Dmitrii Usynin, Daniel Rueckert, Jonathan Passerat-Palmbach †, and Georgios Kaissis †*

Zen and the art of model adaptation: Low-utility-cost attack mitigations in collaborative machine learning

Abstract: In this study, we aim to bridge the gap between the theoretical understanding of attacks against collaborative machine learning workflows and their practical ramifications by considering the effects of model architecture, learning setting and hyperparameters on the resilience against attacks. We refer to such mitigations as *model adaptation*. Through extensive experimentation on both, benchmark and real-life datasets, we establish a more practical threat model for collaborative learning scenarios. In particular, we evaluate the impact of model adaptation by implementing a range of attacks belonging to the broader categories of model inversion and membership inference. Our experiments yield two noteworthy outcomes: they demonstrate the difficulty of actually conducting successful attacks under realistic settings when model adaptation is employed and they highlight the challenge inherent in successfully combining model adaptation and formal privacy-preserving techniques to retain the optimal balance between model utility and attack resilience.

Keywords: privacy, computer vision, federated learning, membership inference, model inversion

DOI 10.2478/popets-2022-0014
Received 2021-05-31; revised 2021-09-15; accepted 2021-09-16.

Dmitrii Usynin: Department of Computing, Imperial College London; Department of Diagnostic and Interventional Radiology, Technical University of Munich, E-mail: du216@ic.ac.uk
Daniel Rueckert: Institute for Artificial Intelligence in Medicine, Technical University of Munich; Department of Computing, Imperial College London; E-mail: d.rueckert@imperial.ac.uk
Jonathan Passerat-Palmbach †: (equal contribution) Department of Computing, Imperial College London; ConsenSys Health, New York, NY, USA, E-mail: j.passerat-palmbach@imperial.ac.uk
***Corresponding Author: Georgios Kaissis †:** (equal contribution) Institute for Artificial Intelligence in Medicine, Technical University of Munich; Department of Computing, Imperial College London, Germany E-mail: g.kaissis@tum.de

1 Introduction

Collaborative learning methods such as federated learning (FL) have become increasingly popular both across industry consortia [1] and larger public deployments [2]. A growing body of work describes attacks against this kind of learning networks, which range from deliberately destroying the utility of the learning process [3] to attempting to extract information from the model itself [4–6], thus compromising the network participants' privacy.

Privacy-preserving (PP) methods have been successfully leveraged in the context of collaborative learning to strengthen the training protocol and guard against attacks from adversarial actors. However, implementing PP techniques that protect *input privacy*, such as Secure Multi-Party Computation (SMPC) and Homomorphic Encryption (HE), adds substantial computational or communication overhead to the process as well as making the setting vulnerable to utility-oriented attacks such as backdoor insertion [7]. Additional computational requirements can be problematic when these clients are edge devices with limited resources. Methods like Differential Privacy (DP) are concerned with *output privacy* and are considered to be less computationally demanding, however they typically drastically reduce the accuracy of the jointly trained model, and as a consequence, limit the benefits of the collaborative learning procedure [8–11]. Moreover, even formal PP methods don't necessarily provide meaningful guarantees of privacy under combined attacks [12], as the adversary can often avail themselves of multiple methods of achieving privacy disclosure [5, 13, 14], while defense mechanisms can only effectively provide protection against only one of these methods at a time. Consequently, we identify a need for mitigation strategies that:

1. Do not degrade the performance of the joint model;
2. Are effective against multiple modes of attack and;
3. Can be deployed in *any* learning context.

In this work, we focus on a broad category of such techniques, which we summarise under the term *model adap-*

tation. We define model adaptation as the selection of specific attributes of the learning context by members of the federation with the aim to maximise robustness against *privacy-focused* attacks. Such attacks are designed to disclose information that the federation does not consent to sharing, whether intentionally or not. Model adaptation encompasses a wide range of methods to empirically achieve improved attack resilience. Some of them can be chosen arbitrarily and even dynamically during the learning process. For example, the training batch size or the selected subset of clients participating in each round belongs to this category. We also consider parameters under the definition which are relevant to the outcome of the learning task, however may not be chosen dynamically, such as the total number of participating clients or the model architecture, which may be task-specific. Recent works [6, 15–18] have highlighted that model adaptations can represent an additional measure of protection in *private-by-design* machine learning workflows despite their lack of formal guarantees. However, model adaptation comes with no *additional* infrastructural or software requirement and can thus serve as a low-cost, low probability-of-failure mechanism, from which *any* learning context can benefit. We position this study as a recommendation to the federation that should be taken into consideration before the learning protocol commences in order to reduce the amount of information unintentionally shared.

1.1 Contributions

In this work, we aim to systematically illuminate the extant techniques for model adaptation through large-scale empirical experimentation encompassing a variety of datasets, and have selected a broad subset that covers a large area of mitigations across multiple prior works [5, 15, 19]. Although we concentrate on collaborative image classification tasks, many of our findings can be applied to other contexts such as semantic segmentation tasks [20]. Our main contributions can be summarised as follows:

1. We identify the feasible attack scenarios in the domain of collaborative machine learning alongside their pertinent threat models (Section 3.2).
2. We perform detailed ablation studies to illuminate which *specific* characteristics of collaborative learning setups facilitate or inhibit privacy attacks to deduce whether it is possible to defend against them by modifying the model or the setup itself (Section 5). Through these experiments, we aim to identify a concrete set of model adaptations which enable the mitigation of multiple attacks at limited or no utility cost. We note that, to the best of our knowledge, our study is the first to consider multiple *simultaneous* attacks on CML.
3. We evaluate the identified attacks and suggested model adaptations in the real-life context of medical image analysis (Section 6), showing that simple modifications of the training protocol can substantially reduce attack effectiveness *in practice*
4. Finally, we contrast model adaptation strategies with formal privacy-preserving mechanisms and identify their corresponding effects on the utility of the joint model (Sections 7 and 9).

2 Related work

This work is motivated by a gap in knowledge between the domains of privacy and practical machine learning. In most previous works that discuss the issues of adversarial interference and mitigations that are deployed to reduce their effectiveness, the learning settings are unrealistically biased against the adversary [13, 21–25]. It is often the case that the adversary is assumed to be oblivious to the fact that such defense mechanisms are deployed, thereby only executing the most basic variations of attacks without accounting for any potential mitigations in place.

Some prior works consider the effects of individual model attributes (such as activation functions) on the resulting privacy leakage during model training, such as the works by Papernot et al. [26] or Avent et al. [27], however, they do not consider these adaptations as a means for defending against reconstruction attacks, but rather investigate their effects on the formal PP mechanisms and associated changes in the utility of the target model.

Works such as Shokri et al. [4] and Chakraborty et al. [28] discuss a broader classification of adversaries in the context of collaborative learning, but they only propose the said classification on a specific type of attack, whereas we explicitly keep our classification generic and applicable to any privacy-oriented attack in the field of CML.

Multiple prior studies [3, 7, 29–31] have investigated the effects of formal PP mechanisms as well as learning setting adaptations on the *utility-based* adversarial attacks, such as backdoor insertion and model poisoning. Our work expands these investigations into the setting of *privacy-focused* attacks, which –so far –have been less studied.

In our study, we demonstrate attacks and mitigations in the setting of medical ML as it represents a highly privacy-sensitive real-life learning context. Prior works have considered formal privacy-preserving mechanisms, as well as limited model adaptation strategies in the domain of medical ML [32, 33], but none were focused on a systematic evaluation of model adaptation strategies in this context, but rather on the derivation of domain-specific insights in the area of medical PPML.

A number of prior works such as [15, 21, 34] assess the effects of single attack types, however they don't discuss the broader context of their findings and the conclusions which can be drawn in the context of reconstruction attack prevention on a larger scale, where such attacks can be applied *simultaneously* to the same ML system. The investigation of simultaneous attacks is a key contribution of our work.

3 Background

We begin by briefly introducing the concepts, terminology and methods utilised in our study.

