2024 IEEE Western New York Image and Signal Processing Workshop (WNYISPW) | 979-8-3315-0555-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/WNYISPW63690.2024.10786500

Towards More Robust Federated Learning with Medical Imaging Model Anomaly Detection

Dmitrii Korobeinikov
Department of Computer Science
Rochester Institute of Technology
Rochester, NY, USA
dk9148@rit.edu

Sergei Chuprov
Department of Computer Science
The University of Texas Rio Grande Valley
Edinburg, TX, USA
sergei.chuprov@utrgv.edu

Leon Reznik
Department of Computer Science
Rochester Institute of Technology
Rochester, NY, USA
leon.reznik@rit.edu

***Abstract*—Federated Learning (FL) is a decentralized approach to Machine Learning (ML) that allows distributed model training locally across different devices, instead data transportation to the aggregation point, thus increasing the data privacy protection. Nevertheless, adversarial attacks still possess a problem for FL in real-world applications, aiming to deteriorate the aggregated ML model performance or reconstruct the training data using the model updates. In this paper, we prove that our Model Anomaly Detection (MAD) technique based on model trust evaluation not only allows for the robust identification of the malicious participants during the distributed ML model training on early aggregation stages, but also improves the model convergence. We verify our findings on the subset of the MedicalMNIST dataset.**

***Index Terms*—federated learning, medical images, model anomaly detection, attack mitigation, medical imaging**

I. Introduction

Federated Learning (FL) was introduced in 2016 [6] and has received significant attention from the research community ever since. The key advantage of the FL over a conventional Machine Learning (ML) is the preservation of the training data privacy. In FL architecture, each data source produces its own local model which training is maintained locally as well. After each training round, local models are sent to the centralized aggregation server where the global model is generated using a specific aggregation strategy. Therefore, the confidential user data is never transferred over a network, enhancing data privacy protection in order to satisfy regulatory requirements. The FL feature is especially important for fields such as medical imaging, where the patient data privacy is heavily regulated. The schematic representation of the FL model training process can be found in Figure 1.

Although user data remains stored locally on individual data sources, the overall system is still vulnerable if these sources get compromised. The potential for intruders to compromise participating devices presents a risk to the global ML model. First, if a malicious entity joins the training process, it will receive global model updates after each aggregation round. As demonstrated in [10], it is feasible to reconstruct training data by accessing only the gradient updates from the centralized server. This vulnerability poses a considerable threat to the privacy of sensitive data, such as medical images and associated labels. Second, as our study reveals, the presence of malicious models within collaborative training not only degrades the accuracy of the centralized model but also negatively affects its convergence.

In this paper, we investigate the domain of medical images, conducting our experiments on the *PneumoniaMNIST* – a dataset which is a part of Medical MNIST dataset [8]. We utilize our Model Anomaly Detection (MAD) technique based on FL model trust evaluation during the collaborative ML model training.

The idea behind MAD consists of the set of techniques to identify and exclude malicious participants from the training process on the aggregation server [7]. By calculating trust values based on each client's historical behavior and the quality of their contributions, this approach allows for the detection and exclusion of clients supplying abnormal models from the aggregation process. MAD approach also incorporates trust thresholds for the client exclusion, enabling greater flexibility for implementation across diverse domains beyond medical imaging. The contributions of this work are as follows:

A. Using the medical imaging dataset, we verify that our MAD algorithm improves the overall FL security by identifying and excluding anomalous models from the joint training process;

B. We show that incorporation of MAD security measures improves the model convergence;

C. We show that the early exclusion of malicious parties allows for more robust aggregated model to be achieved earlier.

II. Model Anomaly Detection Foundation

Our MAD approach is based on the calculation of the trust indicator for each model that is supplied by participating clients [2], [9]. Before generating the final model, the aggregation server uses K-means clustering on the clients' model parameters to group them based on similarity. We then calculate each client's reputation, R, using the normalized Euclidean distance d from the center of the main cluster. Initially, the reputation R_i^{t0} for the i-th client is based on the first training round, where d_i is the difference between one and

This research was partially supported by the USA National Science Foundation (award # 2321652)

