Adversarial attacks on federated learning networks for medical image analysis

Erfan Darzidehkalani[1], Marianna Sijtsema[1] , P.M.A van Ooijen[1]
[1]*Machine learning lab, Data Science Center in Health (DASH)*
University of Groningen, Hanzeplein 1, Groningen, The Netherlands

***Abstract*—Federated learning (FL) can significantly mitigate privacy concerns for Medical image analysis (MIA) systems. However, its decentralized and collaborative nature can bring new ways of attacking which might impose severe threats for participating clients. This paper investigates adversarial attacks where the adversary tries to fool other clients with manipulated data. We investigate the credibility of known threat factors in a federated environment and discuss their importance. We demonstrate that domain-specific settings can lead to higher attacker success on MRI tumor and pathology imaging datasets. In addition, we propose a scenario in which the adversary leverages the federated environment to devise a more powerful attack. We show that using gradient information from previous global model updates enables single-step attacks (e.g., FGSM) to outperform computationally expensive iterative methods so that the adversary reaches the same success rate 20 to 30 times faster.**

***Index Terms*—Adversarial attacks, Federated learning, Medical imaging, Deep learning**

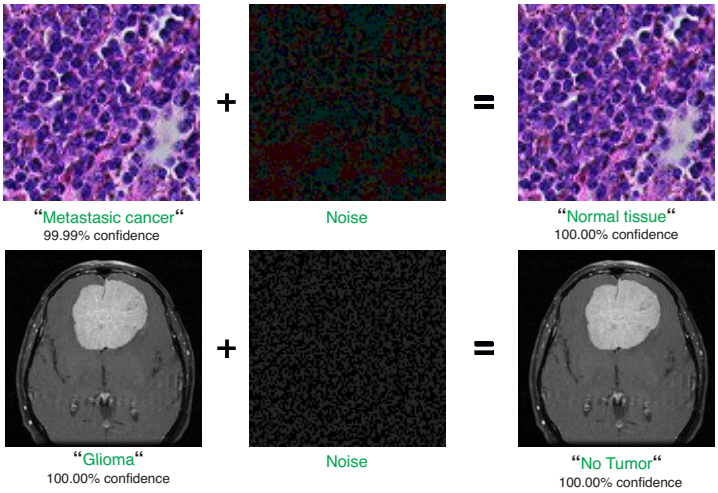


Fig. 1: A schema of adversarial attacks to cancer detection systems

I. Introduction

In the past few years, federated learning (FL) has emerged as one of the mainstream machine learning (ML) paradigms and gained much attention in the field of medical image analysis (MIA). Federated learning enables hospitals and other healthcare providers to train machine learning models jointly with other hospitals without sharing sensitive data. It is applied to a wide scope of MIA tasks, successfully addressing data governance concerns. Notable works include brain tumor classification and segmentation, breast density classification, and covid-19 detection. [1]–[3].

Although FL has mitigated many risks and concerns about multi-institutional collaborations, it is still vulnerable to other privacy issues. Several emerging attacks are proven to threaten FL networks. Adversarial clients might be able to change the model's prediction or gain information about other clients by actively changing its own model parameters or synthesizing fake input data.

This paper analyzes a group of attack scenarios called 'adversarial attacks.' A malicious client aims to cheat the model by adding subtle noise to its examples. The noise is so minimal that the change in the image is imperceptible for a human observer. However, it can cause the model to misclassify it. Adversarial attacks are the main source of vulnerability to the deployed FL models [4]–[7]ånd are designed to fool the models that are already trained passed their test phase and delivered for clinical use.

The FL environment is leveraged to propose a stronger attack and show that it enables adversarial clients to manipulate other benign clients to a high degree. We also show that proposed defense methods such as differential privacy still have limited effect against this attack. In the following sections, we introduce adversarial attacks, sources of vulnerability, and our method.

A. Adversarial attacks

Adversarial attacks are caused by the linear characteristics of high-dimensional data. [8]. In FL setting, these attacks are also known as evasion attacks. [4], [6], [9], [10], and are defined as if a malicious client tries to manipulate other clients with fake data.

At the same time, research on defense methods is also ongoing. Some studies tried to incorporate differential privacy (DP) as a protection measure in FL setting. Differential privacy was initially proposed to protect against privacy attacks, but it has also been applied against adversarial attacks. [4], [11], [12]. Despite the ongoing research, there is no universal defense method against these attacks.

In this paper, adversarial attacks are investigated in the scope of FL and MIA systems. Distinct features of FL environments and MI make their combination highly vulnerable to adversar-

This paper is being prepared with IEEE standards This work was funded in part by NWO under project AMICUS
Erfan Darzidehkalani is with UMCG (e-mail: e.darzidehkalani@umcg.nl).

ial attacks. In the following paragraphs, unique vulnerabilities of such setting are discussed. [13]–[15]

Federated learning and adversarial attacks: FL characteristics are known to come on top of the attacks in the centralized setting, as a result, clients participating in an FL environment still keep their traditional ML vulnerabilities, and are exposed to more threats originated from FL network. [16], [17] The unique vulnerabilities of FL are :

(i) *More data:* Each model received in every round has updated information about other clients. Adversaries can exploit this information to prepare stronger attacks. [18]–[21]

(ii) *Open training:* In FL, the training process involves a lot of participants. Among them, one adversary could act maliciously.

(iii) *Standardized data pipelines:* FAIR data stations and data standardization steps are inherent in many of FL networks in hospitals [22],Digital Imaging and Communications in Medicine (DICOM) is a globally recognized standard for medical data transfer and management,integrated into many federated pipelines. Standardization process will cause clients to have more similar local data distributions. [23]. Similarity in data leads to higher transferability between models. [24]

Medical images and adversarial attacks: Also, medical images have distinct features which makes the networks trained on them more vulnerable.

(i) *Feature representation:* Due to the inherent feature space of medical images, Medical images are more vulnerable to this sort of attack than natural images. As shown in previously published work, medical images have a narrower high-dimensional feature representation of images than natural images, causing trained networks to be over-parameterized. [25] Over-parameterized networks are inherently easier to fool. [26] As a result, deep learning networks trained on medical images are easier to manipulate than ones trained on natural images. [25] Several works have confirmed this by manipulating Funduscupy, Chest X-Ray and Dermoscopy data and succesfully changing models prediction. [27].

(ii) *Unique texture:*Medical images have limited texture diversity, and small texture perturbation in medical images can confuse classifiers to a high degree. [25] This feature can be of advantage to the adversary's attacks, because they can perturb the texture in irrelevant areas and fool the classifier without manipulating the more important parts, e.g., the tumor area. [25]. Which has, for example, been shown to effectively confuse tumor detection classifiers [28]

B. Transferability factors

Despite defense not always being an option, DL systems might be able to gain insight into their points of vulnerability. Some parameters and deployment settings might increase the chances of a successful attack.

Gaining such perception can be directly imported in security analysis so that the IT administrators know where the highest threat can be from, and they do not require a brute-force simulation of all the scenarios and parameter values, which is practically hard and in most cases impossible.

The research is ongoing on attack transferability. [29]–[36]. In the MIA domain, transferability analysis has been done in a centralized ML setting, and the results are found to be domain specific. Factors such as data disparity, perturbation degree, and pre-training are shown to be crucial in attack transferability. However, their extent and optimal values might vary to a large extent. This difference can be attributed to the characteristics of the target imaging domain, texture of images, and knowledge of adversary [25]. [37].

In an FL setting with a higher level of complexity, the findings in centralized setting have limited pertinence. [6] One research question is whether optimal attack settings in an FL environment are concordant with the existing literature obtained from centralized data experiments.