3.1 Privacy-preserving mechanisms

In this section we describe both empirical and provable privacy-preserving mechanisms that we deploy in this work.

3.1.1 Federated learning and split learning

Collaborative machine learning can be conducted in both a centralised and a decentralised fashion [35]. In this work we concentrate on federated learning [36], which allows participants to retain control over their data, as instead of committing raw datasets to the centralised entity, they opt for sharing the model updates that they generate after training the model locally. Once each selected participant completes the local training section, they send their update to the orchestration server, where they are aggregated in order to produce the joint model that is then sent back to the selected clients and the process is repeated. For our experiments, we employed federated averaging (FedAvg) as the aggregation technique for its simplicity and popularity, as it represents the most commonly used aggregation method [35, 37] and is, hence, most likely to be at risk.

Weight-based reconstruction attacks were performed in a setting, which utilised a variant of CML, namely Split Learning (SL) [38]. Here, the amount of information shared between the participants is reduced by only sharing the outputs of a specific layer (*cut layer*), normally the final layer before the classifier. The rest of the computation can be completed on a trusted server. As such, this allows the federation to observe a lower communication cost as well as a better empirical privacy. We note that, although FL and SL possess some privacy-enhancing properties, they are, on their own, not sufficient to protect the federation against privacy-oriented attacks and should primarily be thought of as *governance*-preservation mechanisms, by which we define techniques allowing alternative means of information sharing without assuming control over other client's data, such as access to information through local training and a subsequent model update, e.g. FL.

3.1.2 Differential privacy

A quantifiable measure of how much information an adversary can learn about individuals in the training data can be provided using Differentially Private training [39]. An algorithm is considered differentially private if its output is approximately invariant to the inclusion/exclusion of a single entity (e.g. data record/patient/institution based on the desired privacy guarantee). The privacy guarantees provided (and thus the privacy loss/ *budget*) are determined by a value ϵ, specifying the upper multiplicative bound on the information that can be gained by an adversary and a δ value which can be thought of as the probability that this gain exceeds the ϵ bound. DP training of neural networks is typically performed using the DP-stochastic gradient descent (DP-SGD) algorithm [40], which we also employed in our experimentation using the *pytorch-dp* (now *Opacus*) [41] framework.

3.2 Privacy attacks

3.2.1 Model inversion attacks

This class of adversarial interference concentrates on recovering data that participants have used to train the joint model by reverse-engineering the internal representations of either each individually submitted or of the aggregated models. We chose attacks based on two distinct reconstruction strategies that have proven successful in previous works described below. These attacks can be particularly destructive in CML settings relying on sharing unencrypted or otherwise unprotected gradient or model updates, as they can be conducted from within the learning consortium and are very effective under a white-box access scheme.

Deep leakage from gradients (DLG)

DLG [23] utilises unencrypted gradient updates to reconstruct the training data. The adversary captures an update submitted by an honest participant and runs an optimisation algorithm on the randomly initialised data-label pair that they control in order to mimic the data-label pair that has generated the original model update. In theory, the attack allows the adversary to achieve full disclosure of sensitive training data. In practice however, the attack requires a number of assumptions to hold (e.g. shallow models, small batch sizes, etc.), which makes it amenable to specific model adaptations shown below.

Generative decoder (GD)

The GD attack [6] mimics the structure of an autoencoder. In a collaborative setting, the *encoder* component is the jointly trained model, whereas the *decoder* is the adversarial model that is trained on a disjoint dataset (which often comes from the same distribution). The goal for the attacker is to decode the model outputs into the corresponding training images, while only having access to activations of the shared model. Such a scenario is realistic for e.g. an aggregation server in FL.

3.2.2 Membership inference attacks

Membership Inference Attacks (MIA), proposed by Shokri et al. [5] intend to determine whether or not a specific training record has been part of the training dataset. Various techniques can be utilised to this end, the most widely used being “shadow training”, first proposed by the same authors, in which the adversary trains a number of models that mimic the behaviour of the target model on the disjoint datasets with known ground-truth values. The outputs of these models, alongside the data record, whose membership information is being determined, are fed into a binary classifier that returns the prediction of this data record being in the training dataset.

In environments where the model has been deployed in a black-box setting (e.g. in the setting of *inference-as-a-service* in the cloud) the MIA attack can still succeed, as information about the records that have been used to train the model is encoded into the parameters of the network, thus predictions on these records will have higher confidence, revealing their membership in the original dataset. Here, we utilise the publicly available `mia` library [42] which utilises “shadow training” in order to evaluate the effectiveness of our proposed model adaptations under a black-box adversary.

4 Methodology

For all following privacy-focused attacks we employ an expanded classification based on the work by [4] and distinguish between the following adversarial settings: *Attack Time* (Train or Test), *Position in the Network* (Client or Server or Off-path), *Model Access* (Black-box or White-box) and *Security Model* (Malicious or Honest-but-Curious). In accordance with these categories, we selected three attack scenarios: the *Generative Decoder* [6], *Deep Leakage from Gradients* [23] and *Membership Inference Attack* [5] in order to cover most realistic threat models. We find these attacks to be applicable to most generic collaborative learning environments (such as SL, FL or centralised model trained on securely aggregated data), as they can target both the model that is being trained and the model that has been deployed in an inference setting. All attacks were executed by a single adversarial entity.

4.1 Threat model

GD and DLG were conducted under the following threat model: Train-time attack, client network position (for 2 clients) or server network position (for 3+ clients), white-box model access. Thus, we assumed that there is no Secure Aggregation (SecAgg) of model updates by default, allowing us to visualise the consequences of a federation neglecting SecAgg, facilitating training data reconstruction. Since fully encrypted training [43] is not yet possible without an unacceptable time complexity penalty, we believe this threat model to cover the majority of real-life use-cases. Additionally, for GD attack the threat model can be adapted to He et al. [6] with: Test-time attack, server network position, white-box access in a collaborative inference mode as per original work. However, we note that utilising a test-time server-side attacker, can be considered to be a stronger adversarial assumption due to them being in full control of the pre-deployed model. For MIA, the following threat model was chosen: Test-time attack, off-path position, black-box model access. All attacks utilised an Honest-but-Curious (HbC) security model.

4.2 Learning setting

We define the standard learning setting (unless explicitly stated otherwise) as a classification task with three clients in the federation holding Independently and Identically Distributed (IID) image data (the dimensions of which vary between 1x28x28, 3x32x32 and 3x64x64) with a batch size varying between 1 and 256, depending on the learning task. This setting corresponds to a *cross-silo* horizontal FL scenario, as is widely employed in previous works [35]. In the setting of a classification task used in our study, we define an IID distribution as one in which each client has access to data from all the classes present in the dataset, similar to [44]. In contrast, we define a non-IID distribution as one where each client has access to a subset of individual labels disjoint from the rest of the federation. The optimiser used was Adam with a learning rate of 10^{-4}, the number of training rounds was three, the number of epochs per client was 10 and FedAvg was used as the means of update aggregation. The choice of model depends on the data and the nature of learning task, with most tasks utilising LeNet [45] or AlexNet [46] with ReLU activation functions, unless stated otherwise. Experiments associated with the Generative Decoder additionally employed SL, only allowing the clients to share the outputs of the final pooling layer, with the classification being executed on the aggregation server. The FL process was locally simulated. For the subset of experiments using DP, training was performed using the SGD optimiser.

Over the course of our experiments we evaluated the following learning (hyper-)parameters that are common for all settings: model depth, model width, data dimensionality, data complexity, data distribution among clients, choice of specific layers (including pooling and activations), batch size, number of clients. We additionally deployed DP as a comparative measure in order to evaluate the effectiveness of model adaptations on their own and determine if these are sufficient or if additional PP mechanisms should be employed. Finally, we also investigated the effect of final model utility on attack quality to determine whether attacks on models with higher utility were more successful.

