftüoğlu et al., 2020; Zhang et al., 2021; Ho et al., 2022). The methods refer to training private models. For differentiating the tasks, we assign the classes as COVID-19 (C), Normal (N), or Pneumonia (P). A performance comparison is difficult due to the different characteristics. Baseline shows the best non-private models. Our best private result is based on accuracy for comparison to related work. We further include two proposed additions for filling open gaps: F1-score and MIA.

Related Work	Method	Number of samples	Task	Accuracy in %		ε	F1?	MIA?
				Baseline	Private			
Müftüoğlu et al. (2020)	PATE on EfficientNet-B0	139 COVID-19 234 Normal	binary C | N	94.7	71.0	5.98	×	×
Zhang et al. (2021)	ResNet on images from DP-GAN trained in federated learning	350 COVID-19 2,000 Normal 1,250 Pneumonia	multi-class C | N | P	92.9	94.5	?	×	×
Ho et al. (2022)	DP-SGD in federated learning on custom CNN with spatial pyramid pooling	3,616 COVID-19 10,192 Normal 1,345 Pneumonia	multi-class C | N | P	95.3	68.7	39.4	×	×
Lange et al. (ours)	DP-SGD on ResNet18 with tanh activation and pre-training on Pneunomia	3,616 COVID-19 5,424 Normal	binary C | N	96.8	75.2	1	✓	✓

regulations, which present strict limitations to data sharing between hospitals. All sensitive patient information must be treated confidentially before, during, and after processing. (Balthazar, 2018)

To complicate matters, not only the dataset itself but also the models resulting from ML can compromise privacy. Published models are vulnerable to attacks, including leaking details about their training data (Shokri et al., 2017). Such leaks allow adversaries to potentially deduce sensitive medical facts about individuals in the dataset, for instance by exposing a patient's genetic markers (Homer et al., 2008).

In the case of COVID-19 detection, an attacker could be able to reveal if a person was infected, which would already violate privacy. While the specific risk of X-ray-based attacks might be low, such data should be handled with caution, especially since even with anonymization, results can still be linked to other information like related medications. Furthermore, we cannot rule out an attacker with internal access to images, e.g. doctors utilizing the model in a hospital.

Privacy-Preserving ML (PPML) is a collection of methods for creating trustworthy ML models, enabling, for example, the development of medical applications while maintaining patient privacy. In this work, we apply PPML that satisfies Differential Privacy (DP) (Dwork, 2008) in training a COVID-19 detection model, thus limiting attacks on the resulting classifier from incurring information leakage.

Our investigation is divided into three successive steps: (1) First, a non-private baseline is trained to detect COVID-19 versus normal (no findings) in chest X-rays. (2) The second step then focuses on experiments evaluating ML model architectures and parameters in private training, with the primary objective of finding a feasible utility-privacy trade-off. (3) Finally, model privacy is empirically assessed by attempting to identify training data through black-box Membership Inference Attacks (MIAs), examining to what extent these models leak private information.

Our contributions are:

- We fill open gaps from previous work (Müftüoğlu et al., 2020; Zhang et al., 2021; Ho et al., 2022), where Table 1 shows their characteristics in comparison to our approach. We address the class imbalances and analyze the utility-privacy trade-off more extensively by evaluating multiple and stricter privacy budgets. We further investigate practical privacy by empirically estimating privacy leakage through black-box MIAs. These gaps and our improvements are addressed throughout the following sections.
- We are the first to evaluate if DP helps narrow down MIAs on the COVID-19 detection task. We additionally re-examine this hypothesis on a common benchmarking dataset to reveal connections between the two datasets. Our results point towards identifying the benefits from DP in defending against MIAs as task-dependent and plateauing. We are able to gain better utility-privacy trade-offs at no practical cost. These results thus strengthen the belief that empirical privacy analysis can be a vital tool in supporting attack- and task-specific tuning for privacy.

The following Section 2 provides an overview of essential concepts. We then contextualize our work by examining the existing literature in Section 3, and we present our selected solutions to address open research gaps in Section 4. Section 5 lays out our experimental setup, with their results and discussion in Section 6. In closing, Section 7 provides conclusive thoughts and adds an outlook to possible future work.

2 BACKGROUND

This section establishes a basic understanding of the relevant concepts and algorithms used in this work.

2.1 Differential Privacy

DP offers a guarantee that the removal or addition of a single dataset record does not (substantially) affect the outcome of any analysis (Dwork, 2008). Thus, an attacker is incapable of differentiating from which of two neighboring datasets a given result originates and has to resolve to a random guess. DP's provided guarantee is measured by giving a theoretical upper bound of privacy loss, represented as the privacy budget ε. The metric is accompanied by the probability of privacy being broken by accidental information leakage, which is denoted as δ and depends on the dataset size.

Formally, an algorithm A training on a set S is called (ε,δ)-differentially-private, if for all datasets D and D' that differ by exactly one record:

$$Pr[A(D) \in S] \leq e^{\varepsilon} Pr[A(D') \in S] + \delta \qquad (1)$$

Meaningful privacy guarantees in ML should fulfill $\varepsilon \leq 1$ and $\delta \ll 1/n$, where n is the number of training samples (Nasr et al., 2021; Carlini et al., 2019). The notation $\varepsilon = \infty$ indicates that no DP criteria are met.

