DISBELIEVE: Distance Between Client Models Is Very Essential for Effective Local Model Poisoning Attacks

Indu Joshi[1(✉)], Priyank Upadhya[1], Gaurav Kumar Nayak[2], Peter Schüffler[1], and Nassir Navab[1]

[1] Technical University of Munich, Boltzmannstraße 15, 85748, Garching Bei München, Germany
{indu.joshi,priyank.upadhya,peter.schueffler,nassir.navab}@tum.de
[2] Center for Research in Computer Vision, University of Central Florida, Orlando, FL 32816, USA
gauravkumar.nayak@ucf.edu

Abstract. Federated learning is a promising direction to tackle the privacy issues related to sharing patients' sensitive data. Often, federated systems in the medical image analysis domain assume that the participating local clients are *honest*. Several studies report mechanisms through which a set of malicious clients can be introduced that can poison the federated setup, hampering the performance of the global model. To overcome this, robust aggregation methods have been proposed that defend against those attacks. We observe that most of the state-of-the-art robust aggregation methods are heavily dependent on the distance between the parameters or gradients of malicious clients and benign clients, which makes them prone to local model poisoning attacks when the parameters or gradients of malicious and benign clients are close. Leveraging this, we introduce DISBELIEVE, a local model poisoning attack that creates malicious parameters or gradients such that their distance to benign clients' parameters or gradients is low respectively but at the same time their adverse effect on the global model's performance is high. Experiments on three publicly available medical image datasets demonstrate the efficacy of the proposed DISBELIEVE attack as it significantly lowers the performance of the state-of-the-art *robust aggregation* methods for medical image analysis. Furthermore, compared to state-of-the-art local model poisoning attacks, DISBELIEVE attack is also effective on natural images where we observe a severe drop in classification performance of the global model for multi-class classification on benchmark dataset CIFAR-10.

Keywords: Federated Learning · Model Poisoning Attacks · Deep Learning

I. Joshi and P. Upadhya—These authors contributed equally to this work.

M. E. Celebi et al. (Eds.): MICCAI 2023 Workshops, LNCS 14393, pp. 297–310, 2023.
https://doi.org/10.1007/978-3-031-47401-9_29

1 Introduction

The success of deep models for medical image analysis [13] greatly depends on sufficient training data availability. Strict privacy protocols and limited availability of time and resources pose challenges in collecting sizeable medical image datasets [12]. Although different medical institutions may be willing to collaborate, strict privacy protocols governing patients' information restrict data sharing. Federated learning (FL) offers a promising solution that allows different institutions to share information about these models without revealing personal information about the patients [6,18,20]. Federated Learning is a machine learning paradigm that learns a single shared global model by collaboratively learning from different local models on distributed systems without sharing the data.

A federated learning setup involves multiple clients and a global server [18]. The global server initializes the global model and sends the parameters back to the clients. The clients then train their local models on the data present locally. Once the local models are trained, the parameters are sent to the global model for aggregation. The global model then uses an *aggregation algorithm* to aggregate all the parameter updates and transmits the updated parameters back to the clients, and the cycle repeats until convergence. The federated learning setup allows the clients to preserve the privacy of their data. The success of a federated learning system is majorly dependent on the use of an aggregation algorithm. For example, *Federated Averaging* [18] is an aggregation algorithm in which all the parameters accumulated at the global server from different clients are averaged. However, not all clients would act truthfully in real-world scenarios, and there may be some *byzantine* clients. A client is said to be a byzantine client if it acts malicious intentionally due to the presence of an adversary or unintentionally due to faulty equipment or hardware issues [26]. Studies report that even a single byzantine worker can seriously threaten the FL systems [4].