For the DLG attack we used the L-BFGS optimiser on the pair of adversarially controlled data and label. We utilised the Mean Squared Error (MSE) between the original and the computed gradients and the attack was run for 300 iterations, similarly to the original implementation of DLG [23].

For the Generative Decoder attacks, we designed an adversarial model with the same number of transposed convolution layers as the model under attack had regular convolution layers before the cut layer. The optimiser used by the adversary was Adam with a learning rate of 10^{-4}, using MSE between the reconstruction and the original image from an attacker controlled dataset as a loss function. The decoder was trained for 50 epochs. Similarly to the original work, we executed the attack on the target model that utilised SL in a FL setting. To test the effect of weakly informative priors being used for training the adversarial model, we performed the attack on an MNIST classification task, with adversary having access to the EMNIST dataset to train the decoder. Similarly, we performed an attack on CIFAR-10 while the attacker utilised the MNIST dataset or the disjoint classes of CIFAR-100 as priors. Finally, for the attacks on MedMNIST datasets (PneumoniaMNIST, PathologyMNIST and DermatologyMNIST) we utilised MNIST, ChestMNIST and CIFAR-10 datasets as adversarial priors.

For MIA we used five shadow models that were trained for 30 epochs each using an Adam optimiser with a learning rate of 10^{-4}. Each shadow model consisted of four fully-connected layers with 128 neurons each; utilised ReLU activation functions separated by *drop-out* layers with a probability of 0.2.

In order to evaluate the effectiveness of the GD attack, we employ the following similarity metrics between the original image and the reconstructed image: MSE, Structural Similarity Index Measure (SSIM) and Peak Signal-to-Noise Ratio (PSNR) [47]. To evaluate the effectiveness of the DLG attack we employed the MSE between the generated and the original model updates. For MIA we employed the accuracy metric to quantify how many target images have been classified correctly in regards to their membership.

5 Results

We describe the results ordered by their difficulty of mitigation based on the corresponding threat models.

Gradient-based reconstruction attacks can be mitigated through the smallest number of model adaptations or through application of a single PP mechanism (such as DP). However, the attacker can then adapt their methodology to exploit the intermediate activations, effectively bypassing some of the adaptations made by the federation. Through the addition of more empirical defenses, the federation can mitigate attacks based on capturing intermediate activations, but this is often not sufficient to fully mitigate other privacy-oriented attacks such as MIA. Therefore, we concluded our experimentation section with an analysis of possible adaptations that can be deployed to reduce the effectiveness of MIA, that can be applied in a fully black-box setting without any knowledge of the learning context. We deduce that being able to mitigate this attack provides the federation with strong empirical foundations of privacy, effectively minimizing the amount of information the adversary can obtain irrespective of the learning setting.

5.1 Deep leakage from gradients

Across all experiments, the common trend was that the success of the attack depended was associated with the complexity of the data used to train the model. For instance, we found that if the dimensions of the data are larger than 32x32 pixels, the DLG algorithm (in its original implementation) never converged, not even providing partial reconstructions. In contrast, decoder-based attacks were still able to reconstruct certain features of the training image, albeit incomplete. Hence, this attack can be circumvented if the dataset used to train the model is complex enough. Data complexity comes from both the dimensions of the data as well as from the number of input channels: we note that utilisation of single-channel images leads to better attack performance when compared to three-channel images. Model complexity also has a negative effect on the performance of this attack. The deeper the model is, the more difficult it is for the reconstruction to converge as can be seen in Table 1. For the original models that were deeper than LeNet (such as AlexNet), we failed to obtain correctly reconstructed images at all while utilising the original DLG implementation both in an IID and in a non-IID setting.

Additionally we found this attack to be very sensitive to the choice of batch size, and no reconstruction was possible for batch sizes larger than two. Hence, we deduce that even simple adaptations can mitigate this attack. However, we note that a more practical implementation of this attack by Geiping et al. [15] works for batch sizes up to eight, which we address in the discussion. We note that even in this implementation, larger batch sizes can be used to mitigate the attack, highlighting the importance of these seemingly simple model adaptations. This allows us to put our findings into a real-world perspective and evaluate how identical model adaptations behave under a stronger adversary that we showcase in Section 6.

	MNIST		CIFAR-10	
	LeNet	AlexNet	LeNet	AlexNet
Number of iterations for MSE of 0	31	N/A	84	N/A
Time taken to complete 300 iterations (m:ss)	0:45	0:31	1:54	0:53
Final MSE value	0.02	35112	0.05	30753

Table 1. Results of DLG with varying model architecture

We report that utilisation of certain layers have potential to reduce the effectiveness of this attack: MaxPooling layers, as noted by Geiping et al. [15], are difficult to invert, resulting in incomplete convergence or no convergence for certain inputs. We explored the significance of this finding in a larger context in Section 7. Additionally, we discovered that a higher number of filters has a potential to improve the results of the attack. However, we also found that utilisation of wider convolutional layers in an architecture similar to LeNet did not result in a successful reconstruction. Similarly, if the number of filters was too low, the attack did not generate any meaningful reconstruction either. Thus, we deduce that adaptation of convolutional layers has a potential to mitigate the attack. We summarise these observations in Table 2.

	Conv(16)	Conv(64)	Conv(128)	Conv(256)
Number of iterations for MSE of 0	N/A	84	92	N/A
Time taken to complete 300 iterations (m:ss)	1:03	1:54	2:40	0:52
Final MSE value	2.65	0.02	0.06	69.7

Table 2. Results of DLG with varying number of Conv filters.

Additionally, we note that wider fully connected (FC, also *linear*) layers also have an ability to reduce the effectiveness of the attack. The results for these experiments were summarised in Table 3. Moreover, this attack was extremely sensitive to the complexity of the model overall, as both higher width and depth could mitigate it. Similarly to Geiping et al. [15], we were able to confirm that discontinuous functions such as ReLU in the target model hinder the attacker's convergence while utilising L-BFGS as the adversarial optimiser, and hence models with such activations resulted in no convergence more often when compared to identical models that employed TanH or Sigmoid (logistic) activation functions instead.

	FC(16)	FC(64)	FC(128)	FC(256)
Number of iterations for MSE of 0	43	49	81	N/A
Time taken to complete 300 iterations (m:ss)	1:39	2:56	2:58	2:50
Final MSE value	0.01	0.02	0.06	3.41

Table 3. Results of DLG with varying width of FC layers

When comparing the performance of the attack on a trained and an untrained model, we note that in most scenarios both model types can be successfully inverted. Of note, untrained models normally resulted in faster and more frequent convergence, as the gradient norms of untrained models are typically larger, therefore usually revealing more information about the data at each training step, as has been illustrated in [15, 20]. However, we note that while there was no strong difference between the two trained models, the larger gradient norms caused by the higher loss of an untrained model meant that each update contained more information and thus facilitated reconstruction compared to models which were nearly fully trained and had lower gradient update norms. This difference was particularly apparent when models were trained on more complex datasets such as CIFAR-10. For simpler datasets such as MNIST, we did not observe any substantial difference between a final round model and an untrained model, which we assume to be caused by the simplicity of reconstruction of smaller datasets.

Finally we employed a formal PP mechanism, namely client-level (local) DP, which resulted in full attack mitigation even for large values of $\epsilon > 5.0$, highlighting that, even with relatively weak privacy guarantees, such attacks can be mitigated. We stress that the same does not necessarily hold when noise addition is performed on the aggregated model (*aggregate-level* DP). Thus, an adversary controlling an aggregation server in such a setting, would be able to perform such attacks more easily. Hence, *aggregate-level* DP is only suitable when the adversary is assumed to be a client rather than a central server, or when SecAgg is employed.

Overall, DLG is less likely to be a threat to a collaborative learning protocol due to the number of assumptions that have to be satisfied before the attack can be executed. While this issue has been partially alleviated in more recent and advanced attack implementations ([15, 24]) to e.g. support larger effective batch sizes (associated with a higher client count), the fundamental limitation of the attack remains and does not permit the adversary to reconstruct large batches of images.