In this project, three potential factors that might influence attack transferability are evaluated. Namely, perturbation degree, attack step parameter and the domain of attack are evaluated and the impact degree of each factor is discussed.

C. Our attack

The adversarial client participating in an FL setup has access to the previous model updates. We investigate a scenario where the adversary improves its attack by using the gradient information from previous model updates. We introduce an intermediary noise tensor we call *Cross-round noise (CRN)* which utilizes previously received global model updates to generate noise. To initialize the next round and avoid models' parameter change, we regularize L_2 norm in each noise channel by its mean value.

This way can achieve better attack performance and improve transferability, with a much lower computation burden than the standard attack methods.

We summarize our contributions as follows:

- To the best of our knowledge, this is the first study investigating adversarial attacks on FL in the MIA in a differentially private setting.
- We introduce a new attack and show the superiority of our model compared to the popular models. We show that sometimes single-step calculations can outperform computationally expensive models.
- We discuss how some domain-specific parameters affect attack transferability and estimate their optimal values for different medical imaging tasks.

The rest of this paper is organized as follows: In Section III, adversarial attacks on MIA and FL systems are discussed. Section II introduces FL, differential privacy and adversarial attack models. In Sections IV,V, our proposed attack model and experimental settings are discussed. Finally, Sections VI, VII provide results and discuss their implications.

II. Preliminaries

A. Federated Learning

Federated learning enables multiple data owners with private datasets to jointly train a global model based on local

models. The optimization problem could be formulated as $w = \sum_{i=1}^{N} p_i w_i$, $w_i = \arg\min_{w_i} (\mathcal{L}(\mathcal{D}_i; w))$ where N is the number of data owners, $\mathcal{L}(\mathcal{D}_i; w)$ is a loss function indicating global model parameters w of local datasets. The learning procedure is an iterative process containing local and global steps. Each data owner trains a model received from a global server on its local dataset in local iterations. The global server updates the global model by aggregating the updated local models. Then it sends it back to clients for the next round.

The global server selects a subset of clients at each global round and sends the most recent global model to them. Then each client performs local training over its dataset for a selected number of epochs. The updated local models are calculated on selected batches. Local optimization can be formulated as $w_i \leftarrow w - \eta \cdot \nabla \mathcal{L}(w; \mathcal{D}_i)$, where η is the learning rate. Several local iterations might be required to go over all the local data. Local training procedures can be done for several local epochs. (3) The global model can be updated based on the local models w_i and is shared for aggregation: $w = \sum_{i=1}^{N} p_i w_i$, to update the global model for the next FL round.

B. Differential Privacy

Differential privacy (DP) requires FL parties to ensure an attacker can not distinguish data records. For multi-party systems $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{R}$ mapping from domain $\mathcal{X}$ to target domain $\mathcal{R}$, differential privacy (ϵ, δ)-DP defines a measure to evaluate performance of privacy preserving mechanisms. For two adjacent datasets $\mathcal{D}_i, \mathcal{D}_i'$. DP introduces a bounding parameter $\epsilon > 0$, representing the ratio of probabilities of two datasets bounded by performing the privacy preserving mechanism δ. It can be summarized by the following definition [38]:

Mapping $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{R}$ is (ϵ, δ)-DP, if for all measurable subsets of target domain $\mathcal{S} \subseteq \mathcal{R}$ and for any two adjacent datasets $\mathcal{D}_i, \mathcal{D}_i' \in \mathcal{X}$,

$$Pr[\mathcal{M}(\mathcal{D}_i) \in \mathcal{S}] \leq e^{\epsilon} Pr[\mathcal{M}(\mathcal{D}_i') \in \mathcal{S}] + \delta. \tag{1}$$

In FL setting, (ϵ, δ)-DP can be achieved by adding noise to the updated models. Global differential privacy is a privacy mechanism that imposes a double-sided (ϵ, δ)-DP requirement for both uplink and downlink channels [39]. From the uplink perspective, all clients $1 \leq i \leq N$, clip their updates $\|w_i\| \leq C$, where w_i denotes the updated weights from the i-th client before perturbation and C is the clipping threshold. To satisfy a (ϵ, δ)-DP requirement for the downlink channels, additional noise n_{i} is added by the server, so each client i receives a $\tilde{w_i}$ perturbed model [39].

C. Adversarial attacks

Introduced by [40], adversarial examples are feature space manipulations with respect to linear decision boundaries, [8] and are categorized based on the adversary's goal and knowledge.

Adversary's goal Adversarial attacks can be categorized based on the adversary's goal. *Untargetted attacks* aim to reduce the model performance, regardless of the class to which a test sample belongs. *Targeted attacks* force the model to output certain labels.

Adversary's knowledge Based on the adversary's knowledge, attack can be *white-box*, meaning that the adversary has complete knowledge about other clients' network architecture, gradients, and parameters. In such a setting, it can easily manipulate the model. White-box attacks have been extensively investigated in the literature. [41] [42]. In *black-box* heuristics, the adversary does not have access to the target model. In the black-box setting, however, it can interact with the model. Adversary can feed inputs and receive outputs of the target model and improve the attack by observing the model outputs. Some black box might have limited knowledge about design of the target model. [43]–[45]

D. Attack models

Numerous adversarial attacks have been proposed in the literature in order to fool deep neural networks and produce false predictions. Adversarial attacks work through injecting a guided imperceptible noise that fools the trained deep learning model. Popular attack methods are Projected Gradient Descent (PGD), Fast Gradient Sign Method (FGSM), and Basic Iterative Method (BIM). Each will be discussed in this section.

FGSM: Fast Gradient Sign Method (FGSM) [40]is a fast yet effective method which produces adversary images with one step of calculation. Assuming input x and its corresponding target t, FGSM calculates gradient of x with respect to the loss function $\partial \mathcal{L} / \partial x$.

$$\hat{x} = x + \epsilon \cdot sgn\big(\nabla_x \mathcal{L}(g(x; w)\big) \tag{2}$$

where epsilon ϵ is a hyper-parameter which determines adversarial noise level. $g(x; w)$ is the output of neural network with respect to the input x and parameter set w. $sgn(.)$ is the sign function. The result of sign function goes through a clipping function to impose a maximum bound in change to the perturbation $\epsilon \cdot sgn\big(\nabla_x \mathcal{L}(g(x; w)\big) \in [-1, 1]$.

BIM: Basic Iterative Method (BIM) is an extension of the FGSM method, proposed by Kurakin et al. [46]. BIM repeatedly performs the FGSM process, using a small step-size and $\hat{x}^1 = x$. BIM is stronger than FGSM and requires smaller perturbations.

PGD: Madry et al. proposed their own version of the BIM attack. In Projected Gradient Descent (PGD) [47] the attack starts with a uniform random initialization. The update formula for PGD attack can be written as:

$$\hat{x}^{t+1} = \Pi_{P_\epsilon(x)}\Big(\hat{x}^t + \alpha \cdot sgn\big(\nabla_x \mathcal{L}(g(\hat{x}^t; w), t)\big)\Big) \tag{3}$$

Where $\hat{x}^k$ is the perturbed data in k-th iteration, and $P_\epsilon(x)$ is the projected gradient descent function, which is done by first finding sign values and then projecting the result to a small neighborhood of the input x. These possible parameter spaces determine the set of PGD attack samples that an adversary can use. PGD is one of the strongest attacks and is a universal first-order adversarial method. Technically, PGD is similar to BIM

but with a small random initialization, aside from formulating the problem as a projected gradient. We report results with PGD. With equal iterations, BIM had the same results, or the difference was negligible.

