

# A Survey of Security Issues in Federated Learning

1<sup>st</sup> Yunhao Feng

National University of Defense and Technology  
Changsha, China  
fengyunhaonudt@nudt.edu.cn

2<sup>nd</sup> Yinjian Hou

National University of Defense and Technology  
Changsha, China  
houyinjian18@nudt.edu.cn

**Abstract**—As people’s awareness of the importance of personal privacy protection has grown, there has been a surge of interest in federated learning, which is a machine learning paradigm that enables training without requiring access to users’ private data.

## I. Introduction

The rapid development of digital technology has made the diversification, informationization, and diversity of digital data the main topics of the current era. Meanwhile, deep learning (DL) has demonstrated tremendous success in multiple fields, including computer vision, natural language processing, and graphic networks. Clearly, using diverse data in deep learning models can effectively improve their ability. However, there is also a growing interest in data privacy protection, such as the General Data Protection Regulation (GDPR) [1]. On the other hand, data sources may encounter the challenge of distributed storage, as is the case with data from mobile smart devices or Internet of Things (IoT) scenarios [2], [3]. Therefore, utilizing these data to train models requires overcoming limitations related to distribution and privacy [4].

To solve these problems, federated learning (FL) is a machine learning paradigm proposed as a possible response to these challenges [5]. FL enables collaborative model building among distributed members while ensuring sensitive data remains within each participant’s control [6]. Specifically, federated learning allows two or more participants to collaboratively train a shared global DL model while keeping their training datasets locally. Each participant trains the shared model on its own training data and exchanges and updates model parameters with other participants. Federated learning can improve the training speed and the performance of the shared model while protecting privacy of the participants’ training datasets [7]. Thus, it is a promising technique for the scenarios where the training data is sensitive (e.g., medical records, personally identifiable information, etc.) [8], [9].

Federated learning can be classified based on whether the participating datasets are the same, resulting in two types: homogeneous federated learning and heterogeneous federated learning [10], [11]. In homogeneous federated learning, all participants have datasets with the same characteristics and data distribution, whereas in heterogeneous federated learning, participants’ datasets may differ in their characteristics and data distribution. The second

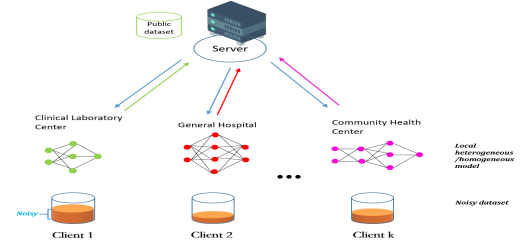


Fig. 1. A schematic of federated learning.

classification of federated learning is based on whether the models involved are the same, resulting in two types: horizontal federated learning and vertical federated learning [12], [13]. In horizontal federated learning, all participants have the same model architecture, but may have different local data [14], while in vertical federated learning, each participant has a different model architecture but they collaborate on processing the same set of data together [15]. The third way to classify federated learning is based on the type of task involved, resulting in several types such as federated learning for clustering [16], [17], federated learning for classification [18], [19], federated learning for regression [20], among others. The fourth way to classify federated learning is based on the optimization approach used between the participants, resulting in several types such as federated averaging [21], [22], federated learning optimization, federated meta-learning [23], and so on.

Federated learning methods currently face significant challenges related to their robustness. This article focuses on three main attacks, including backdoor attacks [24], [25], [26], [27], [28], adversarial attacks [31], [32], [33], [34], and Byzantine attacks [29], [30]. A backdoor attack involves a malicious participant in the federated learning process adding a backdoor to the model being trained, which can be triggered by a specific input pattern, allowing the attacker to control the output of the model in a targeted way. Adversarial attacks, on the other hand, entail adding small, carefully crafted perturbations to the input data to deceive the model and cause it to make incorrect predictions [31], [32].