5.2 Generative decoder

As seen above, even a limited number of model adaptations is sufficient to empirically protect the federation against the gradient-based attacks. As a result, we consider a separate exploitation vector that does not rely on the shared gradients, but on the shared activations instead. We consider this change in adversarial strategy to be realistic, as the federation cannot assume that the attacker to only be limited to a single model inversion technique under the same threat model.

While adapting the attack to train time, in order to enhance privacy of the federation without utilising additional PP techniques, we conducted an experiment with clients employing a combination of FL and SL [38]. This was intended to make sharing updates less of a burden on the network as well as added privacy enhancement, since the model was no longer shared in full. In this study we consider SL [38] which –similar to FL –was proposed as a PP mechanism, however also provides no formal privacy guarantees and is hence an example of learning setting adaptation. For this attack we assumed (unless explicitly stated otherwise) that the federation sends their activations after the second pooling layer, thus limiting the amount of information that the attacker can reconstruct from this data.

We conducted a number of experiments to determine the relationship between the accuracy of the target model and results of the reconstruction. We discovered that the accuracy of the target model greatly affects the outcomes of the attack, as poorly trained target models models with could not be used to generate any meaningful reconstructions. This was particularly noticeable for non-standard multi-channel datasets that require deeper models or a larger number of epochs to train, such as PathologyMNIST or DermatologyMNIST, re-

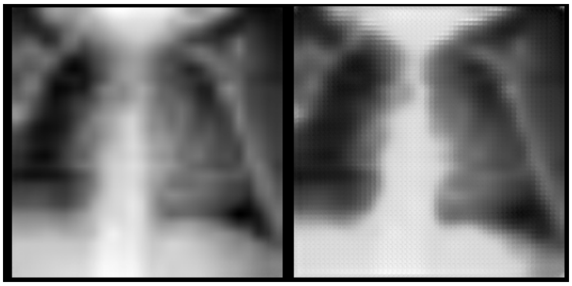

Fig. 1. Reconstruction of PneumoniaMNIST with priors: ChestMNIST (left) and MNIST (right)

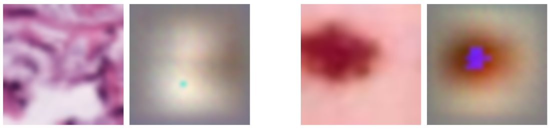

Fig. 2. Reconstruction results for PathologyMNIST (left) and DermatologyMNIST (right)

gardless of the adversarial prior. Reconstruction results for these two datasets are presented in Figure 2. We therefore deduce that, model accuracy seems to represent a vital component of a successful attack, and thus the adversary can achieve improved reconstructions after the joint model gained reasonable accuracy in the training task. In contrast, low accuracy led to meaningless reconstruction results that lacked detailed features and were not humanly recognisable. This conclusion contradicts the results of the DLG attack, where an untrained model reveals more about the training data than a fully-trained one, putting multiple training contexts at risk. We report results associated with the choice of target datasets and corresponding model accuracy in Table 4.

	MNIST	PneumoniaMNIST	CIFAR-10	PathologyMNIST
Model accuracy	91.2%	82.7%	55.2%	35.1%
Reconstruction MSE	397.4	408.3	451.3	548.5
Reconstruction SSIM	22.1	21.9	21.6	19.5
Reconstruction PSNR	0.13	0.12	0.09	0.04

Table 4. Results of GD attack (3 clients, IID, LeNet, 2nd layer SL). Datasets 1 and 2 use FMNIST prior; datasets 3 and 4 use CIFAR-100 prior.

Our next evaluation considered the effects of the relative positioning of the cut layer, which determined how much information is shared with the central server. We found that the cut layer's location relative to the rest of the model can reduce the effectiveness of the attack substantially, but SL alone was not a sufficient protection mechanism. Even placing the cut layer later in the network allowed the attacker to identify the general features associated with the training image. Similarly, early positioning of the cut layer can significantly improve the results of the reconstruction. This is because early layers contain data of higher resolution (including the non-robust, but highly descriptive ones). We present results for two distinct scenarios: cut layer at the first pooling layer and cut layer at the final pooling layer in Figure 3. Consequently, should the federation adopt the approach of sending the activations of the final convolution/pooling layer to the central server, deeper models will be leaking inherently less information about the training data. As a result, we found that deeper models can be effectively deployed to reduce the effectiveness of this attack. However, this can only be the case if employing a deeper model comes with a later cut layer, thus revealing less information. Alternatively, if the cut layer position remains unchanged, deeper models have a much higher accuracy on complex tasks, therefore, making the attack (along with the main learning task) more successful. However, we also note that due to the fact that this attack relies on decoding the features supplied by individual clients, additional width in the convolutional layers improved the reconstruction results. These results are summarised in Figure 4. We also report that additional width in the FC layers did not have a notable impact on the attack.

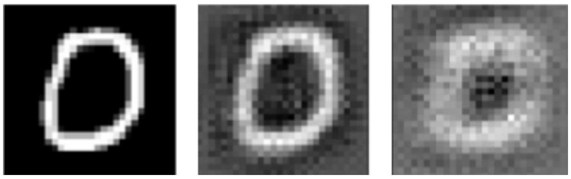

Fig. 3. Reconstruction of MNIST with varying position of the cut layer (LeNet): original image (left), early cut layer (centre) and deeper cut layer (right).

When assessing the relevance of data distribution across the federation, we found that on non-IID data, the attack results were less accurate in comparison to IID data across all datasets. As described in Section 3.2, we consider a scenario in which each client only has access to a disjoint subset of the dataset to be *non-IID*. This observation suggests that potential overfitting in the context of this attack either does not contribute to a more accurate reconstruction or that the effect of overfitting is offset by the significance of the overall model accuracy. We compared the performance of the attack

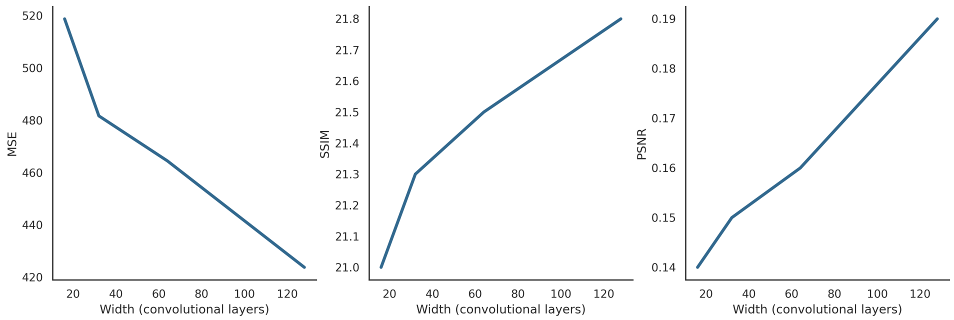

Fig. 4. Effects of model width (convolutional layers) on GD

in the IID setting against the non-IID setting in Figure 5.

A further noteworthy observation is that the attacker only requires data from the same *distribution* as the victims and does not directly from the same *dataset* as the training data for most experiments. This result makes this attack more risky, as it extends the possibilities for the adversary to extract the training data without any prior assumptions on what the data represents, but only knowing its dimensions. We confirmed this observation by performing the attack on the PneumoniaMNIST dataset, utilising MNIST and ChestMNIST as two separate priors. As can be seen in Figure 1, while the latter provided the attacker with more defined features, both attacks resulted in image reconstruction.

	MNIST (MaxPool)	MNIST (AvgPool)	CIFAR10 (MaxPool)	CIFAR10 (AvgPool)
MSE	464.1	429.1	462.4	441.2
PSNR	21.5	21.8	21.7	22.0
SSIM	0.15	0.16	0.08	0.09

Table 5. Results of GD attack (3 clients, IID, LeNet, 2nd layer SL). AvgPool and MaxPool effects on MNIST and CIFAR10.