III. Related work

The effect of adversarial attacks on MIA has been studied in several works. Modalities such as chest X-ray, MRI [11], [25], [27], [37] and CT scan [48] segmentation and classification of medical images are vulnerable to adversarial attacks [49] [11].

Several works worked to discover important parameters on attack transferability. [29]–[36]. Gradients of different samples in one batch [35], extent of data augmentation [29], variation in input gradients [36] are shown to be important in transferability. Zhang et al. analyzed the impact of transfer learning on black-box attack. [50]. Another work [34] utilized the transferability of examples to enhance robustness, although their method was proposed for white-box setting.

In MIA, perturbation degree has been shown previously as a less explored but highly deterministic parameter in attack setting, and might need visual tuning. [37]. Optimal values of standard black-box [37], and white-box [25] are domain-specific. Also iteration steps α, can be highly deterministic in centralized setting. [51] [52].

Methods such to detect the attack [25], [53], [54] or defend the models from adversaries [55]–[59] are also discussed in ML domain. Adding noise to the exchanged model with clipping model updates is effective in several forms of adversarial attacks. [4]. In norm-bound defense, the server enforces an upper-limit norm-bound. Several studies have investigated black-box PGD attacks in an FL environment with norm-bound situation. [18], [20] . In the clinical setting, [60] DP models are used in clinical EHR data [61] [62] and neuroimaging data [63] in multi-site setting.

However, the proposed defense methods have some limitations. Some of them are only usable under specific settings [64]. Some of the algorithms require too much computational power [56], [64], [65], or they lead to drastic model parameter change [34]. Also, recent studies have shown that the existing defense methods are ineffective if the adversary is aware of them [64].

IV. Enhanced adversarial attack

In this section we introduce our method to enhance the existing adversarial attacks in an FL setting.

A. Threat model

In our attack scenario, we assume that an FL client is malicious or is controlled by some malicious adversary. The adversarial client tries to fool the fully trained global model by generating adversarial samples similar to its local real data. We assume central server to be honest and trusted.

Goal of adversary: The adversary's goal is to manipulate the fully trained global model so that the global model has a higher classification error.

Knowledge of adversary: In a realistic scenario, the malicious party only can see its own data D_i, and knows the DNN architecture and the global model weights that it receives each round.

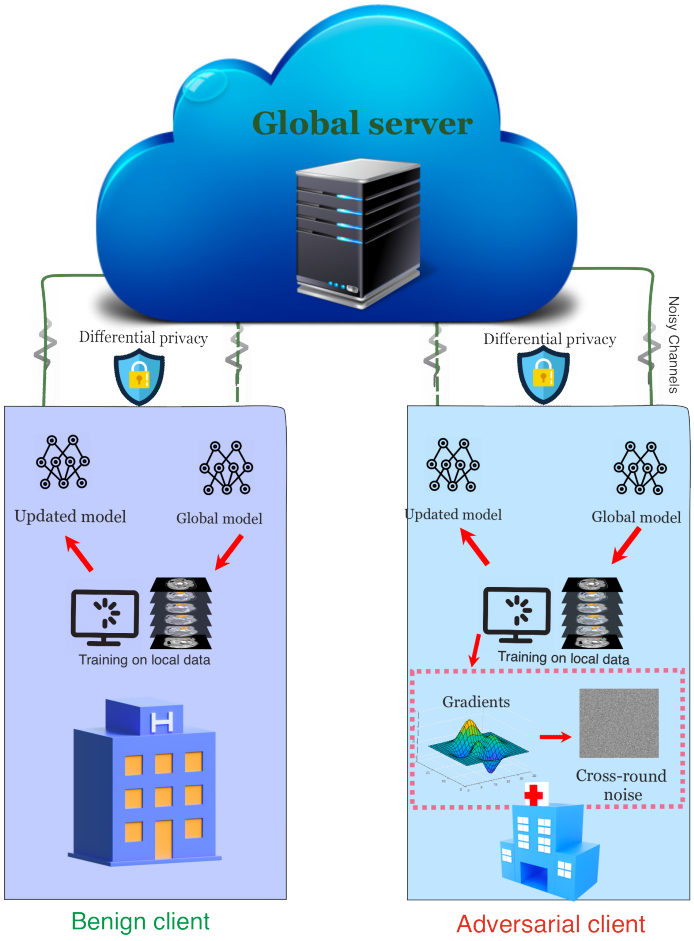


Fig. 2: A schema of our method, at each FL round, the adversary transforms the gradients into cross-round noise, while acting as a benign client and not tampering the training process

Capability of adversary: No special privilege or capability is assumed for the adversary. Similar to other clients, the adversary has control over its own training procedure and data, and can not manipulate other clients' data or the general learning process. (e.g., local computations, communication with central server and aggregation process, DNN architecture, and optimization functions). It can interact with the model, query the outputs, and calculate the gradients. The adversary uses its training data for FL training, and test data to perform adversarial attacks.

B. Cross-round noise

Since the adversarial samples are generated in the inference time, the adversary might not have the necessary compute power as it would have during training. Also, sometimes inference is being done on third-party applications on edge

Algorithm 1: Training procedure

Data: T, β, $w^{(0)}$, μ, ϵ and δ

Initializing parameters: $t = 1$ and $w_i^{(0)} = w^{(0)}$ and $\delta^{(t)} = 0\ \forall i, t$

while $t \leq T$ **do**

  Local training:

  while $\mathcal{C}_i \in \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_N\}$ **do**

    Clients update their models $w_i^{(t)}$ as

    $w_i^{(t)} = \arg\min_{w_i} (\mathcal{L}_i(w_i))$

    Clients clip parameters

    $w_i^{(t)} = w_i^{(t)} / \max\left(1, \frac{\|w_i^{(t)}\|}{C}\right)$

    Clients add Gaussian noise

    $\widetilde{w}_i^{(t)} = w_i^{(t)} + n_i^{(t)}$

  Global update:

  Global server individual models $w^{(t)}$ as

  $w^{(t)} = \sum_{i=1}^{N} p_i \widetilde{w}_i^{(t)}$

  Global server adds Gaussian noise

  $\widetilde{w_i}^{(t)} = w^{(t)} + n_{\mathrm{i}}^{(t)}$

  if $t \geq \beta$ **then**

    Adversarial client $\mathcal{C}_m, \mathcal{X} \in \mathcal{D}_m$:

    Adversary performs gradient regularization

    $\nabla_{\mathcal{X}} \hat{\mathcal{L}} \leftarrow \nabla_{\mathcal{X}} \mathcal{L}(\delta^{(t-1)}, \widetilde{w_m}^{(t)})$

    Adversary projects the gradient

    $\delta^{(t)} \leftarrow \Pi_{P_\epsilon(0)}(\nabla_{\mathcal{X}} \hat{\mathcal{L}})$

  $t \leftarrow t + 1$

Result: $\widetilde{w_i}^{(T)}, \delta^T$

devices without GPUs. Therefore, it is crucial for the adversary to be able to prepare the attack samples with less computations. Iterative methods, such as PGD, require too much computational power to develop the samples. even more than normal training itself. [66] The problem could be worse for high-resolution or whole-slide images like in our pathology example. Also, PGD requires the target and adversary to be highly similar to ensure transferability. However, models being noisy causes high inter-client variability, as a protection measure.

Incorporating prior knowledge already available from the global model, the attack can be initialized from a proper baseline so the computation might be reduced.

An intermediary noise tensor *Cross-round noise* is introduced, which extracts features from received global and is calculated alongside local epochs. The noise gets updated and passed to the next FL round. A function is used for the adversary to produce noise based on L_2 regularized gradients at each FL training round. Then in the test phase, the noise is used to initialize the values for adversarial attack algorithms.