A malicious byzantine worker with an adversary who knows the client's data and model parameters can induce *local poisoning attacks* to degrade the performance of the global model in an FL system. A local poisoning attack in an FL system is a process through which the training of the global model is adversely affected due to either data perturbation or perturbation in model parameters (or gradients) at the local client's side. These attacks are termed as *local data poisoning attacks* or *local model poisoning attacks*, respectively. Several studies indicate that state-of-the-art aggregation methods, for instance, using federated averaging in the presence of a byzantine client, will reduce the performance of the global server. Therefore, to defend against attacks by byzantine clients, the global server uses *robust aggregation algorithms* [25,26]. This research studies the efficacy of state-of-the-art robust aggregation methods for FL systems for medical image analysis and highlights their vulnerability to local model poisoning attacks. We observe that the state-of-the-art robust aggregation methods heavily rely on the distance between malicious and benign client model parameters (or gradients). We argue that some model poisoning attacks can exist when the parameters or gradients of malicious clients are close in Euclidean space to those

of benign clients that circumvent the existing state-of-the-art robust aggregation methods.

Research Contribution: We introduce the DISBELIEVE attack that demonstrates the limitation of state-of-the-art robust aggregation methods for FL on medical images in defending against local model poisoning attacks. The novelty of the proposed attack lies in the fact that it maximizes the objective loss function while ensuring that the Euclidean distance between the malicious parameters and benign parameters is kept marginal. As a result, the attacker can optimally reduce the global model's performance without being detected by the aggregation algorithms. Experiments on three publicly available datasets of different medical image modalities confirm the efficacy of DISBELIEVE attack in significantly reducing the classification performance of the global model (by up to 28%). We also benchmark two current state-of-the-art local model poisoning attack methods and demonstrate that the proposed DISBELIEVE attack is stronger, leading to higher performance degradation. Lastly, we demonstrate that DISBELIEVE attack also effectively works on natural images, as similar trends are reported on the CIFAR-10 dataset.

2 Related Work

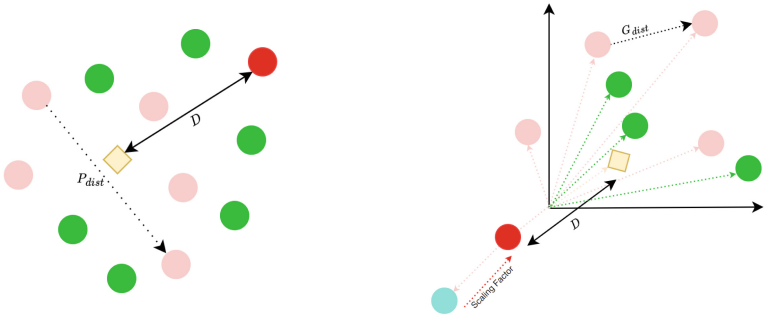
2.1 Robust Aggregation Algorithms

Robust aggregation algorithms are defense methods that prevent malicious clients from significantly affecting parameter updates and global model performance. KRUM [3] is among the earliest methods for robust aggregation and proposes that for each communication round, only one of the clients is selected as an honest participant, and updates from the other clients are discarded. The client that is chosen as honest is the one whose parameters are closer in Euclidean space to a chosen number of its neighbors. On the other hand, Trimmed Mean [26] assumes malicious clients to have extreme values of parameters and proposes to avoid malicious clients by selecting parameters around the median. Recently, the Distance-based Outlier Suppression (DOS) [1] algorithm was proposed to defend against byzantine attacks in FL systems for medical image analysis. DOS proposes to detect malicious clients using COPOD, a state-of-the-art outlier detection algorithm [15]. Subsequently, it assigns less weight to the parameters from those malicious clients. Specifically, it uses Euclidean and cosine distances between parameters from different clients and computes an outlier score for each client. Later, these scores are converted to weights by normalizing them using a softmax function. We note that all these state-of-the-art robust aggregation algorithms assume that malicious clients' parameters (or gradients) are significantly different from benign clients' parameters (or gradients). However, we hypothesize that an attack can be introduced such that parameters (or gradients) of malicious and benign clients are only marginally different, while it can still severely degrade the global model's performance.

2.2 Attacks in Federated Learning

There are various kinds of attacks in a federated learning paradigm, such as *inference attacks, reconstruction attacks, poisoning attacks* [5,11,16]. In inference attacks, the attacker can extract sensitive information about the training data from the learned features or parameters of the model, thus causing privacy issues. Reconstruction attacks, on the other hand, try to generate the training samples using the leaked model parameters [5]. GAN's [7] have successfully extracted private information about the client's data even when model parameters are unclear due to the use of differential privacy [9]. Poisoning attacks in a federated learning paradigm can be categorized as *data poisoning attacks* or *model poisoning attacks*. Both these attacks are designed to alter the behavior of the malicious client's model [17]. In data poisoning attacks, the attacker tries manipulating the training data by changing the ground truth labels or carefully poisoning the existing data [23]. In model poisoning attacks, the attacker aims to alter the model parameters or gradients before sending them to the global server [17].