We observed that similarly to DLG, data with higher resolution was more difficult to invert. This is attributed to assumptions about the data that is available to the attacker to train the decoder. When the prior was not consistent with the data in the cases of 1x28x28 datasets, the loss in quality was insignificant. However, when the attacker attempted to reconstruct larger 3x32x32 datasets, they were not be able to accomplish an accurate reconstruction without a suitable prior as can be seen from Figure 6.

	MNIST (ReLU)	MNIST (Tanh)	CIFAR10 (ReLU)	CIFAR10 (Tanh)
MSE	464.1	558.5	462.4	498.4
PSNR	21.5	20.7	21.7	21.6
SSIM	0.15	0.10	0.08	0.10

Table 6. Results of GD attack (3 clients, IID, LeNet, 2nd layer SL). ReLU and TanH effects on MNIST and CIFAR10

Another adaptation that had a similar effect on the GD attack as on the DLG attack was the usage of pooling layers. Similar to our findings above, Max-Pooling tended to reveal less information in comparison to Average-Pooling; to our knowledge, this finding has not been discussed in prior work in this particular attack setting. The results of these experiments are summarised in Table 5. Thus, as these two experiments showcase, the usage of a specific pooling layer type can reduce the effectiveness of multiple attacks simultaneously. Contrary to DLG, where ReLU had a mitigating effect, we found Sigmoid (logistic) activations to mitigate GD, however this may be due to the, on average, lower accuracy of models utilising this activation function. The results of these experiments are summarised in Table 6.

Finally, we note that increasing the batch size at train time also rendered the attack ineffective, similar to DLG. At inference time, similarly to He et al. [6] we assumed that clients were interested in obtaining the results for one image at a time. If this assumption was violated, the attacker was facing the same difficulty as before and the attack could be circumvented.

In order to evaluate the effects of formal PP mechanisms, we applied client-level Differential Privacy (DP-SGD, adapted from [41]) during training on each individual client node, resulting in inaccurate noisy reconstructions for the adversary. However, we note this result could be attributed to a performance degradation

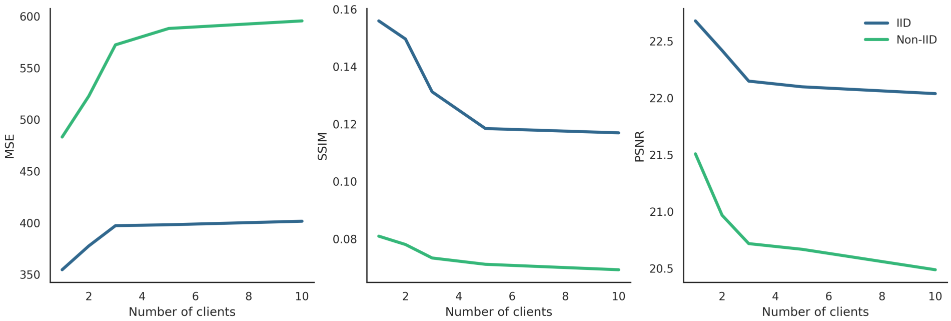

Fig. 5. Effects of data distribution on the reconstruction results

associated with DP as well as to the application of DP itself.

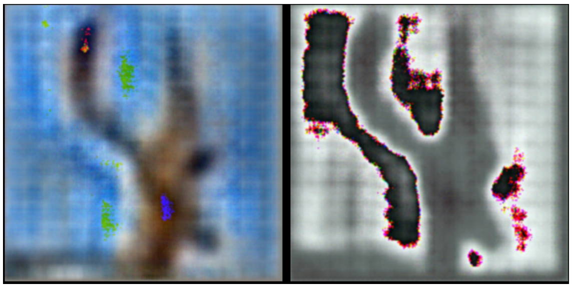

Fig. 6. Reconstruction of CIFAR-10 with priors: CIFAR-100 (left) and MNIST (right)

5.3 Membership inference attack

We now study the situation in which the attacker is prevented from a successful reconstruction attack due to them being positioned outside of the network, or because the federation utilises security and confidentiatlity measures that mask their contributions (such as SMPC). In such cases, attacks such as MIA can still succeed.

In line with previous works [5, 12], we assessed the significance of model complexity on the results of MIA. By model complexity we define the number and width of model layers. Our experiments showed that for models under MIA, the deeper the model is, the less accurate the result of the attack is likely to be. This suggests that an addition of more convolutional layers reduces the amount of information that is memorised by the network about each **individual** training instance, but may improve generalisation, resulting in higher accuracy for the federation and lower accuracy for the attacker. However, as noted by Shokri and colleagues [4], more complex models are at a higher risk of overfitting, resulting in more accurate MIA due to an increased propensity for data memorisation. Our experimental evidence further suggested that increasing the number of filters in the convolutional layers could significantly increase the accuracy of the attack. In contrast, neither an increase in the number or the width of the FC layers resulted in notable difference in MIA results. For this experiment, differently to GD and DLG, we utilised the CIFAR-10 dataset, as simpler datasets produced negligible variation across most experiments, showing very marginal changes in regards to model complexity. Findings of the model complexity experiments are shown in Table 7.

After analysing the effects of the target model itself on the results of the attack, we now consider the effects of the training data. We noted that the more complex the training dataset is, the more uniquely identifiable features each individual training image has that allow the adversary to link it to a specific client. Using the fact that each model *memorises* the data it was trained on and behaves differently during the update phase in comparison to the data it was not trained on, it is easier to exploit images of larger dimensions with three channels in comparison to simpler single-channel images. This is likely to be the case because there are fewer features that help distinguish simple data points from each other in comparison to more complex data points with more diverse feature sets that describe them more uniquely. Additionally, we determined that overfitting of the target model (and hence attacking during an earlier training round in the non-IID setting) results in a more accurate inference. We also noted that the accuracy of the target model has on its own an insignificant effect on the results of the attack, when comparing two trained models. However, since MIA relies on the model being able to memorise information about individual data points, poorly trained or untrained models

	Conv(16)	Conv(32)	Conv(64)	Conv(128)	FC(64)	FC(128)	FC(256)	FC(512)	FC(1024)	FC(2048)
Model accuracy	74.8%	80.1%	85.1%	92.1%	77.3%	76.9%	78.8%	78.1%	78.0%	78.1%
Attack accuracy	54.0%	57.5%	59.2%	64.2%	54.1%	53.5%	55.2%	54.4%	53.8%	54.3%

Table 7. Impact of model width on MIA (3 clients, IID, LeNet, 30 epochs, CIFAR-10)

provide severely worse inference results. Overall for MIA we found it much more difficult to determine the effects of each individual adaptation, since the attack can benefit significantly from overfitting [5] and each of these factors can have an effect on the overfitting of the target model as well as on the attack itself. Results from both the data complexity and model complexity experiments can be found in Table 8

	MNIST	FMNIST	CIFAR-10	CIFAR-100
Model accuracy (LeNet)	99.8%	92.7%	65.5%	43.5 %
Attack accuracy (LeNet)	51.1%	51.4%	60.1%	54.6%
Model accuracy (AlexNet)	99.1%	93.4%	77.6%	62.7%
Attack accuracy (AlexNet)	50.9%	52.5%	56.5%	51.5%

Table 8. Results of MIA with varying model and data complexities (3 clients, 30 epochs)

Finally we note that were no effects on the results of the attack when adapting batch size or activation functions in the target models across all datasets. This distinguishes the attack from the previously discussed ones, as MIA can hence be deployed in settings, which have been adapted to withstand the reconstruction attacks discussed above.

	MNIST	MNIST DP	CIFAR-10	CIFAR-10 DP
Model accuracy	98.8%	98.6%	77.2%	71.5%
Attack accuracy	50.8%	50.4%	59.8%	54.3%

Table 9. Impact of DP on MIA (3 clients, DP-SGD, LR=$1e-2$, 50 epochs, $\epsilon = 1.751$, $\alpha = 14.0$)

After evaluating the empirical defenses, we deployed a provable privacy-preserving strategy, namely DP. The addition of DP reduced the effectiveness of this attack, but for small values of ϵ (up to 1.8) resulted in a performance degradation as can be seen in Table 9. Thus, while DP substantially reduces the success of inference attacks, it also negatively impacts model utility.