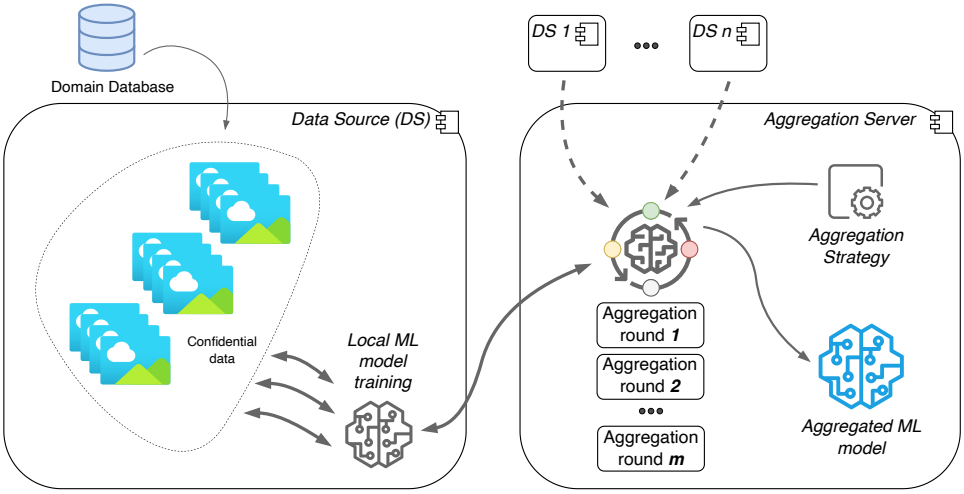


Fig. 1. Federated Learning Model Training Process

the normalized distance. Reputation is updated in each round: if $d \geq \alpha$ (with α set to 0.5 in our case), R increases linearly; if $d < \alpha$, it decreases exponentially. We also apply exponential smoothing to update parameters, combining the current and previous round's reputation with a smoothing factor β (0.75 in our case). The reputation at time t is adjusted based on the previous value R_i^{t-1}, penalizing clients that supply poor-quality models and requiring consistent positive contributions to build reputation.

$$R_i^{t0} = d_i \quad R, d \in [0,1] \subset \mathbb{R} \tag{1}$$

$$R_i^t = \begin{cases} (R_i^{t-1} + d_i) + (R_i^{t-1}/t), & \text{if } d \geq \alpha, \\ (R_i^{t-1} + d_i) - e^{-(1-d)(R_i^{t-1}/t)} & \text{if } d < \alpha \end{cases} \tag{2}$$

$$R_i^t = \begin{cases} 1, & \text{if } R \geq 1 \\ 0, & \text{if } R \leq 0 \end{cases} \tag{3}$$

$$R_i^t = \beta \cdot R_i^t + (1-\beta) \cdot R_i^{t-1} \tag{4}$$

Finally, the trust indicator, derived from R, determines how changes in R affect trust for each client. If the trust score falls below a set threshold β, the client's model is excluded from the current and future aggregation rounds. Trust for the i-th client at time t is calculated using R, the trust value d_i, and the previous round's trust value, with the first round starting at 0 for all clients. We use the same exponential smoothing method for trust, with a β value of 0.85. This ensures that clients with consistently good contributions maintain trust, while those with lower reputation or questionable models are removed from the FL process.

$$Trust_i^t = \sqrt{(R_i^t)^2 + d_i^2} - \sqrt{(1-R_i^t)^2 + (1-d_i)^2} \tag{5}$$

$$Trust_i^t = \beta \cdot Trust_i^t + (1-\beta) \cdot Trust_i^{t-1} \tag{6}$$

$$Trust_i^t = \begin{cases} 1, & \text{if } Trust \geq 1, \\ 0, & \text{if } Trust \leq 0 \end{cases} \tag{7}$$

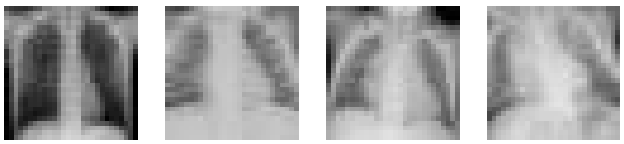
III. Empirical Study

A. Target Domain

For the assessment of our MAD approach, we explore the domain of computer vision applications for medical imaging. FL is especially suitable for medical image processing since its architecture enables local data preservation, thus facilitating the collaboration among hospitals and medical centers [3], as well as ensuring the compliance with privacy regulations such as HIPAA [3]. FL-enabled setups preserve patient privacy and comply with legal and ethical constraints, which is especially important in handling sensitive medical images from MRI, CT, X-ray, and histology. For instance, FL is already increasingly being used for tasks like tumor detection, organ segmentation, and disease diagnosis based on imaging data [4].

B. Data Collection

In our setup, we exercise the task of binary classification utilizing the PneumoniaMNIST dataset, a part of MedicalMNIST [8]. The dataset consists of 5856 images (4708 for training) of the chest x-rays with the labels indicating whether the patient associated with each X-ray was diagnosed with pneumonia. Since each image is stored as an array representing pixel brightness values, we process the dataset in order to convert these arrays into PNG images, which are more appropriate for simulating the real-world conditions. The MedicalMNIST dataset does not contain metadata that could link images to specific data sources due to privacy restrictions. Therefore, we distribute the training partition of the dataset in its entirety among participating clients manually in an IID (Independent and Identically Distributed) manner in such a way so that each client has its own subset of images, and these subsets do not intersect between each other. The IID setting is suitable for theoretical evaluation of our MAD mechanism. Examples of the images from the dataset can be found in Figure 2.