In this research, we design a model poisoning attack that can bypass state-of-the-art robust aggregation algorithms such as DOS, Trimmed Mean, and KRUM. We evaluate the performance of existing state-of-the-art model poi-



(a) Model Poisoning Attack on Parameters **(b)** Model Poisoning Attack on Gradient

Fig. 1. Intuition behind our proposed local model poisoning attack: (a) Green nodes represent the parameters of benign clients, Pink node represent the parameters of malicious clients, Yellow node represents the mean of malicious clients parameters (i.e. average of parameters of Pink nodes), Red node represents the malicious parameters (of model M). We ensure that the shift in parameters of model M from mean is less than the threshold P_{dist} where P_{dist} is the maximum distance between any two attacked clients parameters. (b) Green nodes represent gradients of benign clients, Pink nodes represent the malicious clients gradients, Yellow node represents the mean of malicious clients gradients (i.e. average gradients of Pink nodes), Blue node represents gradient of trained malicious model M, Red node represents gradient of malicious model M after scaling. We ensure that after scaling gradients the distance from mean of gradients is less than threshold G_{dist} where G_{dist} is the minimum distance between any two attacked clients gradients.

Algorithm 1. DISBELIEVE Attack on Parameters

1: Calculate mean of parameters:

$$\mu^{param} = \frac{1}{f}\sum_{i=1}^{f} W_i^{mal}$$

2: Set the threshold value:

$$P_{dist} = Max_{i,k\in f\ i\neq k}||W_i^{mal} - W_k^{mal}||_2^2$$

3: Combine all the training data from malicious clients
4: Initialize the malicious model M with parameters μ^{param}
5: Train M with $Loss = -Loss_{class}$ until:

$$||W_{model}^{mal} - \mu^{param}||_2^2 \leq P_{dist}$$

6: Return W_{model}^{mal}

Algorithm 2. DISBELIEVE Attack on Gradients

1: Calculate the mean of parameters and gradients:

$$\mu^{param} = \frac{1}{f}\sum_{i=1}^{f} W_i^{mal} \qquad \mu^{grad} = \frac{1}{f}\sum_{i=1}^{f} Grad_i^{mal}$$

2: Set the threshold value:

$$G_{dist} = Min_{i,k\in f\ i\neq k}||Grads_i^{mal} - Grads_k^{mal}||_2^2$$

3: Combine all the training data from malicious clients
4: Initialize the malicious model M with parameters μ^{param}
5: Train M with $Loss = -Loss_{class}$
6: $Grads_{model}^{mal} \leftarrow Gradients\ of\ M$
7: $start \leftarrow 0.001$, $end \leftarrow 1000$
8: **while** $|start - end| > 0.01$ **do**
9: $\quad sf \leftarrow (start + end)/2$
10: $\quad Grads_{new}^{mal} = sf * \frac{Grads_{model}^{mal}}{||Grads_{model}^{mal}||}$
11: $\quad diff = ||Grads_{new}^{mal} - \mu^{grad}||_2^2$
12: $\quad$ **if** $diff > G_{dist}$ **then** $start = sf$ **else** $end = sf$
13: **end while**
14: Return the $Grads_{new}^{mal}$

soning attacks such as LIE attack [2] and Min-Max attack [19]. We note that the LIE attack forces the malicious parameters (or gradients) to be bounded in a range $(\mu - z\sigma, \mu + z\sigma)$ where μ and σ are the mean and standard deviation along parameters of the malicious clients, and z is a parameter that sets the lower and upper bounds for deviation around the mean [2]. On the other hand, Min-Max adds deviation to parameters or gradients and then scales them such that

their distance from any other non-malicious parameter is less than the maximum distance between two benign updates. However, instead of relying on standard deviation to approximate the range across which malicious clients' parameters (or gradients) can be manipulated, the proposed attack computes the malicious parameters (or gradients) by maximizing the classification loss (as opposed to minimizing it) to degrade the global model's performance. Additionally, we propose to approximate the range across which the parameters (or gradients) can be perturbed by evaluating the distance between the malicious clients' parameters (or gradients) in Euclidean space.