6 Application to medical imaging datasets

To evaluate our results from benchmark datasets, we finally performed experimentation on medical imaging, which represents a particularly privacy-sensitive domain. This allowed us to place our findings from Section 5 into a real-world context and assess the applicability of model adaptation on sensitive datasets. Initially, we utilised a similar learning setting as described above, with a batch size of one, two clients with IID data and unencrypted non-private gradient sharing. Moreover, to accommodate for a more complex model (ResNet18), we employed an improved version of DLG by Geiping et al. [15] on the publicly available dataset of paediatric chest radiographs originally published by Kermany et al. [48], and recovered chest x-rays that were almost indistinguishable from the original images used for training. This exemplifies that such attacks can lead to catastrophic violation of privacy in standard collaborative settings. Reconstruction results can be found in Figure 7

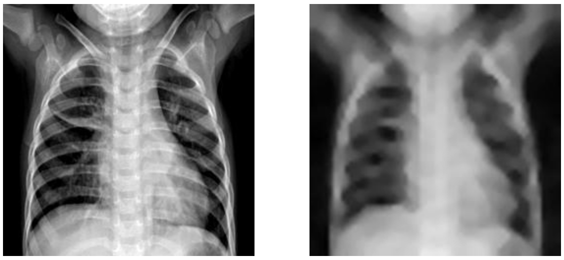

Fig. 7. Reconstruction results for the paediatric pneumonia dataset: original (left) and reconstructed (right)

To showcase the empirical privacy improvements from simple model adaptations, we changed a small number of training settings and compared the results to an unprotected reconstruction. We changed the effective batch size to 30 (batch size of 10 per each client in a federated setting with three clients), which did not result in an accurate reconstruction for the adversary. Additionally, we conducted an attack on a DP-trained model (with $\epsilon = 6.0, \delta = 1.9 \times 10^{-4}, \alpha = 4.4$) for comparison. The results can be found in Figure 8. This allows us to observe that while simple model adaptations are empirical methods, they can reduce the effectiveness of this attack, without the utility penalty typically associated with information-theoretic PP protocols. Moreover, a higher effective batch size can be obtained through two distinct adaptations: an increase in training batch size per client, as well as an increase in the size of the federation. The latter would also allow to empirically mitigate attacks performed by an HbC client, since they would not be able to distinguish the updates submitted by their victim from an aggregated update shared by the central server.

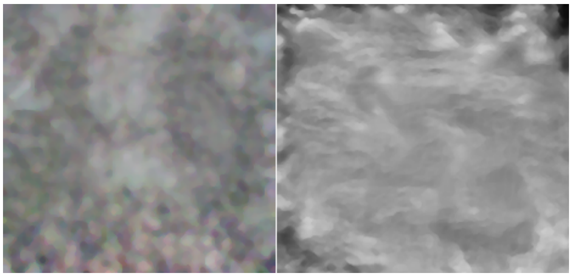

Fig. 8. Reconstruction of 4P while deploying: Model adaptation (left) and DP training (right)

7 Discussion

In this study we experimentally evaluated the effects that simple adaptations of the collaborative learning setup have on privacy-oriented adversarial attacks. After analysing the results of our experiments, we observed a number of model adaptations that reduce adversarial effectiveness under all attack scenarios.

For instance, deeper models reduced the effectiveness of the attack, regardless of which adversarial method is used. Under MIA, these models are more difficult to attack, as they have an improved capacity for model generalisation and do not memorise the features that distinguish specific data points or clients. However, at the same time as noted by Shokri et al. [5], more complex models tend to be prone to overfitting, generating additional risk of data memorisation, which should be taken into account when training the model in data-deficient regimes. When compared to other attacks, GD relies on the filters of the attacked model having captured sufficient information from the dataset to reconstruct the training data and therefore, the deeper the model is (and while utilising SL, the further away from the input the cut layer is), the more difficult it is for the attacker to achieve reasonable reconstruction accuracy. This is due to a loss of features through additional pooling layers. For a DLG adversary, a deeper model means a higher computation cost associated with the attack, resulting in no convergence for simpler attack implementations. Utilisation of specific pooling techniques is also associated with a reduction in adversarial information gain across all reconstruction attacks. This finding becomes significant in the context of application of formal PPML techniques: utilisation of Max-Pooling layers remains non-trivial when computed privately through protocols such as SMPC and HE for which the necessary comparison operators are not easily or efficiently implemented. The federation is then forced to either adapt their models' architectures to be compatible with such protocols by replacing layers, or opt for an empirical model adaptation approach instead, without taking an associated utility penalty. Lifting limitations to unlock the full extent of model architectures when implemented over blind computing protocols is an active area of research both in the fields of SMPC [49, 50] and HE [51].

Similarly, for a larger number of clients, all attacks we discuss failed to obtain meaningful results, be it a high confidence inference prediction or an image similar to the one used during the training process. These results are summarized in Table 10 and highlight a suitable baseline for most CML contexts that are empirically shown to reduce the effectiveness of privacy-centred attacks. This allows collaborators to benefit from a higher empirical privacy, while not paying the associated performance penalty associated with formal PPML mechanisms. While this can be considered as a *setting adaptation* rather than a model adaptation, such finding allows us to recommend that smaller federations employ formal PPML mechanisms, as they are more vulnerable to all attack types that we have discussed in this work, instead of relying on model adaptations alone. One important observation is that deploying adaptations such as increasing client population or making changes to the batch size, can result in attacker's threat model being ineffective. This issue arises when the adversary can no longer reliably determine which update corresponds to which data-client pair as an HbC client-side attacker, forcing them to assume the position of an HbC server, which they often cannot do in practise, rendering the attack unsuccessful in such settings. However, we note that adaptations that involve changing the learning setting, rather than the model itself can be more challenging to implement

	Model Adaptations						Privacy-Preserving Mechanisms	
	Increasing the number of clients	Increasing data dimensionality	Deeper models	Wider models	Large batch sizes	Discontinuous activations	Differential Privacy	Secure Aggregation
Generative Decoder	Yes	Yes	Yes	No	Yes	No	Yes	Yes
Deep Leakage from Gradients	Yes	Yes	Yes	Yes	Yes	Yes	Yes[1]	Yes
Membership Inference	Yes	No	Yes[2]	No	No	No	Yes	No

Table 10. Proposed mitigations

[1] Client/patient-level DP

[2] Can be ameliorated through model overfitting

in practice, thus making them only effective under a number of assumptions are satisfied.

Additionally, as can be seen from Table 10, certain adaptations that are deployed can conflict with each other when it comes to mitigating different attacks simultaneously. In a case of privacy based attacks, MIA benefits from more complex data, whereas DLG can fail to converge on a data that's dimensions are too large. There are also factors such as batch size that have an ability to fully mitigate reconstruction attacks to a certain extent, while being completely irrelevant under a MIA adversary. The most notable adaptation that does not have a singular defined affect on the attacks is the model width. While wider convolutional layers tend to provide the adversary with more information about the training data, they also make certain reconstruction tasks, such as DLG, more challenging. Additionally, there is no clear affect of the deployment of wider FC layers on the attacks, as wider layers either do not affect certain attack at all or, in fact, marginally reduce the attack's effectiveness, but are offset by the effects of wider convolutional layers in the context of more complex models overall. Consequently, we deduce that forming a single set of adaptations that would hold effective against *any* attack type is still an open challenge, but we hope that our work provides a suitable baseline and outlines directions that the future work may take. As a first step, some of our findings are effective not only against a number of privacy-centred attacks discussed above, but also against their derivatives relying on similar exploitation vectors. For example, the utilisation of larger models is not only effective against the MIA implementation our work is based on [5] but is also effective against variants [52, 53] which, like MIA, are based on the concept of shadow training.