2.2 Differentially-Private Stochastic Gradient Descent

The Differentially-Private Stochastic Gradient Descent (DP-SGD) algorithm introduced by Abadi et al. (2016) takes widely used SGD and applies a gradient perturbation strategy. Gradient perturbation adds enough noise to the intermediate gradients to obfuscate the largest value, since that original sample inhibits the highest risk of exposure. To generally bound the possible influence of individual samples while training, DP-SGD clips gradient values to a predefined maximum Euclidean norm before adding noise. The noisy gradients are then used to update the parameters as usual. The total noise added through the algorithm is composed over all training iterations and determines the resulting privacy budget.

2.3 Membership Inference Attacks

In black-box MIAs, an attacker feeds data samples to a target model and thereby tries to figure out each sample's membership or absence in the model's training set based solely on the returned confidence values. This technique takes advantage of the differences in predictions made on data used for training versus unseen data, where the former is expected to output higher confidence values due to memorization (Carlini et al., 2019). As proposed by Shokri et al. (2017), such attacks can utilize multiple shadow models specifically mimicking a target model's predictions, to train an attack model able to elicit the desired membership information. Salem et al. (2019) relaxed the need for shadow models, by finding that simply using the original model's predictions on given samples can be sufficient to deduce their membership. By revealing the membership of an individual's record in the dataset, an adversary might in turn disclose sensitive information on them.

3 RELATED WORK

In the following, we first describe gaps left open by related work in Section 3.1. We then show mitigation strategies for MIAs and methods of practical privacy analysis in Sections 3.2 and 3.3, respectively.

3.1 Private COVID-19 X-Ray Detection

In Table 1, existing works on private COVID-19 detection from X-rays are summarized and compared to our approach. There are multiple factors that impede a fair comparison, which mainly lie in the differences in datasets, tasks, and privacy guarantees (ε). In this section, we show open gaps and then give elaborations in Section 4 on how we address them.

Datasets. A problem regarding (Müftüoğlu et al., 2020) is that their results are based on only a small dataset of 139 COVID-19 scans. The COVID-19 Radiography Database used by (Ho et al., 2022) and us, provides a better basis in terms of dataset size. However, the class imbalances result in a rather skewed data basis, which is left unaddressed but could influence MIA threat (Jayaraman et al., 2021). With the FedDPGAN approach, Zhang et al. (2021) try to enlarge and balance their small dataset using synthetic images, but the quality of the generated distribution is left unanswered. This is particularly problematic because GANs trained on imbalanced input data tend to produce data with similarly disparate impacts (Ganev et al., 2022). As a general problem with skewness, the mentioned works solely assess performance using accuracy, although this metric is known to undervalue false negatives for minority classes and could favor classifiers that are actually worse in detecting the COVID-19 minority class (Bekkar et al., 2013).

Privacy budgets. The used ε-values of 5.98 and 39.4 by Müftüoğlu et al. (2020) and Ho et al. (2022) respectively, are significantly weaker than the privacy

budget of $\varepsilon \leq 1$, which is commonly assumed to provide strong privacy (Nasr et al., 2021; Carlini et al., 2019). Furthermore, the results by Zhang et al. (2021) lack comparability, since they do not provide their privacy budget. Using their parameters and noise in a standard DP-SGD analysis results in $\varepsilon > 5 * 10^{13}$ for a client after 500 rounds[2] of federated training. Even with their most private setting they still accumulate $\varepsilon = 19.6$. Thus, no model adheres to $\varepsilon \leq 1$ and they instead only offer weaker or unclear guarantees.

Practical privacy. Regarding practical privacy, prior work does not include actual attack scenarios. It is therefore left open to what extent the provided models and ε-guarantees retain patient privacy against real adversaries. Such analysis helps in assessing the defense capabilities provided by the achieved privacy budgets and could reveal room for tuning them.

3.2 Repelling MIAs

Related work suggests multiple strategies for reducing MIA threats. Shokri et al. (2017) show that limiting the model outputs to only class labels instead of explicit confidence values can be an effective remedy. However, in medical tasks such as COVID-19 detection, where the use case is to help medical professionals in diagnosing a disease, the confidence value is an integral part that indicates how likely a patient is affected. Shokri et al. (2017) also find that model architecture can contribute to MIA defense and Salem et al. (2019) demonstrate that even the training process can hinder MIAs through e.g. model stacking.

DP should limit and oppose the success of MIAs by design, with Jayaraman and Evans (2019) supplying the corresponding reasoning: "[DP], by definition, aims to obfuscate the presence or absence of a record in the data set. On the other hand, [MIAs] aim to identify the presence or absence of a record [...]." Rahman et al. (2018) test this hypothesis by evaluating MIAs on different privacy levels. They find their model's MIA resistance to gradually increase when lowering the allowed privacy budget and explain it with less overfitting when adding more noise. Yeom et al. (2018) prove that overfitting in ML models is sufficient to enable MIAs, but at the same time show that overfitting is not a necessary criterion, and stable models can still be vulnerable.