1) Gradient calculation: The noise is calculated by the gradients of model at each round. The adversary uses the loss:

$$\nabla_x \mathcal{L}(g(x + \delta^{(t)}; w^{(t)}), g(x; w^{(t)})) \tag{4}$$

To update the stored values for the noise. In which $g(.; w^t)$ refers to received global model in communication round t. The adversary iteratively updates the noise by maximizing the loss between the model output for the noisy and the clean test data. The global model parameters change for the test dataset being the same in each round. [67] Saving noise could be started after an arbitrary number of rounds (β) is passed and be calculated alongside local training with similar epochs. In this paper, we consider initializing from the last 10 (CRN10) and last 5 (CRN5) rounds.

2) L_2 regularization: Transferring gradient information between models might lead to drastic parameter change. [67]–[69]. Unlike some methods which manually reset the gradients, so they lose information [34], we regularize the change in CRN by subtracting the mean value in each channel of the input noise. We apply a function for L_2 regularization the gradients.

$$\nabla_{x_i} \hat{\mathcal{L}} = \nabla_{x_i} \mathcal{L} - \frac{1}{M} \sum_{j=1}^{M} \nabla_{x_{i,j}} \mathcal{L} \tag{5}$$

The resulting value is channel-wise regularized loss, where i indicates index of channel, and j refers to individual pixels. This results in a smaller L_2 norm of gradient values, which is a proven measure against explosion or outlier gradients. [70] [71].Then the final gradients are projected to a bounded L_∞=ϵ norm around zero.

C. Inference phase attack

Attacks are performed after training is done. FGSM and BIM use zero initialization, and PGD uses random initialization, then they compute the adversarial example according to Equations 2 and 3.

Cross-round noise is added to the input test data as an initial point for these methods,

V. Experiment

A. Datasets

In our experiments, the samples are non-independent and identically distributed (non-IID) across clients. The datasets used in our experiment are combination of multiple datasets, and images are This way, the data assigned to each client differs from other clients.

To allocate samples in non-IID manner, datasets were split into non-overlapping chunks, each chunk representing a tiny portion of the general distribution. Each client samples are taken only from specific chunks.

Brain cancer classification: Dataset for detecting brain cancer is downloaded from Kaggle. It contains brain MRI images categorized into four classes: Three types of brain tumor, and one healthy no Tumor class. For Meningioma and Glioma detection, 1437 and 1426 samples were used, respectively, and they were split as train/ test data. Transforms used are rotation, Flipping, and normalizing. Images were resized to 100×100.

Histopathologic cancer detection: The samples are metastasic tissue images of lymph node cancer. Images are in size

,96 $\times$ 96, and the task is to classify the tissue samples. Each positive sample has a metastasic region, which is located in 32×32 neighborhood of each image sample. They were categorized into cancer and non-cancer classes. 2150 samples were chosen and randomly assigned to clients, and each split 62% for train and 38% for the test. We used horizontal flipping, and images were normalized.

B. Network architectures

Deep learning models: The deep learning model used is a Convolutional Neural Network (CNN) with six layers of convolution stacked to 5 fully connected layers. The activation function used is ReLU. And dropout parameter (0.25) Models are all trained and converged before implementing adversarial attacks. For each dataset, a classifier is trained with Cross-Entropy loss and with an SGD optimizer.

Federated setting: Three clients are defined for the FL setup. The data is assigned randomly, and the clients have non-IID data distribution. The FedAVG method is used to aggregate the models. The aggregation is weighed based on the length of the training dataset. Each client is trained for 20 epochs at the communication round. The total FL rounds are 50.

C. Attack setting

The role of adversarial client is assigned in turn to each of the clients. To perform the attack, each of the clients is therefore assigned 100 data samples per use case, and the results are averaged at the end. We use four metrics, Clean accuracy is defined as the performance of models on uncorrupted test images, *Attack Success Rate (ASR)*; how much an adversary can change the predicted labels produced by each model. We refer to the averaged ASR among all the clients and all of their test data as *Average Attack Success Rate (AASR)*.

Average Error Transferability Rate (AETR) is also defined to evaluate transferability. An adversary might perform an attack by only selecting the samples that successfully fooled its own model. AETR measures how successfully these samples changed the target models' correct predictions. So unlike ASR, this metric doesn't count the examples that were initially misclassified by benign clients' model. [57]

Here, we expand the previous findings to the FL settings and discuss whether change in ϵ can lead to transferability.

We perform three analyses to find important factors in attack success:

- We investigate dependency of attacker on ϵ. By visual inspection and ASR evaluation, and as a comparison to previously found optimal known values in centralized setting.
- We compare efficiency of models, by discussing their attack preparation time and their ASR.
- We also see how attack step α can determine ASR.

VI. Results

A. Efficiency analysis

The models are trained in an FL enviroment. Average test results, or clean accuracy shows the average performance of clients on unperturbed test data. Here we compare the computational complexity of iterative algorithms. The comparison is based on the time required to train one batch of adversarial examples, with $\epsilon = 0.05$ and $\alpha = 0.001$. As table I shows, the higher iteration rounds take more computation and lead to higher AASR.

Also, CRN does not require additional computation. Similar to single-step methods, computational load is small but increases on high-resolution images. As an example, the pathology dataset requires more calculation than MRI data. However, the overall time is far less than iterative models, and also it could enhance the transferability much better. The effect is consistent among different datasets.

TABLE I: Comparison of iterative models and their computational efficiency, on performing computations on one batch of data. ACC shows average client performance for unperturbed test data. AASR is average ASR on all clients.

Dataset	Attack type	ACC	AASR	time (sec)
Meningioma	**PGD-1+CRN10**	84.12%	**63.92%**	**0.217**
	PGD-20		27.52%	3.423
	PGD-40		32.99%	6.794
Pathology	**PGD-1+CRN10**	77.01%	**90.92%**	**0.354**
	PGD-20		60.98%	5.464
	PGD-40		82.27%	10.843
Glioma	**PGD-1+CRN10**	61.84%	**70.55%**	**0.216**
	PGD-20		51.83%	3.420
	PGD-40		63.77%	6.793

B. Effect perturbation degree

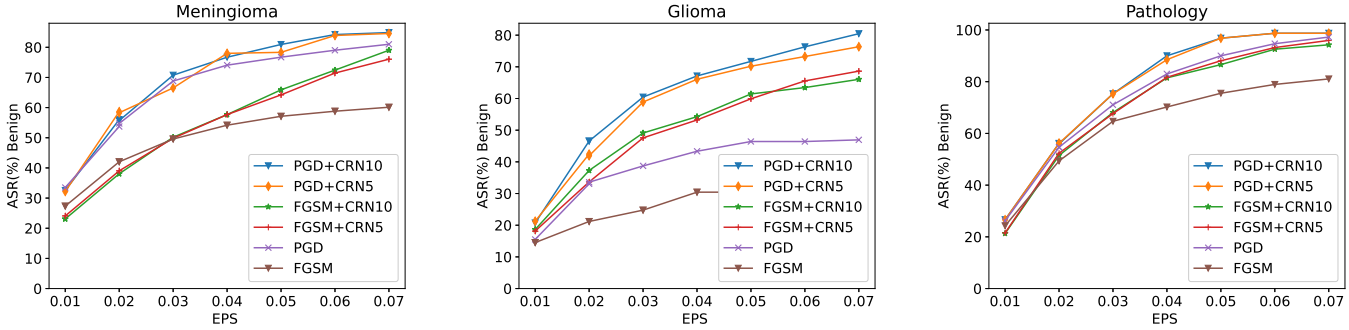
We performed attack scenarios to evaluate the effect of perturbation degree ϵ on attack success on the model. We compared the baseline attack to CRN- enabled attacks. ASR is calculated on benign and adversary clients and is shown in Fig 3.