(a) (b) (c) (d)

Fig. 2. Examples of dataset images: (a), (b) – chest x-rays of patients without diagnosed pneumonia; (c), (d) – chest x-rays of patients with diagnosed pneumonia

C. Attack Model

We implement label flipping as a poisoning attack in our testing setup. We consider this specific type of attack, because, on the one hand, in the case of an intrusion this type of attack is easy for a malicious party to implement [5], and, on the other hand, it can significantly degrade the global model performance [5]. We implement 100% label flipping for 20% of participating clients as we see this as a realistic partition when considering a medical institution collaboration in the medical image processing domain.

D. Federated Learning Testbed Setup

For our experiments, we employ our FL testbed that is based on the Flower framework [1] that incorporates a custom aggregation strategy featuring MAD algorithm. This setup involves 10 participating clients and a range of 100 to 1000 aggregation rounds, depending on the experiment. Our testbed enables the comparison of metrics, such as aggregated loss and accuracy, across different simulation strategies with the consistent number of aggregation rounds. We vary the aggregation round at which the detection and exclusion of malicious clients is initiated, and execute FL simulations under these conditions. Subsequently, we analyze the collected history of loss and evaluation accuracy for each strategy.

IV. Results

Figure 3(a) illustrates the history of average client loss over rounds across strategies where the MAD algorithm is engaged from aggregation rounds 5, 20, 40, 60, and 80. Additionally, we illustrate the history of average client loss for the case when there is no attack. During each simulation, our MAD algorithm was able to detect and exclude both malicious clients that performed the label flipping attack on their data subsets. On the one hand, as clearly displayed on this figure, the exclusion of anomalous models leads to the immediate and significant decrease in the loss (in our experiments we utilized cross-entropy loss function). It is also clear from this graph that the presence of anomalous models entirely deteriorates the model training, as the average loss does not converge to an optimal value, instead slightly increasing with each aggregation round. On the other hand, as one can see in Figure 3(b), the removal of malicious clients immediately improves the average training accuracy.

In order to further showcase the importance of removal of malicious clients for the FL robustness with the regard to its accuracy, we conduct another set of simulations with the greater number of aggregation rounds. Specifically, we run the simulation for a total of 1000 rounds, comparing the average evaluation accuracy history of benign clients when the MAD algorithm was activated at round 10 versus round 500. The outcomes, shown in Figure 3(c), indicate that the presence of malicious clients in the aggregation process adversely impacts the accuracy of benign clients. The architecture of FL explains this effect: during centralized aggregation, anomalous models influence the global model, which is then redistributed to all clients at the end of each round. Moreover, the average accuracy in the presence of anomalous models is not only less in comparison to the case when such models were removed at the round 10, but also fluctuates from round to round until the MAD algorithm engages and the round 500, rendering the lesser robustness of the FL setup as a whole.

V. Conclusion

In this study we showed the importance of the implementation of MAD techniques in FL, with a focus on applications within the domain of medical imaging. We employed our MAD algorithm that is based on the client trust evaluation, utilizing the PneumoniaMNIST dataset, a subset of MedicalMNIST, for our experiments. The simulations were conducted within our FL testbed that is based on the Flower framework, incorporating our custom aggregation strategy. Across all of our experiments, the MAD algorithm successfully detected and excluded clients that performed the label flipping attack on their training subsets, from the training process. The removal of malicious clients increased the overall security of FL setup, as these clients were no longer able to receive global model updates, thereby preventing them from conducting membership inference attacks. Furthermore, our results showed that early exclusion of malicious clients improves the evaluation accuracy of benign client models, contributing to a more robust FL setup – an essential factor in sensitive fields such as medical imaging.

References

[1] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernandez-Marques, Y. Gao, L. Sani, K. H. Li, T. Parcollet, P. P. B. d. Gusmão, and N. D. Lane, "Flower: A friendly federated learning research framework," no. arXiv:2007.14390, Mar. 2022, arXiv:2007.14390. [Online]. Available: http://arxiv.org/abs/2007.14390

[2] S. Chuprov, I. Viksnin, I. Kim, E. Marinenkov, M. Usova, E. Lazarev, T. Melnikov, and D. Zakoldaev, "Reputation and trust approach for security and safety assurance in intersection management system," *Energies*, vol. 12, no. 23, p. 4527, 2019.

[3] F. R. da Silva, R. Camacho, and J. M. R. S. Tavares, "Federated learning in medical image analysis: A systematic survey," *Electronics*, vol. 13, no. 11, p. 47, Jan. 2024.