3.3 Practical Privacy Analysis

Multiple works examined the possibilities of estimating the practical privacy for ML models by performing an empirical study through attacks, e.g. MIAs. Jagielski et al. (2020) and Nasr et al. (2021) conclude that the assumed theoretical upper bound privacy loss for DP, given in the privacy budget ε, gives a tight worst-case analysis on attack proneness and thereby limits MIA success. However, in many cases actual attacks extract significantly less information than assumed by the theoretical bound, which is also supported by Malek et al. (2021) and Jayaraman and Evans (2019). This discrepancy could possibly enable better utility-privacy trade-offs, but Jayaraman and Evans (2019) warn that privacy always comes at a cost and reducing privacy could ultimately promote information leakage. Malek et al. (2021) propose that a realistic lower bound on the amount of revealed information by a model can be determined by "[considering] the most powerful attacker pursuing the least challenging goal" and that in the case of standard DP, such would be an attacker powerful enough to successfully perform membership inference.

4 METHODS

As seen in Section 3.1 and Table 1, related work on COVID-19 detection lacks comparability and leaves open research gaps. We therefore do not solely focus on enhancing the performance of former solutions but rather suggest improvements by filling existing gaps, ultimately proposing the following improvements.

Datasets. Since the dataset used by us and Ho et al. (2022) provides a good amount of COVID-19 samples, we instead aim for better handling of the problems arising from the skewed nature of the class distribution. In a first effort, we employ random undersampling and class weights to elevate the underrepresented COVID-19 class, both in database construction and in training, respectively. Furthermore, since accuracy is not representative in cases of skewed data, we improve the evaluation by using the more balanced F1-score metric (Bekkar et al., 2013).

Privacy budgets. We investigate the utility-privacy trade-off by evaluating multiple and stricter privacy budgets of $\varepsilon = [\infty, 10, 1, 0.1]$. To find the best private model and extend the pool of evaluated methods, we propose untested architectural experiments relevant to private DP-SGD training in Section 5.4.

Practical privacy. As seen in the works discussed in Section 3.3, we investigate the practical implications of DP regarding the defense against black-box MIAs by undertaking an empirical analysis through actual attacks, and therefore give a more realistic lower bound to the resulting privacy leakage (Jagielski et al., 2020; Malek et al., 2021). We thereby

[2] They do not state their exact number of rounds but their graphs show 500 rounds.

Table 2: Summary of the experiment parameters. Each combination from left to right constitutes a possible setup (resulting in $2*2*2*3*4 = 96$ setups).

Dataset	Architecture	Activation	Pre-training	ε
COVID-19	ResNet18	ReLU	None/Standard	∞
MNIST	ResNet50	tanh	ImageNet	10
			Pneumonia	1
				0.1

provide the first attack results in the field of private COVID-19 detection and evaluate possible room for tuning the utility-privacy trade-off. An additional evaluation regarding the privacy leakage of our models on the MNIST database enables us to formulate takeways regarding similarities and disproportions regarding the attack-specific privacy on both datasets. Evaluating another dataset is a first step towards generalization and MNIST is particularly interesting because related works (Rahman et al., 2018; Nasr et al., 2021) previously investigated the connection between DP and MIA on this task.

5 EXPERIMENTAL SETUP

In this section we provide details on the setups used in our experiments, which are summarized in Table 2. Reference code is available from our repository[3].

5.1 Environment

We use Python with the Keras, Tensorflow, and Tensorflow Privacy libraries. To enable reproducible results any random seeds are set to a fixed value of 42. Hardware-wise our machines are equipped with 64GB RAM and an NVIDIA Tesla V100 GPU.

5.2 Datasets

For a comprehensible dataset creation, we provide details on the different public datasets we used.

- The *COVID-19 Radiography Database*[4] (Chowdhury et al., 2020; Rahman et al., 2021) is the most comprehensive collection of COVID-19 chest X-ray images, stemming from different databases around the web. In total, this image collection offers chest X-rays of 3,616 COVID-19 positive, 10,192 Normal, and 1,345 Pneumonia cases. For our binary task, we omit the pneumonia samples and employ undersampling to directly reduce class imbalances. Dataset construction takes all COVID-19 scans but only 1.5× the amount for Normal (5,424), resulting in 9,040 images total. When testing hyperparameters, this ratio showed to elevate performance (F1) and reduce privacy risk due to less overfitting (Rahman et al., 2018).
- The *Chest X-Ray Images (Pneumonia)*[5] (Kermany et al., 2018) offers X-ray images divided into two classes with 1,583 Normal and 4,273 Pneumonia samples. Here, we again apply undersampling to achieve similar class ratios and take all Normal scans but just 1.5× the amount for Pneumonia (2,374). This pneumonia dataset is also part of the COVID-19 Radiography Database, constituting 13% (1,341) of its Normal class images. To fix this issue and enable its use as a public dataset for our private transfer learning approach without compromising privacy, we exclude duplicates when sampling images for the COVID-19 task.
- The *ImageNet* (Deng et al., 2009) is a vast collection counting 14 million images and covering 20,000 categories from general (mammal) to specific (husky). Non-private models benefit from using this massive dataset for pre-training, introducing many differentiating concepts to a neural network before training on the target data.
- With their *MNIST Database* (LeCun et al., 1998), LeCun et al. offer a large image collection of handwritten digits. The database provides 60,000 images for training and 10,000 for testing. Even though MNIST does not contain COVID-19 related images, it is a commonly used benchmark in image classification and PPML, making it a perfect candidate for comparing results.