More successful attacks on adversaries led to higher transferability to benign clients. Low values of ϵ can decrease transferability to a large extent. ϵ = 0.01 has poor performance on all the clients. Generally, Higher ϵ values lead to higher ASR in all scenarios. In pathology images, Iterative models can reach perfect accuracy on the adversarial client and have very high transferability.

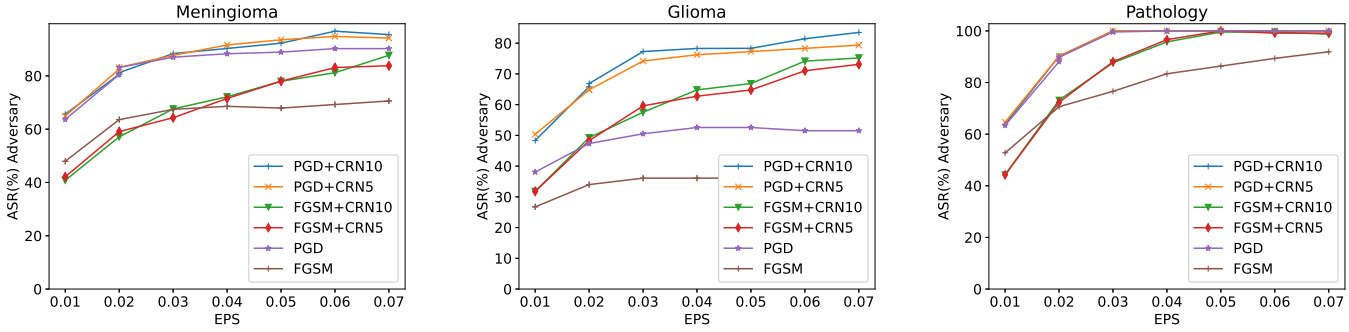
CRN has a positive effect on transferability in all scenarios. Also, CRN-enabled models are more dependent on ϵ; their extent can vary depending on the dataset, attack method, and value of ϵ. Tuning ϵ might be tricky since performance and imperceptibility should be considered together.High ϵ values mean more distortion in the image. We performed an FGSM attack with varying ϵ values to evaluate the perturbation effect visually. Fig 5 shows the results of perturbed images with different perturbation degrees and their comparison with the unperturbed image. Note that perturbation values below $\epsilon = 0.05$ cause limited distortion. Iterative models can reach high ASR very quickly with increasing ϵ.

C. Effect perturbation step

Another critical parameter is the step of change in iterative methods (α), which is shown to be highly deterministic in

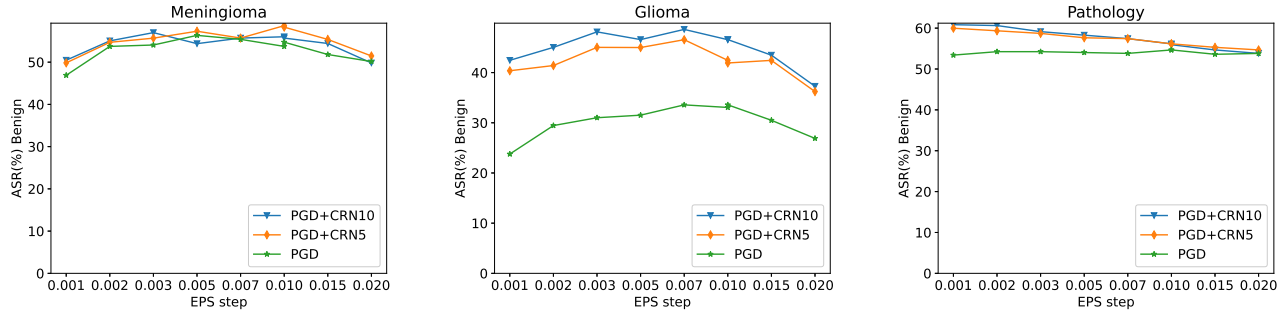


(a) Effect of ϵ on Average ASR on benign clients

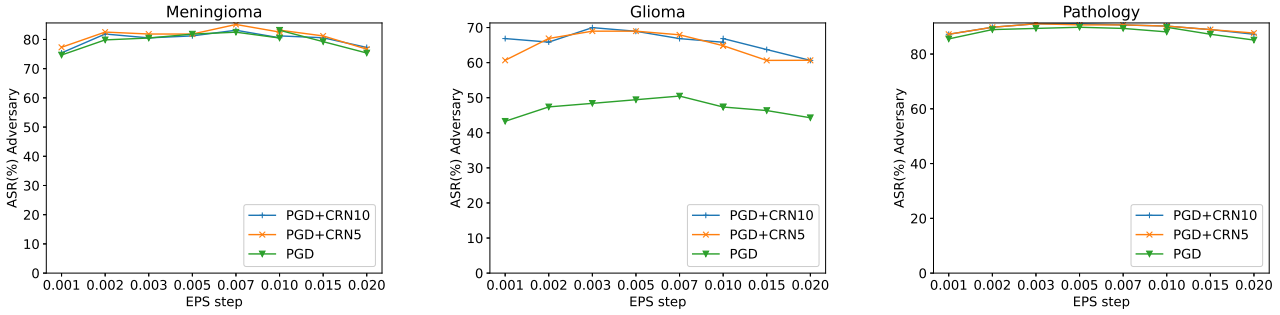


(b) Effect of ϵ on ASR on the advesarial client

Fig. 3: Effect of Error perturbation degree ϵ on attack transferability. FGSM and PGD attacks with and without CRN initialization were performed.ASR is calculated on benign and adversary clients. The higher ASR on benign clients shows higher transferability



(a) Average ASR on benign clients



(b) Average ASR on advesarial client

Fig. 4: Effect of Error perturbation step α on attack transferability. PGD attack with and without CRN initialization was performed. ASR is calculated on benign and adversary clients.The higher ASR on benign clients shows higher transferability

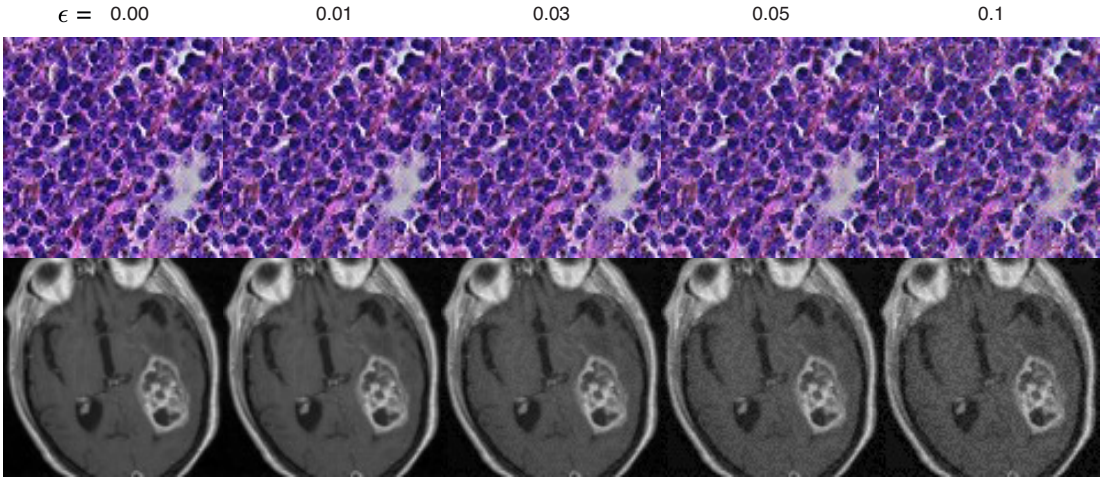


Fig. 5: From (a) to (f), normal tumor image, and petrurbed images with FGSM attack, with perturbation parameters: ϵ = 0.01, 0.03, 0.05, 0.10, respectively

specific settings but is not yet investigated in MIA domain. To evaluate how α can affect transferability, we examined the clients with $\epsilon = 0.03$ in various steps. Generally, there is an optimal middle range the same as ϵ. A step size of 0.007 led to the highest transferability in MRI images. Larger steps led to a sharp decrease in transferability. For pathology images, midspans, about 0.007 led to the best ASR on the adversary, and higher values decreased both ASR on adversary and transferability. However, smaller steps also had high ASR values.