3 Proposed Method

Formally, we assume a total of n federated learning clients out of which f clients ($1 < f < n/2$) have been compromised such that rather than improving global models' accuracy, the compromised clients work towards decreasing the performance of the global model. We further assume that all the attackers corresponding to different malicious clients are working together or that a single attacker controls all the malicious clients. The attacker thus has access to all the malicious client's model parameters and training data. Our goal is to create malicious parameters or gradients that can bypass the robust aggregation algorithms and reduce the performance of the global model. In this direction, this research introduces a model poisoning attack (DISBELIEVE attack) that creates a single malicious model (M) with access to parameters, gradients, and training data of all the f clients. M serves as a proxy for f clients and aims towards pushing the output of the global model away from the distribution of the ground truth labels.

To be specific, the malicious model (M) is trained to generate malicious parameters or gradients by minimizing the loss $L_{model} = -L_{class}$ as opposed to benign clients where the loss given by $L_{model} = L_{class}$ is minimized. Here L_{class} refers to cross-entropy loss. Once the malicious parameters (or gradients) are computed, M forwards these malicious values to all the f clients, which then transmit these values to the global model. Note that all the f clients receive the same malicious parameters (or gradients) from M. Our work leverages the shortcomings of robust federated learning aggregation algorithms such as KRUM [3] and DOS [1], which are based on the assumption that malicious parameters or gradients are significantly different from the parameters or gradients of benign clients in euclidean space respectively. Therefore, to reduce the defense capabilities of these aggregation algorithms, it is essential to perturb the parameters (or gradients) so that their Euclidean distance from benign clients' parameters (or gradients) does not become significant. This can be ensured if the Euclidean distance between the malicious parameters (or gradients) and the mean of benign clients' parameters (or gradients) remains bounded. Due to the normal distribution of data, it is safe to assume that the mean of parameters (or gradients) of clients controlled by the attacker is closer to the mean of benign clients parameters (or gradients) respectively in the Euclidean space [2].

The local model poisoning attack can be introduced on model parameters or gradients [1,2]. However, the critical difference between parameters and gradients is that gradients have direction and magnitude, whereas parameters only have magnitude. Hence, we propose different attacks on parameters and gradients. Details on the strategy for attacking parameters or the gradients are provided in Sect. 3.1 and Sect. 3.2, respectively. The attacker initially chooses the clients it wants to attack and accumulates the chosen clients' model parameters, gradients, and training data. Subsequently, the attacker computes the mean of chosen (attacked) clients' model parameters (μ^{param}) and gradients (μ^{grad}) and initializes a new malicious model M with these mean values.

$$\mu^{param} = \frac{1}{f}\sum_{i=1}^{f} W_i^{mal} \qquad \mu^{grad} = \frac{1}{f}\sum_{i=1}^{f} Grad_i^{mal}$$

Here, W_i^{mal} and $Grad_i^{mal}$ refer to the model parameters or gradients of the i^{th} malicious client respectively.

3.1 DISBELIEVE Attack on Parameters

The initialized malicious model, M, is trained on the accumulated training data for minimizing the loss function $L_{model} = -L_{class}$ until the Euclidean distance between the malicious model's (M) parameters and the mean values is less than the maximum distance between any two attacked client's parameters.

$$||W_{model}^{mal} - \mu^{param}||_2^2 \leq P_{dist} \quad where, \quad P_{dist} = Max_{i,k\in f\ i\neq k}||W_i^{mal} - W_k^{mal}||_2^2$$

Here, W_{model}^{mal} refers to the malicious parameters after training of the malicious model M, and P_{dist} refers to a threshold. The threshold P_{dist} is critical to ensure a successful attack as it controls how far the malicious parameters can be from the mean of parameters in Euclidean space. Through the proposed attack, we suggest setting this value to the maximum Euclidean distance between any two malicious client parameters. Intuitively this is a reliable value within an upper bound on the malicious parameters by which they can deviate within a fixed bounded Euclidean space around the mean (see Fig. 1a). The pseudo-code for the attack is given in Algorithm 1.