5.3 Pre-processing

To build our final splits for model training, we employ necessary sampling and pre-processing steps. Both X-ray datasets, for COVID-19 and pneumonia, use a train-validation-test split of 80% training, 5% validation, and 15% test set. All datasets are handled with three color channels. We therefore convert the MNIST grey-scale images into the RGB space, as this is vital to allow the models pre-trained on color images to still work with the input data. X-ray images are downscaled to 224x224 pixels, while MNIST images keep their size of 28x28. Both however undergoes an image normalization on the factor of x=1/255.

To combat overfitting, training sets are shuffled and training images from the X-ray datasets are subjected to data augmentation (Shorten and Khoshgof-

[3]https://github.com/luckyos-code/mia-covid

[4]https://www.kaggle.com/tawsifurrahman/covid19-radiography-database

[5]https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

taar, 2019). To further reduce the imbalance regarding Sickness (COVID-19, Pneumonia) and Normal class frequencies, we apply class weights during training to help artificially balance each sample's impact. This process yields class weights of 0.83 for the Normal and 1.25 for the Sickness classes.

5.4 Architectural Experiments

In the following, we describe our different architectural choices we used for experiments.

5.4.1 Model Size in DP-SGD

Non-private and private classification perform differently depending on the underlying model architecture (Papernot et al., 2021). Performance is greatly dependent on model size with non-private training typically benefiting from using bigger models. Using DP-SGD, the same models can suffer from accuracy loss when increasing in size. Taking these findings into account in our experiments, we used two differently-sized architectures from the model family of Residual Networks (He et al., 2016), or short ResNets. For one, the ResNet50 with 50 layers, as well as the ResNet18, which is smaller at 18 layers.

5.4.2 tanh Activation

A disruptive discovery in DP-SGD research was made by Papernot et al. (2021). In their work, they determined that replacing the de facto standard ReLU activation function with the tanh function in model layers improves performance in DP-SGD. To achieve this boost, they utilize the fact that the tanh activation generally results in smaller gradients than the ReLU function, which in turn reduces the information loss from gradient clipping.

5.4.3 Pre-training

A commonly applied strategy to improve performance for non-private classification relies on pre-training using the extensive ImageNet collection. As another method, Abadi et al. (2016) state that DP-SGD models can further profit from pre-training in a domain closely related to the target task. While ImageNet resembles a general choice for image-based tasks, pre-training for pneumonia detection is closer to our COVID-19 task due to the similarity in symptoms (Speranskaya, 2020; Lange, 2022).

The pre-training on the Pneumonia dataset is performed using the same settings as on the COVID-19 set, while the ImageNet variants are provided by a library for Keras models[6]. For our tanh variants we take the take the pre-trained ReLU models and change the activation function in each trainable layer before training on our target datasets.

5.5 Privacy Experiments

In this section, we elaborate on the used settings and hyperparameters for evaluating privacy.

5.5.1 Private Training Settings

For our non-private baseline, we employ Adam (Kingma and Ba, 2015) optimization with batch sizes of 32 and train for 20 epochs using a learning rate of $\alpha = 1e{-}3$, which decays down to a minimum of $\alpha = 1e{-}6$ on plateaus. Afterwards, we apply DP-SGD (or here DP-Adam) training with a clipping norm of 1.0 and the appropriate noise to all private models to achieve a candidate for each ε-guarantee. DP-SGD training for COVID-19 uses ResNet50 and ResNet18 variants with batch sizes of 8 and 16, instead of 32 respectively. We aim at privacy budgets of $\varepsilon \leq 1$, since such values present strong privacy guarantees (Nasr et al., 2021; Carlini et al., 2019). We also evaluate budgets neighboring this setting by an order of magnitude, to gain further insights into the performance and estimated privacy on different DP levels. Due to the dataset size, the DP analysis uses $\delta = 1e{-}4$ for COVID-19 ($n = 9,040$) and $\delta = 1e{-}5$ for MNIST ($n = 60,000$).

5.5.2 MIA Settings

For selecting the most potent MIA each run, we try four different attack types based on logistic regression, multi-layered perceptron, k-nearest neighbors, and threshold. These attacks are an implementation of the single shadow model black-box attack proposed by Salem et al. (2019), that directly relies on target model predictions instead of training several shadow models. Given a target model, MIAs utilize two types of data: (1) the original training data to be inferred and (2) unseen but similar data to differentiate non-training data. In our case, we want to fully empower the attacker for estimating the practical worst-case in an optimal black-box setting (Malek et al., 2021). We satisfy this condition by giving access to the full training and test sets with their corresponding labels, thus, handing the attacker the largest input regarding (1) and the most similar input regarding (2).

[6] https://github.com/qubvel/classification_models

5.5.3 Measuring Privacy Leakage

Like Jayaraman and Evans (2019), our used metric for measuring privacy leakage through MIAs is the attacker's membership advantage as introduced by Yeom et al. (2018). The adversarial game is based on an attacker's capabilities in differentiating the membership of a sample that is chosen uniformly at random to originate from the training set or not. The resulting difference in True Positive Rate (TPR) and False Positive Rate (FPR) is given as the attacker's membership advantage: $\mathsf{Adv}^{\mathsf{M}} = TPR - FPR$.