[4] N. Hernandez-Cruz, P. Saha, M. M. K. Sarker, and J. A. Noble, "Review of federated learning and machine learning-based methods for medical image analysis," *Big Data and Cognitive Computing*, vol. 8, no. 99, p. 99, Sep. 2024.

[5] N. M. Jebreel, J. Domingo-Ferrer, D. Sánchez, and A. Blanco-Justicia, "Defending against the label-flipping attack in federated learning," no. arXiv:2207.01982, Jul. 2022, arXiv:2207.01982. [Online]. Available: http://arxiv.org/abs/2207.01982

[6] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," *arXiv e-prints*, pp. arXiv–1602, 2016.

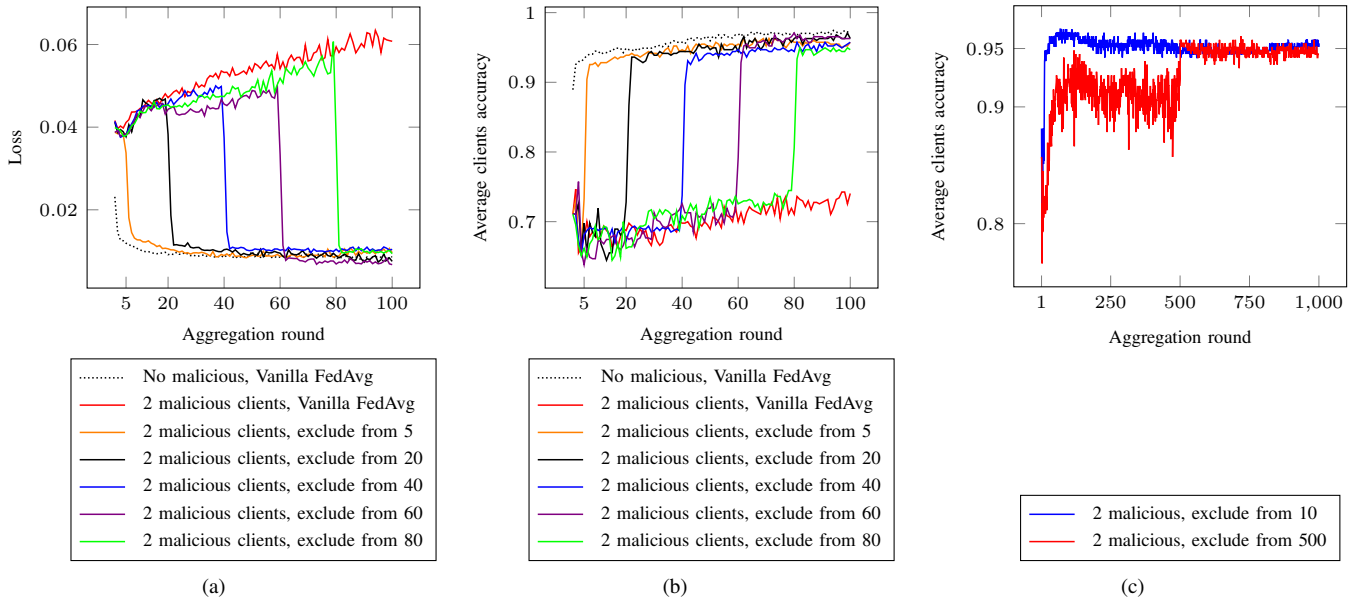


Fig. 3. Comparison of collected metrics among conventional FL and aggregation strategies with MAD applied at different aggregation round. (a) - history of average loss among clients, (b) - history of average accuracy among clients, (c) - history of average accuracy of benign clients.

[7] H. Patel, S. Chuprov, D. Korobeinikov, R. Zatsarenko, and L. Reznik, "Improving federated learning security with trust evaluation to detect adversarial attacks"." [Online]. Available: https://par.nsf.gov/biblio/10532526

[8] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, "Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification," *Scientific Data*, vol. 10, no. 1, p. 41, 2023.

[9] R. Zatsarenko, S. Chuprov, D. Korobeinikov, and L. Reznik, "Trust-based anomaly detection in federated edge learning," in *2024 IEEE World AI IoT Congress (AIIoT)*, May 2024, p. 273–279. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10578967

[10] L. Zhu, Z. Liu, and H. Song, "Deep leakage from gradients," *arXiv.org*, Dec. 2019, university: Cornell University Library arXiv.org. [Online]. Available: https://www.proquest.com/docview/2245977798?parentSessionId=TDcEiWYDOHLKHUFQaggZSxcOyUDxsecVVdNHBnxOX7o%3D&pq-origsite=summon&sourcetype=Working%20Papers