3.2 DISBELIEVE Attack on Gradients

For attacking gradients, as described in Algorithm 2, we train the malicious model M with the similar loss function, $Loss = -Loss_{class}$, however, without any thresholding. Once the model M is trained, we accumulate the malicious gradients ($Grads_{model}^{mal}$) and scale them by a scaling factor sf to make sure that their distance from the mean of gradients of malicious clients (μ^{grad}) is smaller than the minimum distance between any two malicious client's gradients (G_{dist}) (see Fig. 1b).

$$G_{dist} = Min_{i,k\in f\ i\neq k}||Grads_i^{mal} - Grads_k^{mal}||_2^2$$

To find the optimum scaling factor (sf), we use a popular search algorithm known as binary search [19]. We initialize a start value of 0.001 and an end value of 1000. An optimal sf is computed using the divide and conquer binary search algorithm in between these values, which makes sure that after scaling the unit gradient vector, its distance to the mean of gradients (μ^{grad}) is less than G_{dist}

$$||sf * \frac{Grads^{mal}_{model}}{||Grads^{mal}_{model}||} - \mu^{grad}||_2^2 \leq G_{dist}$$

For calculating gradients, the minimum distance (G_{dist}) is preferred over the maximum distance (P_{dist}) when attacking parameters. This preference arises because maximizing the objective loss function results in gradients pointing in the opposite direction compared to the direction of benign gradients. By using the minimum distance, we can prevent malicious gradients from becoming outliers.

4 Experiments

4.1 Datasets

CheXpert-Small: CheXpert [10] is a large publicly available dataset containing over 200,000 chest X-ray images for 65,240 patients. However, consistent with the experimental protocol used by state-of-the-art DOS [1], we use the smaller version of CheXpert, also known as CheXpert-small, that contains 191,456 X-Ray images of the chest. The dataset contains 13 pathological categories. A single observation from the dataset can have multiple pathological labels. Each sample's pathological label is classified as either negative or positive. Consistent with the state-of-the-art aggregation method DOS [1], we preprocess all the images by rescaling them to 224×224 pixels using the torchxrayvision library.

Ham10000: Ham10000 [24] or HAM10k is a publicly available benchmark dataset containing dermatoscopic images of common pigmented skin lesions. It is a multi-class dataset with seven diagnostic categories and 10000 image samples. As suggested in [1], we use this dataset to evaluate the model performance in non-iid settings where each image is resized to 128×128.

Breakhis: The breakhis dataset [22] is a public breast cancer histopathological database that contains microscopic images of breast cancer tissues. The dataset contains 9109 images from 82 different patients. The images are available in magnifying scales such as 40X, 100X, 200X, and 400X. Each image is a 700 × 460 pixels sized image, and we rescale each image to 32 × 32 for our classification task. We use this dataset for binary classification of 400X magnified microscopic images where we classify cancer present in images as either benign or malignant.

CIFAR-10: The Cifar-10 [14] is a popular computer vision dataset that contains 60000 natural images of size 32×32. The dataset contains ten classes, and each class has 6000 images. 50000 images are reserved for training, and 10000 images are used for testing.

4.2 Experimental Setup and Implementation Details

The experimental setup used in this research is consistent with the experimental protocols suggested in [1]. Subsequently, we use Chexpert-Small [10] and Ham10k datasets [24] for parameter-based attacks. Likewise, the CheXpert-small dataset is used to train the Resnet-18 [8] model with a batch size of 16 for 40 communication rounds, and the number of local epochs is set to 1, whereas the Ham10k dataset is trained on a custom model with two convolutional layers and three fully connected layers with a batch size of 890 for 120 communication rounds and the number of local epochs were set to 3. For both datasets, the number of clients is fixed at 10, the number of attackers is fixed at 4, and the learning rate is set to 0.01.