Yeom et al. (2018) show that if a learning algorithm satisfies ε-DP, then the adversary's membership advantage is bounded by $\mathsf{Adv}^{\mathsf{M}} \leq e^{\varepsilon} - 1$ in their attack scenario. Transferring the theorem to (ε,δ)-DP given by Equation (1), the upper bound can be derived as:

$$\mathsf{Adv}^{\mathsf{M}} \leq e^{\varepsilon} - 1 + \delta \tag{2}$$

Because the theoretical assumption relies on Gaussian distributed training errors and a balanced prior data distribution probability, it might not provide reliable bounds given our differing practical scenario.

Since individual MIA results are subject to variability, they need to be experimentally stabilized. Like Malek et al. (2021), we achieve this by running 100 entire MIAs and calculating the corresponding 95% Confidence Interval (CI) for the obtained results.

6 DISCUSSION

We now revisit the open gaps from the related work discussed in Section 3 and review the outcomes of our proposed solutions from Section 4. We refer to Tables 3 and 4 for our evaluation results and to Table 1 for an organized showcase of results from related work. We evaluate our proposed improvements:

Datasets. We achieve a more balanced data basis than before by utilizing undersampling and class weights. To evaluate on the still skewed data, we add the F1-score metric. The advantage to accuracy is visible in the COVID-19 results, where both metrics differ regularly and F1 thus reveals models that perform better on the minority class COVID-19. That F1 accounts for the underlying class distribution is further demonstrated on the more balanced MNIST dataset, where accuracy and F1-score are almost identical.

Privacy budgets. In contrast to related work, we are able to achieve working COVID-19 detection models, while adhering to strong privacy budgets of $\varepsilon \leq 1$. By additionally evaluating different architectures over multiple privacy levels, we deduce favorable architectural decisions for keeping good utility-privacy trade-offs in DP-SGD. Our findings for training private models are summarized in Section 6.1.

Practical privacy. By including an empirical study on practical privacy though MIAs, we gain insights into the relationship between DP and privacy leakage. In Section 6.2 we derive the implications stemming from our empirical analysis. The results allow us to improve the utility-privacy trade-off while keeping the same practical privacy.

6.1 Building Better DP-SGD Models

With our experiments, we transfer the results from Papernot et al. (2021) to deeper and pre-trained networks and are able to confirm the tanh advantage over ReLU in low ε DP-SGD training. Our private models with strong ε-guarantees of $\varepsilon = 1$ and $\varepsilon = 0.1$ rely on this change, while the non-private and less private models still prefer the ReLU activation function.

A commonality between our best performers is that they were subjected to pre-training. While all best non-private models are pre-trained on ImageNet, this trend only continues in all private models on the MNIST database. The same ImageNet-based models underperform on the COVID-19 task, when looking at settings of $\varepsilon = 1$ and $\varepsilon = 0.1$, which might be related to the different contents in both tasks. On the COVID-19 task, we introduce task-specific pre-training on pneumonia images, that leads to superior performance in our most private settings.

We could not fully confirm that larger models perform worse in DP-SGD (Papernot et al., 2021). The ResNet50 especially wins at the most private setting of $\varepsilon = 0.1$. Model size, however, seems to play a role, when examining the earlier failure of the private ReLU models in the bigger ResNet50.

In summary, our results support the existing belief that model architectures should be specifically adjusted for private DP-SGD training, where established standards from non-private training do not necessarily provide the same advantages (Papernot et al., 2021; Abadi et al., 2016). Examples are the switch from ReLU to tanh activation and the superiority of Pneumonia pre-training to ImageNet pre-training in the private COVID-19 models.

6.2 Insights Regarding Practical DP

We now refer to Figure 2 that plots our estimated privacy leakage from MIAs at the different ε-budgets.

In both Figures 2a and 2b, we include the (ε,δ)-DP bound on $\mathsf{Adv}^{\mathsf{M}}$ from Equation (2), which is based on Yeom et al. (2018). The bound already surpasses our plotted maximum of 0.5 $\mathsf{Adv}^{\mathsf{M}}$ long before $\varepsilon = 1$,

Table 3: Experimental results on the COVID-19 dataset. The Standard, ImageNet and Pneunomia models rely on the ReLU activation function, which is then changed to tanh in the respective counterparts. Model variants are evaluated across multiple DP budgets ε, where $\varepsilon = \infty$ translates to non-private training. They are matched by accuracy and F1-score in %, as well as empirical privacy leakage from MIAs, measured by the membership advantage ($\mathsf{Adv}^{\mathsf{M}}$) and given as a 95% CI over 100 attacks. If training resulted in an F1-score of 0.0, no feasible model was derived, making accuracy and attacks obsolete (NA).