D. Error tranferability

Another measure is the average error transfer rate (AETR). Although PGD outperforms FGSM in ASR, in AETR, the results are close. PGD is trained for 40 iterations, and the models are compared with and without CRN. Both FGSM and PGD having high AETR suggest that although FGSM might not be able to produce high-quality samples in general, the good examples which can fool the adversary are also highly transferable to benign clients. And can be better than PGD samples. CRN generally increases the AETR.

VII. Discussion

Our results confirm prior research about vulnerability of FL networks [4], [6], [16]–[24] and MIA systems [25], [27], [28], [37] to adversarial attacks.

We also found that differential privacy might have partial effect on attacker's success. In fact, DP caused the lower ASR in benign than the adversary clients. Although some research supports DP, our findings suggest that it is an option but not very reliable. [4]. Also, the accuracy compromise coming with noise should be considered. [72] [73].

In our experiments the attacker did not tamper with the training procedure. This is unlike backdoor or poisoning attacks. [7] [5], where the attacker performs malicious activities during training, and opens the way for detection methods. Also, CRN initialization led to faster attacks with higher transferability. This section discusses our findings, how they could be important in the MI domain, and what to consider when setting up FL MI infrastructure.

A. Transferability

Our experiments aimed to asses some factors in transferability. However, other than investigated factors, we found three other parameters might highly predict transferability.

Benign/Adversary correlation: Our results show that ASR on benign and adversary clients are highly correlated. Hence, for the adversary to ensure it can fool other clients in a black-box setting, it needs to obtain good results on its own data and model. So tuning the hyperparameters, data preprocessing can be all effective as long as they improve ASR on adversary.

AETR/ASR difference:, Attacks have higher AETR than ASR, which enables the adversary to subsample examples to achieve higher transferability.

Attack domain: We observed that transferability is specific within imaging domains, which is consistent with prior findings. [37] Attacks on Pathology images were consistently more successful than MRI images.

Perturbation parameter Level of ϵ can substantially change transferability. Increasing ϵ does not improve ASR from a certain point, but adding CRN can increase the dependency. The best perturbation range in our experiments is around 0.03-0.05. These results suggest that conclusions in a centralized setting might not apply to an FL environment. And it shows different transferability behavior or parameter dependence.

Step parameter Our findings show that common α values have a negligible effect on the final results, aside from too small or too large values. The reason could be that α similar to ϵ bounds the change in each iteration, but generally, its values are one order of magnitude less than ϵ.

B. Efficiency

Our results show that computational complexity has a linear dependence on the number of iterations. However, the increase

TABLE II: Result of Average Error Transfer Rate (AETR) for FGSM and PGD methods, with and without CRN initalization

Dataset	FGSM			PGD		
	Baseline	CRN5	CRN10	Baeline	CRN5	CRN10
Meningioma	84.80%	**86.31%**	84.02%	83.73%	**84.25%**	84.16%
Glioma	82.32%	87.06	**89.98**	74.89%	**83.64%**	82.32%
Pathology	**90.67%**	84.81%	85.98%	79.86%	86.79%	**86.93%**

in transferability can vary. PGD with random intialization requires high computational power which might be burdensome for High-resolution or large-batch training. Also, if PGD performs poorly initially, having higher iterations does not help much. Initialization can be influential in computational efficiency. Using CRN leads to much fewer computational requirements. Single-step attacks can be $20 \sim 30\times$ less expensive than iterative models with 40 iterations. Also, FGSM with CRN can outperform PGD with random initialization.
With less GPU requirements, adversaries might be capable of attacks using high-volumes of data or with multiple devices by using CRN.
Standard ways of simulating attacks, like adaptive models and sanity checks, [74] might be unfeasible in their traditional way. An adversary might have more knowledge than what is assumed in evaluation models, hence using more efficient methods or attacks with unexpected devices.

C. Practical implications and suggestions

In the following paragraphs, we discuss the motivations that cause adversarial attacks .

The large healthcare economy causes more benefit for malicious behavior, as there are existing reports of pervasive fraud in healthcare. Medical record manipulation for financial purposes [25] fake trial reports [75] in radiology [76] and pathology [77] are existing practices, and already made billions of dollars profit [78] for fraudsters. Also, advanced fabrication algorithms and hard to detect methods are always intriguing for fraudsters. Existing Manipulations like altering the visual features of images, or using photo editing software, [76] can result in images of benign subjects classified as malignant. [79] [80], but with visual distortion. In contrast, adversarial attacks are imperceptible, do not require manual intervention, and have high transferability. That gives a significant incentive for potential fraudsters to utilize adversarial attacks for high potential revenue.
For example, consider a hypothetical case in which insurance companies utilize AI-based systems to approve a disease and get refund. A malicious person could add adversarial noise to images to manipulate insurance companies. Or consider if a doctor tries to bill insurance companies by reporting falsified surgical procedures. He can add adversarial noise to MRI images to forge evidence for his claimed diagnosis.
These vulnerabilities have some implications for clinical decision-makers, stakeholders, and insurance companies, who consider deploying FL or AI pipelines , which will be discussed in the following.

(i) First, manipulation and fraud is pervasive in healthcare and is reinforced by financial and personal incentives. This should be taken into consideration. [27]
(ii) Second, having a secure infrastructure in training phase does not guarantee safety in the deployment phase. So the difference between these two should be clear, and redundant ways might be considered to incorporate AI in MIA pipelines.
(iii) Third, size and level of trust should be regarded as imporant parameters in collaborative/FL networks. Smaller networks with trusted parties have lower probability of having potential adversary.
(iv) Fourth, Not enforcing same data pre-processing pipelines and using diverse data hinders the adversary. However, having disparate data comes with a compromise: it affects performance or even convergence of FL network. [81]. So healthcare managers should consider whether it is worth sacrificing performance for security.

For developers of MIA systems these findings can have consider four things :

(i) Developers can help end-users and clinicians by providing them more information other than model outputs. For example, adding explainable system reports helps the clinicians evaluate the legitimacy of predictions.
(ii) There are some suggestions by [74] to test the adversarial robustness in centralized setting. Notably, performing adaptive attacks, with standard scenarios to disclose vulnerabilities. However, those ways are not recommended in an FL setting. Developers should consider scenarios where adversary traces the global model, and has more knowledge, or attacks with unexpected devices.
(iii) We saw that attack setting, and values of parameters like ϵ, can affect transferability. Adversary might enhance its attack by choosing the right set of parameters and setting. However, There is no universal setting that guarantees a better performance. Hence, developers are encouraged to use the results from prior research to find the optimal parameter set for their desired domain, or they should brute-force search to find upper-bounds of vulnerability. [74]
(iv) Despite not having a universal defense method, there are limited case-specific defense models shown to work on some datasets, [78] so developers might consider looking into potnetial defense method for their usecase.