For preserving the privacy of clients and their data, federated learning setups usually share gradients instead of model parameters. Hence, we also evaluate our attack for gradient aggregation on the Breakhis [22]. Furthermore, to assess the generalization ability of the proposed DISBELIEVE attack on natural images, we evaluate the proposed DISBELIEVE attack on the CIFAR-10 dataset with a gradient aggregation strategy at the global server. For experiments on Breakhis dataset, VGG-11 [21] model is trained for binary classification. Training occurs for 200 communication rounds with a batch size of 128 and a learning rate 0.0001. For the CIFAR-10 dataset, we use the VGG-11 [21] model with ten output classes

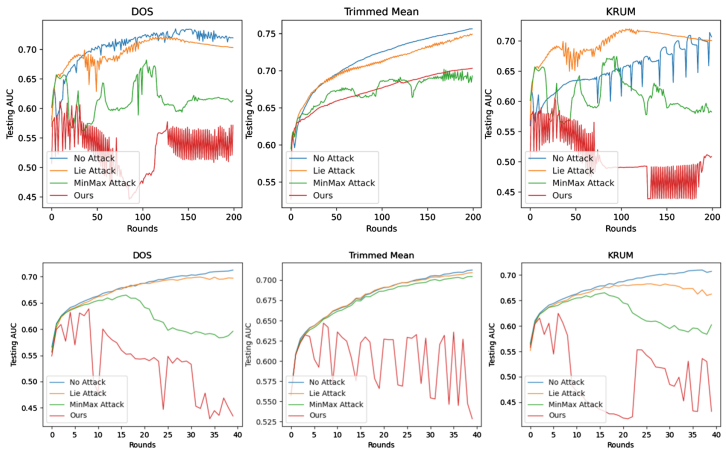


Fig. 2. Performance of different attacks on Ham10k (top-row) and CheXpert (bottom-row) datasets under different parameter aggregation methods. Left to right (in order): AUC scores when attacks are made on DOS, Trimmed Mean and Krum.

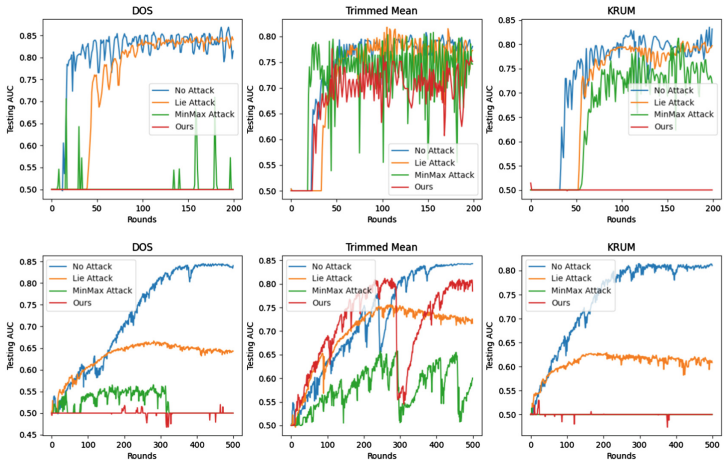


Fig. 3. Performance of different attacks on Breakhis (top-row) and CIFAR-10 (bottom-row) datasets under different gradient aggregation methods. Left to right (in order): AUC scores when attacks are made on DOS, Trimmed Mean and Krum.

for 500 communication rounds with a batch size of 1000 and a learning rate of 0.001. Adam optimizer was used for both datasets. The total number of clients and attackers for both datasets is fixed at 10 and 3, respectively.

Table 1. Area Under the Receiver Operating Characteristic Curve (AUC) scores with different types of poisoning attack on model parameters

Dataset	Attack	DOS	Trimmed Mean	KRUM
Ham10k	No Attack	0.72	0.75	0.70
	LIE Attack	0.70	0.74	0.70
	Min-Max Attack	0.61	0.68	0.58
	Ours	0.52	0.70	0.51
CheXpert	No Attack	0.71	0.71	0.70
	LIE Attack	0.69	0.71	0.65
	Min-Max Attack	0.59	0.70	0.59
	Ours	0.44	0.52	0.43

Table 2. Area Under the Receiver Operating Characteristic Curve (AUC) scores with different types of poisoning attack on model gradients