	$\varepsilon = \infty$			$\varepsilon = 10$			$\varepsilon = 1$			$\varepsilon = 0.1$		
Variant	%-Acc.	%-F1	$\mathsf{Adv}^{\mathsf{M}}$	%-Acc.	%-F1	$\mathsf{Adv}^{\mathsf{M}}$	%-Acc.	%-F1	$\mathsf{Adv}^{\mathsf{M}}$	%-Acc.	%-F1	$\mathsf{Adv}^{\mathsf{M}}$
ResNet18												
Standard	91.4	89.5	0.22–0.24	71.2	57.8	0.22–0.24	NA	0.0	NA	NA	0.0	NA
ImageNet	**96.8**	**95.9**	0.25–0.27	**85.5**	**79.4**	0.25–0.27	NA	0.0	NA	NA	0.0	NA
Pneumonia	92.2	89.8	0.22–0.24	71.5	57.2	0.23–0.25	70.5	54.3	0.22–0.24	71.3	61.4	0.22–0.23
tanh-Standard	85.1	82.8	**0.21–0.23**	71.8	67.9	0.21–0.23	71.5	62.5	0.20–0.22	68.0	63.0	0.20–0.22
tanh-ImageNet	91.4	89.8	0.22–0.24	57.5	65.2	**0.19–0.21**	44.5	58.9	0.20–0.22	50.8	61.0	**0.19–0.21**
tanh-Pneumonia	79.9	78.6	**0.21–0.23**	73.9	73.1	0.21–0.24	**75.2**	70.5	0.22–0.24	72.9	65.8	0.21–0.22
ResNet50												
Standard	91.6	89.3	0.25–0.27	NA	0.0	NA	NA	0.0	NA	NA	0.0	NA
ImageNet	95.6	94.4	0.25–0.26	NA	0.0	NA	NA	0.0	NA	NA	0.0	NA
Pneumonia	91.4	89.6	0.24–0.26	NA	0.0	NA	NA	0.0	NA	NA	0.0	NA
tanh-Standard	78.8	78.0	**0.21–0.23**	72.3	63.4	0.22–0.23	70.4	62.9	**0.19–0.21**	68.6	62.0	**0.19–0.21**
tanh-ImageNet	88.8	84.9	0.23–0.25	47.9	60.1	**0.19–0.21**	46.0	59.3	**0.19–0.21**	50.8	60.7	**0.19–0.21**
tanh-Pneumonia	81.3	80.1	**0.22–0.23**	72.0	72.7	0.21–0.23	72.0	**72.5**	0.21–0.23	**73.0**	**69.4**	0.21–0.23

Table 4: Experimental results on the MNIST database. See Table 3 caption for details. F1-score is given as the macro average over the 10 classes. Pneumonia pre-trained models are omitted from the evaluation, since the tasks are not closely related.

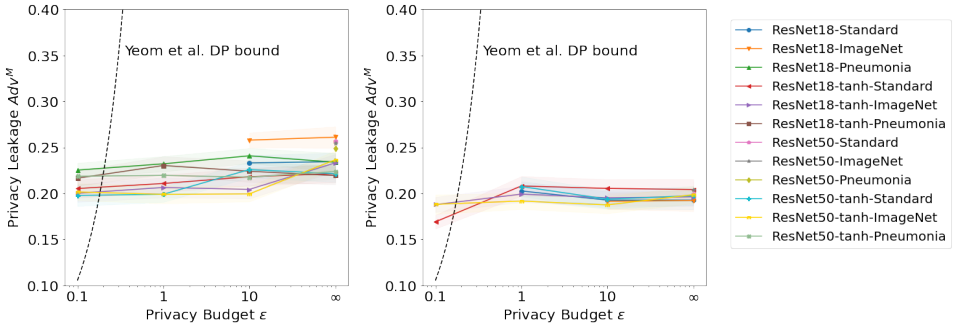
	$\varepsilon = \infty$			$\varepsilon = 10$			$\varepsilon = 1$			$\varepsilon = 0.1$		
Variant	%-Acc.	%-F1	$\mathsf{Adv}^{\mathsf{M}}$	%-Acc.	%-F1	$\mathsf{Adv}^{\mathsf{M}}$	%-Acc.	%-F1	$\mathsf{Adv}^{\mathsf{M}}$	%-Acc.	%-F1	$\mathsf{Adv}^{\mathsf{M}}$
ResNet18												
Standard	**99.5**	**99.5**	**0.18–0.20**	95.9	95.9	0.18–0.20	90.2	90.1	0.19–0.21	24.3	19.9	0.13–0.15
ImageNet	**99.5**	**99.5**	**0.18–0.20**	95.2	95.1	0.18–0.20	38.0	35.2	**0.13–0.14**	14.7	12.7	0.15–0.17
tanh-Standard	99.2	99.2	0.19–0.22	95.1	95.0	0.20–0.22	92.4	92.3	0.20–0.22	72.6	71.5	0.16–0.18
tanh-ImageNet	99.0	99.0	0.19–0.21	**97.8**	**97.8**	0.19–0.20	**96.7**	**96.7**	0.19–0.21	90.9	90.8	0.18–0.20
ResNet50												
Standard	**99.5**	**99.5**	0.19–0.21	16.0	14.5	**0.14–0.15**	12.7	11.6	0.14–0.16	10.9	9.5	0.15–0.17
ImageNet	98.5	98.5	0.19–0.21	11.2	10.1	0.14–0.16	10.0	8.9	0.15–0.17	9.5	8.0	0.14–0.16
tanh-Standard	99.3	99.3	0.19–0.21	93.3	93.3	0.18–0.20	85.0	84.7	0.20–0.22	27.6	25.4	**0.13–0.14**
tanh-ImageNet	99.0	99.0	0.19–0.21	97.7	97.7	0.18–0.20	96.6	96.6	0.18–0.20	**93.3**	**93.2**	0.18–0.20