Future lines of research could be to improve existing defense methods, [18], [20], [34], [56], [59]–[65] , towards a universal defense algorithm. Another line of research could be how to utilize distributed nature of FL networks to protect all participating clients, or if collaboration can be used to enhance the current defense methods.

VIII. Conclusion

This paper investigated adversarial attacks on federated MIA systems and discussed the crucial parameters on their transferability. It also proposed a new method that leverages the federated setting to improve the attack success rate and reduce the computation burden. Our results indicate that tuning the parameters can substantially improve transferability. It was also observed that adequately using noise from previous model updates can effectively improve computational load and has the potential to be well integrated into the existing attack methods.

We hope this research could benefit healthcare institutions and hospitals considering bringing in AI or joining FL networks to be warier of the threat of adversarial examples. For medical institutions and managers of hospitals and insurance companies, this research suggests that they should pay extra attention to the AI pipelines and FL deployments. Our results indicate that security experts should re-asses standard vulnerability analysis and consider less common scenarios where a capable adversary proposes fast and efficient attacks.

Acknowledgement

This research is supported by KWF Kankerbestrijding and the Netherlands Organisation for Scientific Research (NWO) Domain AES, as part of their joint strategic research programme: Technology for Oncology IL. The collaboration project is co-funded by the PPP allowance made available by Health Holland, Top Sector Life Sciences Health, to stimulate public-private partnerships.

References

[1] M. J. Sheller, B. Edwards, G. A. Reina, J. Martin, S. Pati, A. Kotrotsou, M. Milchenko, W. Xu, D. Marcus, R. R. Colen, *et al.*, "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.

[2] I. Dayan, H. R. Roth, A. Zhong, A. Harouni, A. Gentili, A. Z. Abidin, A. Liu, A. B. Costa, B. J. Wood, C.-S. Tsai, *et al.*, "Federated learning for predicting clinical outcomes in patients with covid-19," *Nature medicine*, vol. 27, no. 10, pp. 1735–1743, 2021.

[3] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, *et al.*, "The future of digital health with federated learning," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–7, 2020.

[4] N. Bouacida and P. Mohapatra, "Vulnerabilities in federated learning," *IEEE Access*, vol. 9, pp. 63 229–63 249, 2021.

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

[6] G. Costa, F. Pinelli, S. Soderi, and G. Tolomei, "Covert channel attack to federated learning systems," *arXiv preprint arXiv:2104.10561*, 2021.

[7] L. Lyu, H. Yu, X. Ma, L. Sun, J. Zhao, Q. Yang, and P. S. Yu, "Privacy and robustness in federated learning: Attacks and defenses," *arXiv preprint arXiv:2012.06337*, 2020.

[8] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[9] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2013, pp. 387–402.

[10] M. A. Ayub, W. A. Johnson, D. A. Talbert, and A. Siraj, "Model evasion attack on intrusion detection systems using adversarial machine learning," in *2020 54th Annual Conference on Information Sciences and Systems (CISS)*. IEEE, 2020, pp. 1–6.

[11] S. Asgari Taghanaki, A. Das, and G. Hamarneh, "Vulnerability analysis of chest x-ray image classification against adversarial attacks," in *Understanding and interpreting machine learning in medical image computing applications*. Springer, 2018, pp. 87–94.

[12] M. Wu, X. Zhang, J. Ding, H. Nguyen, R. Yu, M. Pan, and S. T. Wong, "Evaluation of inference attack models for deep learning on medical data," *arXiv preprint arXiv:2011.00177*, 2020.

[13] 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.

[14] H. Chen, C. Fu, J. Zhao, and F. Koushanfar, "Deepinspect: A black-box trojan detection and mitigation framework for deep neural networks." in *IJCAI*, vol. 2, no. 5, 2019, p. 8.

[15] Y. Ji, X. Zhang, S. Ji, X. Luo, and T. Wang, "Model-reuse attacks on deep learning systems," in *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, 2018, pp. 349–363.

[16] M. Goldblum, D. Tsipras, C. Xie, X. Chen, A. Schwarzschild, D. Song, A. Madry, B. Li, and T. Goldstein, "Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses," *arXiv preprint arXiv:2012.10544*, 2020.

[17] P. Liu, X. Xu, and W. Wang, "Threats, attacks and defenses to federated learning: issues, taxonomy and perspectives," *Cybersecurity*, vol. 5, no. 1, pp. 1–19, 2022.

[18] Z. Sun, P. Kairouz, A. T. Suresh, and H. B. McMahan, "Can you really backdoor federated learning?" *arXiv preprint arXiv:1911.07963*, 2019.

[19] M. Fang, X. Cao, J. Jia, and N. Gong, "Local model poisoning attacks to {Byzantine-Robust} federated learning," in *29th USENIX Security Symposium (USENIX Security 20)*, 2020, pp. 1605–1622.

[20] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J.-y. Sohn, K. Lee, and D. Papailiopoulos, "Attack of the tails: Yes, you really can backdoor federated learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 070–16 084, 2020.

[21] M. Song, Z. Wang, Z. Zhang, Y. Song, Q. Wang, J. Ren, and H. Qi, "Analyzing user-level privacy attack against federated learning," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 10, pp. 2430–2444, 2020.

[22] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[23] P. van Ooijen, E. Darzidehkalani, and A. Dekker, "Ai technical considerations: Data storage, cloud usage and ai pipeline," *arXiv preprint arXiv:2201.08356*, 2022.

[24] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records," *Scientific reports*, vol. 6, no. 1, pp. 1–10, 2016.

[25] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, and F. Lu, "Understanding adversarial attacks on deep learning based medical image analysis systems," *Pattern Recognition*, vol. 110, p. 107332, 2021.

[26] S. Ye, K. Xu, S. Liu, H. Cheng, J.-H. Lambrechts, H. Zhang, A. Zhou, K. Ma, Y. Wang, and X. Lin, "Adversarial robustness vs. model compression, or both?" in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[27] 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.

[28] D. Gupta and B. Pal, "Vulnerability analysis and robust training with additive noise for fgsm attack on transfer learning-based brain tumor detection from mri," in *Proceedings of the International Conference on Big Data, IoT, and Machine Learning*. Springer, 2022, pp. 103–114.

[29] C. Gao and W. Wu, "Boosting the transferability of adversarial examples with more efficient data augmentation," in *Journal of Physics: Conference Series*, vol. 2189, no. 1. IOP Publishing, 2022, p. 012025.

[30] I. A. Elaalami, S. O. Olatunji, and R. M. Zagrouba, "At-bod: An adversarial attack on fool dnn-based blackbox object detection models," *Applied Sciences*, vol. 12, no. 4, p. 2003, 2022.

[31] J. Dai and L. Shu, "Fast-uap: An algorithm for expediting universal adversarial perturbation generation using the orientations of perturbation vectors," *Neurocomputing*, vol. 422, pp. 109–117, 2021.

[32] M. Duan, K. Li, J. Deng, B. Xiao, and Q. Tian, "A novel multi-sample generation method for adversarial attacks," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 4, pp. 1–21, 2022.

[33] X. Du, J. Yu, Z. Yi, S. Li, J. Ma, Y. Tan, and Q. Wu, "A hybrid adversarial attack for different application scenarios," *Applied Sciences*, vol. 10, no. 10, p. 3559, 2020.

[34] H. Zheng, Z. Zhang, J. Gu, H. Lee, and A. Prakash, "Efficient adversarial training with transferable adversarial examples," in *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1181–1190.