Dataset	Attack	DOS	Trimmed Mean	KRUM
Breakhis	No Attack	0.81	0.78	0.83
	LIE Attack	0.84	0.77	0.79
	Min-Max Attack	0.50	0.74	0.72
	Ours	0.50	0.75	0.50
CIFAR-10	No Attack	0.83	0.84	0.81
	LIE Attack	0.64	0.71	0.60
	Min-Max Attack	0.50	0.60	0.50
	Ours	0.50	0.78	0.50

5 Results and Discussions

5.1 Baselines

The DISBELIEVE attack is evaluated against three state-of-the-art defense methods: DOS [1], Trimmed Mean [26], and KRUM [3]. Comparisons are also made with prominent attacks, including LIE [2] and Min-Max [19], under different defense methods. Under any defense, AUC scores are highest in the absence of attacks. The LIE attack slightly reduces AUC scores while remaining relatively weaker due to parameter bounding. Conversely, introducing noise and scaling parameters makes the Min-Max attack more potent, consistently reducing AUC scores more significantly across various aggregation methods.

5.2 Vulnerability of State-of-the-Art Defense Methods

The proposed DISBELIEVE attack reveals the vulnerability of the current state-of-the-art robust aggregation algorithms (Trimmed Mean [26], KRUM [3], and DOS [1]) over local model poisoning attacks. We empirically validate that our proposed local model poisoning attack (DISBELIEVE attack) can successfully circumvent all three state-of-the-art robust aggregation algorithms (refer Figs. 2, 3). For both parameters and gradient aggregation, DISBELIEVE attack consistently reduces the global model's area under the curve (AUC) scores on all three benchmark medical image datasets. Furthermore, to assess the effectiveness of the proposed DISBELIEVE attack on natural images apart from the specialized medical images, we additionally conduct DISBELIEVE attack on a popular computer vision dataset, CIFAR-10. For natural images, we also find (refer Fig. 3) that the DISBELIEVE attack reduces the global model's AUC score for different state-of-the-art aggregation algorithms DOS, Trimmed Mean, and KRUM. Tables 1 and 2 show that when subjected to DISBELIEVE attack, the AUC scores fall drastically for all datasets compared to the AUC scores in case of no attack. Therefore, these results demonstrate the vulnerability of state-of-the-art robust aggregation methods to the proposed local model poisoning attack.

5.3 Superiority of DISBELIEVE Attack over State-of-the-art Local Model Poisoning Attacks

The state-of-the-art robust aggregation algorithm for medical images DOS is only evaluated against additive Gaussian noise, scaled parameter attacks, and label flipping attacks. We additionally benchmark the performance of two state-of-the-art model poisoning attacks, namely Min-Max [19] and LIE [2] on all the three medical image datasets (refer Figs. 2 and 3). Results establish the superiority of the proposed DISBELIEVE attack over state-of-the-art model poisoning attacks on different medical image datasets. While using DOS and KRUM aggregation, the DISBELIEVE attack reduces the global model's AUC score by a more significant margin than both Min-Max and LIE for all the datasets. In the case of trimmed mean, the results of DISBELIEVE attack are comparable on Ham10k (parameter aggregation) and Breakhis (gradient aggregation) datasets with the Min-Max attack and better on CheXpert (parameter aggregation) dataset when compared to the Min-Max and LIE attacks. To compare the effectiveness of DISBELIEVE attack with state-of-the-art model poisoning attacks on the natural image dataset (CIFAR-10), we observe that DISBELIEVE attack performs better than LIE and Min-Max on DOS and KRUM defenses. Tables 1 and 2 compare state-of-the-art model poisoning attacks and the proposed DISBELIEVE attack under different state-of-the-art robust aggregation algorithms for parameter and gradient aggregation, respectively.

6 Conclusion and Future Work

This research highlights the vulnerability of state-of-the-art robust aggregation methods for federated learning on medical images. Results obtained on three public medical datasets reveal that distance-based defenses fail once the attack is designed to ensure that the distance between malicious clients and honest clients' parameters or gradients is bounded by the maximum or minimum distance between parameters or gradients of any two attacked clients, respectively. Moreover, we also demonstrate that the proposed DISBELIEVE attack proves its efficacy on natural images besides domain-specific medical images. In the future, we plan to design a robust aggregation algorithm for federated learning in medical images that can withstand the proposed local model poisoning attack.