Furthermore, we note that based on our experimental evaluation of model adaptations under MIA, the results that we obtain raise a number of questions on the dependency between overfitting, model architecture and data complexity. Particularly, we find it challenging to determine to what extent each one of these components individually affects the results of the attack. Such challenge arises from the fact that these components are often tightly coupled and it is difficult to isolate their effects when it comes to evaluating the results of a seemingly "more successful" attack. Consequently, we see an open area of future work in determining the precise relationship between these components in regards to MIA accuracy.

Our experimental results raise questions about the effectiveness of solely adjusting the learning settings as opposed to leveraging methods that can be deployed alongside the training protocol such as SMPC or DP. However, most of these formal protocols do not have the ability to circumvent all privacy-centred attacks either: protecting data from reconstruction does not provide the methods that achieve perfect secrecy and confidentiality, such as HE or SMPC, with provable guarantees of privacy when under MIA. Indeed, most attacks that target membership information only require black-box access to the model [4, 54–57] and achieve high attack accuracy without utilising any additional information about the victim. The utilisation of DP can result in an unacceptable utility penalty and HE is typically very computationally expensive for the training of deep neural networks [58–60]. However, the utilisation of SecAgg (even without fully encrypted training) can still mitigate reconstruction attacks, and may integrate better in the context of collaborative learning. This is due to a smaller effect on the performance of the model [10, 61] and the attacker's inability to utilise shares of individual updates to achieve privacy violation. We note that a hybrid system that utilises DP and SecAgg is an optimal choice for tasks that prioritise confidentiality over performance, as such a combination of algorithms allows the clients to be provably defended against both reconstruction and inference attacks simultaneously [20].

Moreover, similarly to other works [62] we note that, in addition to performance degradation, DP is associated with further challenges, such as the selection of optimal noise parameters [63], issues of unfairness to underrepresented datasets during training [64] and a *false sense* *of privacy* for large values of ϵ. This last issue can prove problematic as it leads clients to believe that their data is secure and maintains high utility, while it may be insufficient to prevent adversarial interference [65]. Thus, we highlight the requirement for further investigation to facilitate the large-scale deployment of DP.

8 Limitations

We consider a number of limitations to model adaption, which arise from the experimental evidence and discussion above, especially compared to formally security and/or privacy mechanisms. Firstly, we highlight that the adaptations we discuss are mostly applicable to privacy-centred adversarial contexts, therefore providing limited protection against utility-centred attacks such as backdoor insertion [7] or model poisoning attacks [3]. However, we note that a similar line of discussion is also applicable to formal PETs such as SMPC or HE which would, in fact, even allow the adversary to remain concealed while performing such utility-centred attacks. Secondly, as mentioned above, a subset of the discussed adaptations cannot be dynamically altered as the learning progresses. Thus, adversaries can, in principle, choose the attack method they deploy based on prior knowledge of which adaptations have been deployed by the training consortium. In addition, we highlight that there exist other variations of privacy-centred attacks such as *Attribute Inference Attacks* [4] or Generative Adversarial Model-based inversion attacks [13], which we have not discussed in this study. We consider these a promising area of future work investigating mitigations that are specific to exploitation vectors that the adversaries rely on in these contexts. For example, perturbation of the weights of the *trained* model can prevent attribute inference attacks, but is inapplicable to attacks occurring during model training. Finally, although we have covered a plethora of techniques (such as model width and depth, batch size, etc.) which represent fundamental choices in the learning process, we have omitted a detailed treatment of more situational techniques, such as examining the large number of available regularisers, or the use of mixed precision training and quantisation, which are not yet common practice in every machine learning context.

9 Conclusion

This work investigated the use of model adaptation techniques to mitigate privacy attacks in a collaborative learning setting. We selected three attacks from the literature that were designed to extract sensitive data in a joint learning task. We then evaluated the possibility of tuning parameters of the learning task itself to mitigate the attacks. Our experiments show the impact of factors such as model depth or layers forming its architecture on diminishing the success of an attack. However, we also noted that adaptations such as data complexity, model width, accuracy of the target model or batch size do not have a consistent effect on the fidelity of attacks when the threat model can be adapted to incorporate multiple attack simultaneously. Should the same adversary be able to utilise multiple attacks from various entry points, they would find certain reconstruction attacks benefiting from simpler data, whereas membership attacks might fail to achieve meaningful results.

We conclude that both model adaptations and formal PP techniques follow the exact same trade-off principle: prevention strategies deployed against privacy-oriented attacks increase the adversarial gain in scenarios of utility destruction and malicious model augmentation. Future work will thus explore the unified strategies combining model adaptation and privacy-preserving methods [16–18, 32] to simultaneously protect the federation from both privacy- and utility-focused attacks.

10 Acknowledgements

G.K. received funding from the Technical University of Munich, School of Medicine Clinician Scientist Programme (KKF), project reference H14. D.U. received funding from the Technical University of Munich/Imperial College London Joint Academy for Doctoral Studies. This research was supported by the UK Research and Innovation London Medical Imaging and Artificial Intelligence Centre for Value Based Healthcare. The funders played no role in the design of the study, the preparation of the manuscript or the decision to publish.

References

[1] "MELLODDY consortium." https://cordis.europa.eu/project/rcn/223634/factsheet/en. Accessed: November 21, 2020.

[2] T. Yang, G. Andrew, H. Eichner, H. Sun, W. Li, N. Kong, D. Ramage, and F. Beaufays, "Applied federated learning: Improving google keyboard query suggestions," *arXiv preprint arXiv:1812.02903*, 2018.

[3] L. Muñoz-González, K. T. Co, and E. C. Lupu, "Byzantine-Robust Federated Machine Learning through Adaptive Model Averaging," *arXiv preprint arXiv:1909.05125*, 2019.

[4] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 739–753, IEEE, 2019.

[5] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, IEEE, 2017.

[6] Z. He, T. Zhang, and R. B. Lee, "Model inversion attacks against collaborative inference," in *Proceedings of the 35th Annual Computer Security Applications Conference*, pp. 148–162, 2019.

[7] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *International Conference on Artificial Intelligence and Statistics*, pp. 2938–2948, PMLR, 2020.

[8] A. Team, "Learning with privacy at scale," *Apple Mach. Learn. J*, vol. 1, no. 9, 2017.

[9] P. Kairouz, K. Bonawitz, and D. Ramage, "Discrete distribution estimation under local privacy," *arXiv preprint arXiv:1602.07387*, 2016.

[10] E. Hesamifard, H. Takabi, and M. Ghasemi, "Cryptodl: Deep neural networks over encrypted data," *arXiv preprint arXiv:1711.05189*, 2017.

[11] O. Goldreich, S. Micali, and A. Wigderson, "How to play any mental game, or a completeness theorem for protocols with honest majority," in *Providing Sound Foundations for Cryptography: On the Work of Shafi Goldwasser and Silvio Micali*, pp. 307–328, 2019.

[12] L. Song, R. Shokri, and P. Mittal, "Membership inference attacks against adversarially robust deep learning models," in *2019 IEEE Security and Privacy Workshops (SPW)*, pp. 50–56, IEEE, 2019.

[13] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song, "The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks," *arXiv preprint arXiv:1911.07135*, 2019.

[14] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, "The secret sharer: Evaluating and testing unintended memorization in neural networks," in *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pp. 267–284, 2019.

[15] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting Gradients–How easy is it to break privacy in federated learning?," *arXiv preprint arXiv:2003.14053*, 2020.

[16] N. Papernot, S. Chien, S. Song, A. Thakurta, and U. Erlingsson, "Making the shoe fit: Architectures, initializations, and tuning for learning with privacy," 2020.

[17] L. Lyu, H. Yu, and Q. Yang, "Threats to federated learning: A survey," *arXiv preprint arXiv:2003.02133*, 2020.

[18] E. De Cristofaro, "An Overview of Privacy in Machine Learning," *arXiv preprint arXiv:2005.08679*, 2020.