which shows the large discrepancy between the theoretically assumed worst-case and practice. Simultaneously, no model actually trained for $\varepsilon = 0.1$ is able to conform to the calculated bound. Such inconsistencies can also be found in related work (Yeom et al., 2018; Jayaraman and Evans, 2019). As an explanation, Yeom et al. (2018) unveil that, the training set error distributions are not exactly Gaussian in practice, sometimes leading to better attack performance than predicted. Even though COVID-19 and MNIST have rather opposing priors, where the former's classes are skewed and the latter's roughly balanced, we see the same inconsistencies in both evaluations. Thus, the given theoretical bound does not seem reliable for deriving a limit on the real world threat in our case.

For both COVID-19 and MNIST, the leakage almost describes a flat line with just negligible changes over all privacy settings. We spot a few outliers[7], that see a bigger drop in leakage risk, which however, is mainly attributed to their gravely lowered performance (>20% F1 loss) and accordingly reduced memorization (Rahman et al., 2018). Even the non-private models exhibit almost the same leakage as the private models and thus, including DP-guarantees does not imply the expected improvement to practical MIA proneness. The plateau in privacy leakage can enable the use of lesser DP-guarantees, while still providing the same practical privacy The MNIST models show to generally leak slightly less than on COVID-19, leading to stronger privacy needs for COVID-19. The existing difference in MIA risk between COVID-19 and MNIST suggests, that privacy estimation can be an important tool for assessing task- and data-dependent threats from attacks. Thus, such estimates can in turn support tuning trade-offs according to task-specific privacy needs.

The findings suggest room for utilizing weaker DP-guarantees on both tasks when defending against our MIA-specific setting. Practical privacy is already

[7]On COVID-19 the outliers are both tanh-ImageNet models, which reduce their leakage from non-private to $\varepsilon = 10$, and the ResNet50 tanh-Standard doing the same from $\varepsilon = 10$ to $\varepsilon = 1$. There is also one outlier on MNIST, where the Resnet18-tanh-Standard improves privacy at $\varepsilon = 0.1$.



(a) COVID-19 (b) MNIST

Figure 2: Empirical privacy leakage results from MIAs are given as our 95% CI membership advantage (Adv^M) and plotted across the different privacy budgets. Model variants can be distinguished with the legend. We exclude data points with <50% F1 because low performance disproportionately reduces leakage. A dotted line shows the DP bound from Yeom et al. (2018).

strong in our less private and even non-private models. We are thus able to improve the utility-privacy trade-off on both datasets at no practical privacy cost. We could introduce even stronger guarantees to possibly further improve MIA defense, however, this would lead to an even bigger utility loss and in turn result in impractical performance.

We want to emphasize that there is still a need for strong theoretical privacy guarantees (Nasr et al., 2021). As stated in Section 3.2, ε-guarantees from DP limit the maximum amount of possible information leakage. In actual attacks, however, the theoretical ceiling might differ notably from the practical threat as shown in this and other works presented in Section 3.3. Thus, we would rather choose a COVID-19 model at $\varepsilon = 10$ than at $\varepsilon = \infty$, even though both exhibit almost the same practical privacy levels. The model at $\varepsilon = 10$ performs better than the one at $\varepsilon = 1$ and, in contrast to $\varepsilon = \infty$, still provides a provable DP guarantee to limit future adversaries.

7 CONCLUSION

Within this piece of work, we close several open gaps in the field of private COVID-19 detection from X-ray images. In comparison to related work on the topic, we improve data handling regarding imbalances, deliver a more robust privacy evaluation, and are the first to investigate the implications concerning practical privacy (Müftüoğlu et al., 2020; Zhang et al., 2021; Ho et al., 2022).

We introduce a selection of yet untested architectural ML model choices to the COVID-19 task. Through our evaluation, we are able to compare the setups in a common environment. Since well-known practices from non-private training are not always transferable to DP-SGD training, it is important to gather a wide range of results for finding the best models. We are therefore making a noticeable contribution by exploring a range of different architectures on the COVID-19 and MNIST tasks.

Our practical privacy analysis reveals that assessing attack-specific threats from black-box MIAs in a practical scenario helps finding appropriate privacy attributes and can thus improve the utility-privacy trade-off at no practical cost. On both the COVID-19 and MNIST datasets, we found just minor improvements from the provided theoretical DP-guarantee regarding practical defense against our MIAs. Instead, our tested models almost showed the same strong repelling properties across all privacy levels—even for non-private models. By confirming this plateau for both datasets, we are able to reduce the required DP guarantees for both tasks without sacrificing attack-specific practical privacy. Our attacks are slightly more successful on the COVID-19 task, showing that it needs stricter privacy than MNIST and that practical privacy analysis is important for identifying the task-specific initial MIA threat.