And adversarial attacks can occur in federated learning when a malicious participant intentionally sends adversarial examples to the central server in an attempt to

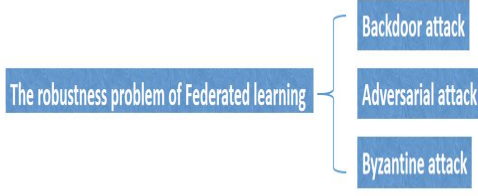


Fig. 2. The robust threat to federal learning

bias the model towards their own interests. This can be particularly problematic in applications such as personalized advertising or credit scoring, where the malicious participant may be motivated to gain an unfair advantage. Finally, Byzantine attacks involve one or more malicious participants in the federated learning process sending incorrect or misleading updates to the central server to disrupt the training process [35].

While federated learning can be vulnerable to certain types of attacks, there are techniques and approaches that can be used to improve the robustness and security of the process. It is important to carefully consider these issues when designing and implementing federated learning systems [38], [39]. For instance, knowledge distillation is a technique that can mitigate backdoor attacks by training a smaller [36], distilled model using the output of the original model as the target labels. This can help remove any backdoor triggers that may have been added to the original model, as the smaller model won't be able to identify them. Another technique to mitigate backdoor attacks is model erasure [44], where the model is trained to ignore specific input patterns that may be associated with the backdoor.

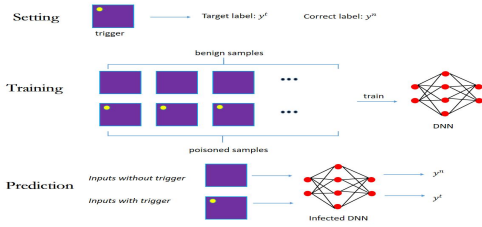


Fig. 3. Backdoor Attack

Adversarial training is a technique that involves explicitly training the model to resist adversarial examples by adding adversarial perturbations to the training data [31], [32]. This can improve the model's ability to detect and resist adversarial attacks in federated learning settings. Clustering can be used to identify malicious clients in federated learning systems subject to Byzantine attacks [35], [36]. The idea is to group participating clients based on the similarity of their updates, and to identify any

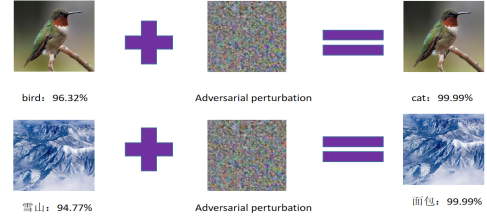


Fig. 4. Adversarial Attack

clients whose updates are significantly different from the others. These clients can then be excluded from the training process, or their updates can be treated with greater suspicion to minimize the impact of their malicious behavior [37]. This paper provides an overview

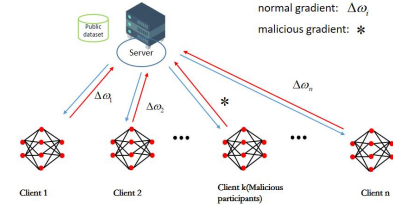


Fig. 5. Byzantine Attack

of methods to increase the robustness of federated learning models, with the aim of enhancing the credibility and security of federated learning. While previous work has addressed the security of federated learning [39], [40], [41], [42], [43], it has primarily focused on privacy leakage or backdoor attacks, with relatively few studies and reports on adversarial attacks. Building on prior work, this paper summarizes the attacks and defense methods of adversarial, backdoor, and Byzantine attacks in federated learning. A new classification method is proposed, supplementing the deficiencies of previous work on adversarial attacks. Moreover, this paper investigates a multi-level defense system against these attacks, and identifies open problems and future research directions for improving the robustness of federated learning.

## II. Thraet Model

Prior to delving into the details of the threats to federal learning, it is essential to establish the connections between these threats based on different criteria. Specifically, we can categorize these threats into two main stages: the training phase and the inference phase. Additionally, we can differentiate between untargeted attacks and targeted attacks based on whether a specific target is present or not. [38], [39], [40], [41].

## A. Training Phase and Inference Phase

1) Training Phase: Attacks that occur during the model training process are intended to either disrupt or impact the federated learning model itself. Backdoors are inserted into the model during the training phase to influence the resulting model outcomes [45], [46]. On the other hand, Byzantine attacks disrupt the convergence of the model by utilizing malicious clients or servers [29].