Acknowledgment. This work was done as a part of the IMI BigPicture project (IMI945358).

References

1. Alkhunaizi, N., Kamzolov, D., Takáč, M., Nandakumar, K.: Suppressing poisoning attacks on federated learning for medical imaging. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII, vol. 13438, pp. 673–683. Springer (2022). https://doi.org/10.1007/978-3-031-16452-1_64

2. Baruch, M., Baruch, G., Goldberg, Y.: A little is enough: circumventing defenses for distributed learning (2019)
3. Blanchard, P., El Mhamdi, E.M., Guerraoui, R., Stainer, J.: Machine learning with adversaries: byzantine tolerant gradient descent. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017)
4. Blanchard, P., Mhamdi, E.M.E., Guerraoui, R., Stainer, J.: Byzantine-tolerant machine learning (2017)
5. Chen, Y., Gui, Y., Lin, H., Gan, W., Wu, Y.: Federated learning attacks and defenses: a survey (2022)
6. Dayan, I., et al.: Federated learning for predicting clinical outcomes in patients with COVID-19. Nat. Med. **27**(10), 1735–1743 (2021)
7. Goodfellow, I.J., et al.: Generative adversarial networks (2014)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)
9. Hitaj, B., Ateniese, G., Perez-Cruz, F.: Deep models under the GAN: information leakage from collaborative deep learning (2017)
10. Irvin, J., et al.: CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison (2019)
11. Jere, M.S., Farnan, T., Koushanfar, F.: A taxonomy of attacks on federated learning. IEEE Secur. Priv. **19**(2), 20–28 (2021). https://doi.org/10.1109/MSEC.2020.3039941
12. Joshi, I., Kumar, S., Figueiredo, I.N.: Bag of visual words approach for bleeding detection in wireless capsule endoscopy images. In: Campilho, A., Karray, F. (eds.) Image Analysis and Recognition, pp. 575–582. Springer, Cham (2016)
13. Joshi, I., Mondal, A.K., Navab, N.: Chromosome cluster type identification using a Swin transformer. Appl. Sci. **13**(14), 8007 (2023). https://doi.org/10.3390/app13148007, https://www.mdpi.com/2076-3417/13/14/8007
14. Krizhevsky, A., et al.: Learning multiple layers of features from tiny images (2009)
15. Li, Z., Zhao, Y., Botta, N., Ionescu, C., Hu, X.: COPOD: Copula-based outlier detection. In: 2020 IEEE International Conference on Data Mining (ICDM). IEEE (2020). https://doi.org/10.1109/icdm50108.2020.00135
16. Lyu, L., Yu, H., Yang, Q.: Threats to federated learning: a survey (2020)
17. Lyu, L., Yu, H., Zhao, J., Yang, Q.: Threats to Federated Learning, pp. 3–16 (2020)
18. McMahan, H.B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data (2023)
19. Shejwalkar, V., Houmansadr, A.: Manipulating the byzantine: optimizing model poisoning attacks and defenses for federated learning. In: NDSS (2021)
20. Sheller, M.J., et al.: Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci. Rep. **10**(1), 1–12 (2020)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015)
22. Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L.: A dataset for breast cancer histopathological image classification. IEEE Trans. Biomed. Eng. **63**(7), 1455–1462 (2016). https://doi.org/10.1109/TBME.2015.2496264
23. Tolpegin, V., Truex, S., Gursoy, M.E., Liu, L.: Data poisoning attacks against federated learning systems. In: Chen, L., Li, N., Liang, K., Schneider, S. (eds.) Computer Security - ESORICS 2020, pp. 480–501. Springer, Cham (2020)
24. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data **5**(1) (2018). https://doi.org/10.1038/sdata.2018.161, https://doi.org/10.1038

25. Xie, C., Koyejo, O., Gupta, I.: Generalized byzantine-tolerant SGD (2018)
26. Yin, D., Chen, Y., Ramchandran, K., Bartlett, P.: Byzantine-robust distributed learning: towards optimal statistical rates (2021)