[19] Y. Kaya, S. Hong, and T. Dumitras, "On the Effectiveness of Regularization Against Membership Inference Attacks," *arXiv preprint arXiv:2006.05336*, 2020.

[20] G. Kaissis, A. Ziller, J. Passerat-Palmbach, T. Ryffel, D. Usynin, A. Trask, I. Lima, J. Mancuso, F. Jungmann, M.-M. Steinborn, A. Saleh, M. Makowski, D. Rueckert, and R. Braren, "End-to-end privacy preserving deep learning on multi-institutional medical imaging," *Nature Machine Intelligence*, May 2021.

[21] S. Hidano, T. Murakami, S. Katsumata, S. Kiyomoto, and G. Hanaoka, "Model inversion attacks for prediction systems: Without knowledge of non-sensitive attributes," in *2017 15th Annual Conference on Privacy, Security and Trust (PST)*, pp. 115–11509, IEEE, 2017.

[22] Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, and H. Qi, "Beyond inferring class representatives: User-level privacy leakage from federated learning," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pp. 2512–2520, IEEE, 2019.

[23] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Advances in Neural Information Processing Systems*, pp. 14747–14756, 2019.

[24] B. Zhao, K. R. Mopuri, and H. Bilen, "iDLG: Improved Deep Leakage from Gradients," *arXiv preprint arXiv:2001.02610*, 2020.

[25] T. Orekondy, S. J. Oh, Y. Zhang, B. Schiele, and M. Fritz, "Gradient-leaks: Understanding and controlling deanonymization in federated learning," *arXiv preprint arXiv:1805.05838*, 2018.

[26] N. Papernot, A. Thakurta, S. Song, S. Chien, and Ú. Erlingsson, "Tempered sigmoid activations for deep learning with differential privacy," *arXiv preprint arXiv:2007.14191*, 2020.

[27] B. Avent, J. Gonzalez, T. Diethe, A. Paleyes, and B. Balle, "Automatic discovery of privacy-utility pareto fronts," *arXiv preprint arXiv:1905.10862*, 2019.

[28] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial attacks and defences: A survey," *arXiv preprint arXiv:1810.00069*, 2018.

[29] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv preprint arXiv:1712.05526*, 2017.

[30] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, "Certified robustness to adversarial examples with differential privacy," in *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 656–672, IEEE, 2019.

[31] M. Mozaffari-Kermani, S. Sur-Kolay, A. Raghunathan, and N. K. Jha, "Systematic poisoning attacks on and defenses for machine learning in healthcare," *IEEE journal of biomedical and health informatics*, vol. 19, no. 6, pp. 1893–1905, 2014.

[32] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Machine Intelligence*, pp. 1–7, 2020.

[33] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019.

[34] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing," in *23rd {USENIX} Security Symposium ({USENIX} Security 14)*, pp. 17–32, 2014.

[35] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, *et al.*, "Advances and Open Problems in Federated Learning," *arXiv preprint arXiv:1912.04977*, 2019.

[36] A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage, "Federated learning for mobile keyboard prediction," *arXiv preprint arXiv:1811.03604*, 2018.

[37] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *arXiv preprint arXiv:1812.06127*, 2018.

[38] P. Vepakomma, O. Gupta, T. Swedish, and R. Raskar, "Split learning for health: Distributed deep learning without sharing raw patient data," *arXiv preprint arXiv:1812.00564*, 2018.

[39] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2013.

[40] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, Oct 2016.

[41] "PyTorch-DP." https://https://github.com/facebookresearch/pytorch-dp. Accessed: June 29, 2020.

[42] B. Kulynych and M. Yaghini, "mia: A library for running membership inference attacks against ML models," Sept. 2018.

[43] E. Hesamifard, H. Takabi, M. Ghasemi, and R. N. Wright, "Privacy-preserving machine learning as a service," *Proceedings on Privacy Enhancing Technologies*, vol. 2018, no. 3, pp. 123–142, 2018.

[44] Q. Li, Y. Diao, Q. Chen, and B. He, "Federated learning on non-iid data silos: An experimental study," 2021.

[45] Y. Le Cun, L. D. Jackel, B. Boser, J. S. Denker, H. P. Graf, I. Guyon, D. Henderson, R. E. Howard, and W. Hubbard, "Handwritten digit recognition: Applications of neural network chips and automatic learning," *IEEE Communications Magazine*, vol. 27, no. 11, pp. 41–46, 1989.

[46] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[47] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th international conference on pattern recognition*, pp. 2366–2369, IEEE, 2010.

[48] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[49] S. Wagh, D. Gupta, and N. Chandran, "SecureNN: Efficient and Private Neural Network Training," p. 44.

[50] T. Ryffel, D. Pointcheval, and F. Bach, "ARIANN: Low-Interaction Privacy-Preserving Deep Learning via Function Secret Sharing," *arXiv:2006.04593 [cs, stat]*, June 2020. arXiv: 2006.04593.

[51] I. Chillotti, M. Joye, and P. Paillier, "Programmable bootstrapping enables efficient homomorphic inference of deep neural networks." Cryptology ePrint Archive, Report 2021/091, 2021. https://eprint.iacr.org/2021/091.

[52] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models," *arXiv preprint arXiv:1806.01246*, 2018.

[53] A. Sablayrolles, M. Douze, C. Schmid, Y. Ollivier, and H. Jégou, "White-box vs black-box: Bayes optimal strategies for membership inference," in *International Conference on Machine Learning*, pp. 5558–5567, PMLR, 2019.

[54] M. A. Rahman, T. Rahman, R. Laganière, N. Mohammed, and Y. Wang, "Membership Inference Attack against Differentially Private Deep Learning Model," *Transactions on Data Privacy*, vol. 11, no. 1, pp. 61–79, 2018.

[55] S. Truex, L. Liu, M. E. Gursoy, L. Yu, and W. Wei, "Demystifying Membership Inference Attacks in Machine Learning as a Service," *IEEE Transactions on Services Computing*, 2019.

[56] C. A. C. Choo, F. Tramer, N. Carlini, and N. Papernot, "Label-only membership inference attacks," *arXiv preprint arXiv:2007.14321*, 2020.

[57] Y. Park and M. Kang, "Membership Inference Attacks Against Object Detection Models," *arXiv:2001.04011 [cs]*, Jan. 2020. arXiv: 2001.04011.

[58] R. Gilad-Bachrach, N. Dowlin, K. Laine, K. Lauter, M. Naehrig, and J. Wernsing, "Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy," in *International Conference on Machine Learning*, pp. 201–210, 2016.

[59] B. D. Rouhani, M. S. Riazi, and F. Koushanfar, "Deepsecure: Scalable provably-secure deep learning," in *Proceedings of the 55th Annual Design Automation Conference*, p. 2, ACM, 2018.

[60] M. Barni, C. Orlandi, and A. Piva, "A privacy-preserving protocol for neural-network-based computation," in *Proceedings of the 8th workshop on Multimedia and security*, pp. 146–151, ACM, 2006.

[61] P. Mohassel and Y. Zhang, "SecureML: A system for scalable privacy-preserving machine learning," in *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 19–38, IEEE, 2017.

[62] S. L. Garfinkel, J. M. Abowd, and S. Powazek, "Issues encountered deploying differential privacy," in *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, pp. 133–137, 2018.

[63] J. Lee and C. Clifton, "How much is enough? choosing ε for differential privacy," in *International Conference on Information Security*, pp. 325–340, Springer, 2011.

[64] T. Farrand, F. Mireshghallah, S. Singh, and A. Trask, "Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy," *arXiv preprint arXiv:2009.06389*, 2020.

[65] J. Zhao, T. Wang, T. Bai, K.-Y. Lam, Z. Xu, S. Shi, X. Ren, X. Yang, Y. Liu, and H. Yu, "Reviewing and improving the gaussian mechanism for differential privacy," *arXiv preprint arXiv:1911.12060*, 2019.