We still advocate the use of DP and would not recommend to risk publishing non-private COVID-19 detection models. Instead, if justified by a practical privacy analysis, the ε-guarantee can be tuned to a more favorable utility-privacy trade-off that through the inclusion of a reasonable DP-guarantee still limits the worst-case privacy leakage from future attacks.

As a brief outlook into possible future work, it would be beneficial to extend our evaluation by applying practical privacy analysis to more datasets, especially with different underlying tasks. Another venture could be to derive best practices and ultimately a taxonomy regarding advantageous architectural decisions when training DP-SGD models.

ACKNOWLEDGMENTS

We thank the reviewers for their helpful feedback and our colleagues for insights on earlier drafts. The authors acknowledge the financial support by the Federal Ministry of Education and Research of Germany and by the Sächsische Staatsministerium für Wissenschaft Kultur und Tourismus in the program Center of Excellence for AI-research "Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig", project identification: ScaDS.AI.

REFERENCES

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *ACM SIGSAC*.

Albert, E., Torres, I., Bueno, F., Huntley, D., Molla, E., F.-F., M., Martínez, M., Poujois, S., Forqué, L., Valdivia, A., S. A., C., Ferrer, J., Colomina, J., and Navarro, D. (2021). Field evaluation of a rapid antigen test (panbio™ covid-19 ag rapid test device) for covid-19 diagnosis in primary healthcare centres. *CMI*, 27(3).

Balthazar, T. (2018). Sharing health-data between hospitals and other care-providers: Towards legal clarity about what can be communicated to whom. *NTvG*, 74.

Bekkar, M., Djemaa, H. K., and Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. *Rev. inf. eng. appl.*, 3(10).

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. (2019). The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security Symposium (USENIX Security 19)*.

Chowdhury, M. E., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M. A., Mahbub, Z. B., Islam, K. R., Khan, M. S., and Iqbal, A. (2020). Can ai help in screening viral and covid-19 pneumonia? *IEEE Access*, 8.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *CVPR*. IEEE.

Dwork, C. (2008). Differential privacy: A survey of results. In *TAMC*, pages 1–19. Springer.

Ganev, G., Oprisanu, B., and De Cristofaro, E. (2022). Robin hood and matthew effects: Differential privacy has disparate impact on synthetic data. In *ICML*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*. IEEE.

Ho, T., Tran, K., and Huang, Y. (2022). Fedsgdcovid: Federated SGD COVID-19 detection under local differential privacy using chest x-ray images and symptom information. *Sensors*, 22(10):3728.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F., and Craig, D. W. (2008). Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. *PLOS Genetics*, 4(8):e1000167.

Jagielski, M., Ullman, J., and Oprea, A. (2020). Auditing differentially private machine learning: How private is private sgd? *NeurIPS*, 33.

Jayaraman, B. and Evans, D. (2019). Evaluating differentially private machine learning in practice. In *28th USENIX Security Symposium (USENIX Security 19)*.

Jayaraman, B., Wang, L., Knipmeyer, K., Gu, Q., and Evans, D. (2021). Revisiting Membership Inference Under Realistic Assumptions. *PETS*, 2021(2).

Kermany, D., Goldbaum, M., Cai, W., Valentim, C., Liang, H., Baxter, S., McKeown, A., Yang, G., Wu, X., and Yan, F. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *ICLR*.

Lange, L. (2022). Privacy-Preserving Detection of COVID-19 in X-Ray Images. Master's thesis, Leipzig Univ.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Malek, M., Mironov, I., Prasad, K., Shilov, I., and Tramer, F. (2021). Antipodes of label differential privacy: Pate and alibi. *NeurIPS*, 34.

Müftüoğlu, Z., Kizrak, M. A., and Yildlnm, T. (2020). Differential privacy practice on diagnosis of covid-19 radiology imaging using efficientnet. In *INISTA*. IEEE.

Nasr, M., Song, S., Thakurta, A., Papernot, N., and Carlini, N. (2021). Adversary instantiation: Lower bounds for differentially private machine learning. In *S&P*. IEEE.

Papernot, N., Thakurta, A., Song, S., Chien, S., and Erlingsson, Ú. (2021). Tempered sigmoid activations for deep learning with differential privacy. In *AAAI*.

Rahman, M. A., Rahman, T., Laganière, R., Mohammed, N., and Wang, Y. (2018). Membership inference attack against differentially private deep learning model. *Transactions on Data Privacy*, 11(1):61–79.

Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Kashem, S. B. A., Islam, M. T., and Al Maadeed, S. (2021). Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Comput. Biol. Med.*

Salem, A., Zhang, Y., Humbert, M., Fritz, M., and Backes, M. (2019). Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *NDSS*. Internet Society.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *S&P*. IEEE.

Shorten, C. and Khoshgoftaar, T. (2019). A survey on image data augmentation for deep learning. *J. Big Data*, 6(1).

Speranskaya, A. (2020). Radiological signs of a new coronavirus infection covid-19. *Dia. rad. and rad.*, 11(1).

Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. (2018). Privacy risk in machine learning: Analyzing the connection to overfitting. In *CSF*. IEEE.

Zhang, L., Shen, B., Barnawi, A., Xi, S., Kumar, N., and Wu, Y. (2021). Feddpgan: Federated differentially private generative adversarial networks framework for the detection of covid-19 pneumonia. *Inf. Syst. Front.*