[35] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial training for free!" *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[36] H. Qiu, Y. Du, and T. Lu, "The framework of cross-domain and model adversarial attack against deepfake," *Future Internet*, vol. 14, no. 2, p. 46, 2022.

[37] G. Bortsova, C. González-Gonzalo, S. C. Wetstein, F. Dubost, I. Katramados, L. Hogeweg, B. Liefers, B. van Ginneken, J. P. Pluim, M. Veta, *et al.*, "Adversarial attack vulnerability of medical image analysis systems: Unexplored factors," *Medical Image Analysis*, vol. 73, p. 102141, 2021.

[38] C. Dwork, A. Roth, *et al.*, "The algorithmic foundations of differential privacy." *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3-4, pp. 211–407, 2014.

[39] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020.

[40] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[41] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," *arXiv preprint arXiv:1705.07204*, 2017.

[42] H. Xu, Y. Ma, H.-C. Liu, D. Deb, H. Liu, J.-L. Tang, and A. K. Jain, "Adversarial attacks and defenses in images, graphs and text: A review," *International Journal of Automation and Computing*, vol. 17, no. 2, pp. 151–178, 2020.

[43] Z. Yue, Z. He, H. Zeng, and J. McAuley, "Black-box attacks on sequential recommenders via data-free model extraction," in *Fifteenth ACM Conference on Recommender Systems*, 2021, pp. 44–54.

[44] Y. Xiang, Z. Chen, Z. Chen, Z. Fang, H. Hao, J. Chen, Y. Liu, Z. Wu, Q. Xuan, and X. Yang, "Open dnn box by power side-channel attack," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 67, no. 11, pp. 2717–2721, 2020.

[45] S. Cheng, Y. Dong, T. Pang, H. Su, and J. Zhu, "Improving black-box adversarial attacks with a transfer-based prior," *Advances in neural information processing systems*, vol. 32, 2019.

[46] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 99–112.

[47] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[48] F. Navarro, C. Watanabe, S. Shit, A. Sekuboyina, J. C. Peeken, S. E. Combs, and B. H. Menze, "Evaluating the robustness of self-supervised learning in medical imaging," *arXiv preprint arXiv:2105.06986*, 2021.

[49] U. Ozbulak, A. Van Messem, and W. D. Neve, "Impact of adversarial examples on deep learning models for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 300–308.

[50] Y. Zhang, Y. Song, J. Liang, K. Bai, and Q. Yang, "Two sides of the same coin: White-box and black-box attacks for transfer learning," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 2989–2997.

[51] Y. Tashiro, Y. Song, and S. Ermon, "Diversity can be transferred: Output diversification for white-and black-box attacks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4536–4548, 2020.

[52] Q.-Z. Cai, M. Du, C. Liu, and D. Song, "Curriculum adversarial training," *arXiv preprint arXiv:1805.04807*, 2018.

[53] M. Yin, S. Li, Z. Cai, C. Song, M. S. Asif, A. K. Roy-Chowdhury, and S. V. Krishnamurthy, "Exploiting multi-object relationships for detecting adversarial attacks in complex scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7858–7867.

[54] N. Drenkow, N. Fendley, and P. Burlina, "Attack agnostic detection of adversarial examples via random subspace analysis," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 472–482.

[55] W.-Y. Lin, F. Sheikholeslami, L. Rice, J. Z. Kolter, *et al.*, "Certified robustness against adversarial patch attacks via randomized cropping," in *ICML 2021 Workshop on Adversarial Machine Learning*, 2021.

[56] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 9, pp. 2805–2824, 2019.

[57] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017, pp. 506–519.

[58] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[59] Y. Li, M. Cheng, C.-J. Hsieh, and T. C. Lee, "A review of adversarial attack and defense for classification methods," *The American Statistician*, pp. 1–17, 2022.

[60] R. Shao, H. He, H. Liu, and D. Liu, "Stochastic channel-based federated learning for medical data privacy preserving," *arXiv preprint arXiv:1910.11160*, 2019.

[61] Z. Li, K. Roberts, X. Jiang, and Q. Long, "Distributed learning from multiple ehr databases: contextual embedding models for medical events," *Journal of biomedical informatics*, vol. 92, p. 103138, 2019.

[62] J. Ma, Q. Zhang, J. Lou, J. C. Ho, L. Xiong, and X. Jiang, "Privacy-preserving tensor factorization for collaborative health data analysis," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 1291–1300.

[63] X. Li, Y. Gu, N. Dvornek, L. H. Staib, P. Ventola, and J. S. Duncan, "Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: Abide results," *Medical Image Analysis*, vol. 65, p. 101765, 2020.

[64] M. Yin, S. Li, C. Song, M. S. Asif, A. K. Roy-Chowdhury, and S. V. Krishnamurthy, "Adc: Adversarial attacks against object detection that evade context consistency checks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 3278–3287.

[65] J. Uesato, B. O'donoghue, P. Kohli, and A. Oord, "Adversarial risk and the dangers of evaluating against weak attacks," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5025–5034.

[66] D. Zhang, T. Zhang, Y. Lu, Z. Zhu, and B. Dong, "You only propagate once: Accelerating adversarial training via maximal principle," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[67] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *International conference on machine learning*. PMLR, 2019, pp. 7472–7482.

[68] I. Pan, H. H. Thodberg, S. S. Halabi, J. Kalpathy-Cramer, and D. B. Larson, "Improving automated pediatric bone age estimation using ensembles of models from the 2017 rsna machine learning challenge," *Radiology: Artificial Intelligence*, vol. 1, no. 6, p. e190053, 2019.

[69] T. Dettmers, "8-bit approximations for parallelism in deep learning," *arXiv preprint arXiv:1511.04561*, 2015.

[70] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International conference on machine learning*. PMLR, 2013, pp. 1310–1318.

[71] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646–1654.

[72] C. L. Canonne, G. Kamath, and T. Steinke, "The discrete gaussian for differential privacy," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 676–15 688, 2020.

[73] Y. Wang, "Privacy-preserving average consensus via state decomposition," *IEEE Transactions on Automatic Control*, vol. 64, no. 11, pp. 4711–4716, 2019.

[74] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin, "On evaluating adversarial robustness," *arXiv preprint arXiv:1902.06705*, 2019.

[75] S. L. George and M. Buyse, "Data fraud in clinical trials," *Clinical investigation*, vol. 5, no. 2, p. 161, 2015.

[76] A. Chowdhry, K. Sircar, D. B. Popli, and A. Tandon, "Image manipulation: Fraudulence in digital dental records: Study and review," *Journal of forensic dental sciences*, vol. 6, no. 1, p. 31, 2014.

[77] S. Suvarna and M. Ansary, "Histopathology and the 'third great lie'. when is an image not a scientifically authentic image?" *Histopathology*, vol. 39, no. 5, pp. 441–446, 2001.

[78] A. Graese, A. Rozsa, and T. E. Boult, "Assessing threat of adversarial examples on deep neural networks," in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2016, pp. 69–74.

[79] T. Xia, A. Chartsias, and S. A. Tsaftaris, "Pseudo-healthy synthesis with pathology disentanglement and adversarial learning," *Medical Image Analysis*, vol. 64, p. 101719, 2020.

[80] L. Sun, J. Wang, Y. Huang, X. Ding, H. Greenspan, and J. Paisley, "An adversarial learning approach to medical image synthesis for lesion detection," *IEEE journal of biomedical and health informatics*, vol. 24, no. 8, pp. 2303–2314, 2020.
[81] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," *arXiv preprint arXiv:1907.02189*, 2019.