2) Inference Phase: Attacks that occur during the reasoning phase are typically intended to alter the model's reasoning outcomes and deceive it into generating incorrect outputs [47]. During the training stage, backdoor attacks involve the insertion of a backdoor into the model, whereas input deception models with triggers are utilized during the reasoning stage to cause the model to generate incorrect results. Adversarial attacks, on the other hand, leverage the model's vulnerability to disturbances and utilize samples with adversarial perturbations as input to the model, causing it to produce erroneous outcomes.

## B. Untargeted and Targeted

1) Untargeted attack: Untargeted attacks are designed to compromise the integrity of the target model in an arbitrary manner. Byzantine attack is one form of an untargeted attack that involves uploading malicious gradients to the server in an arbitrary manner, with the goal of causing the global model to fail [48], [49], [50], [51].

2) Targeted Attack: A targeted attack is executed with the aim of inducing the model to produce the target label specified by the adversary for specific testing examples, while keeping the testing error for other testing examples unaffected.

## III. Backdoor Attack

A backdoor attack on deep neural networks entails surreptitiously implanting a malicious backdoor within the model. This enables the model to function normally when processing benign inputs, but triggers a pre-defined malicious behavior when presented with a specific malicious trigger. The first neural backdoor in centralized settings can be traced back to 2013 [52]. A common strategy used by attackers is to embed an attacker-chosen trigger into clean samples, creating trigger samples, and then modify the labels of these samples to match the targeted class. By training the model on both normal and trigger samples, the model learns to associate the trigger with the targeted class while maintaining a similar level of accuracy on clean samples as a normal model would. This approach allows the attacker to achieve their goal without significantly impacting the model's performance on clean data [53].

## IV. Defenses against Backdoor Attack

### V. Byzantine Attack

## VI. Defenses against Byzantine Attack

### VII. Adversarial Attack

## VIII. Defenses against Adversarial Attack

### IX. Hybrid Defenses

## X. Advanced Research and Problems

## XI. Conclusion

## References

- [1] M. Goddard, The EU General Data Protection Regulation (GDPR): European regulation that has a global impact, *International Journal of Market Research* 59 (2017) 703–705.
- [2] O. Gómez-Carmona, D. Casado-Mansilla, F. A. Kraemer, D. L. de Ipiña, J. GarcíaZubia, Exploring the computational cost of machine learning at the edge for human-centric internet of things, *Future Generation Computer Systems* 112 (2020) 670–683.
- [3] J. Zhang, D. Tao, Empowering things with intelligence: A survey of the progress, challenges, and opportunities in artificial intelligence of things, *IEEE Internet of Things Journal* 8 (2021) 7789–7817.
- [4] C. Ma, J. Konečný, M. Jaggi, V. Smith, M. Jordan, P. Richtárik, M. Takáč, Distributed optimization with arbitrary local solvers, *Optimization Methods and Software* 32 (2017) 813–848.
- [5] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 20–22 Apr 2017.
- [6] GlobalFederatedLearningMarketbyApplication (Drug Discovery, Industrial IoT, Risk Management), Vertical (Healthcare Life Sciences, BFSI, Manufacturing, Automotive Transportation, Energy Utilities), and Region - Forecast to 2028, "researchandmarkets.com", Accessed date: May 12, 2023.
- [7] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, H. Yu, *Federated Learning*, Synthesis Lectures on Artificial Intelligence and Machine Learning, 2019.
- [8] M. Househ, E. Borycki, and A. Kushniruk. *Multiple Perspectives on Artificial Intelligence in Healthcare*. Springer, 2021.
- [9] R. Rau, R. Wardrop, and L. Zingales. *The Palgrave Handbook of Technological Finance*. Springer, 2021.
- [10] Fang, Xiuwen, and Mang Ye. "Robust federated learning with noisy and heterogeneous clients." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
- [11] Tang, Zhenheng, et al. "Virtual homogeneity learning: Defending against data heterogeneity in federated learning." *International Conference on Machine Learning*. PMLR, 2022.
- [12] Kairouz, Peter, et al. "Advances and open problems in federated learning." *Foundations and Trends® in Machine Learning* 14.1–2 (2021): 1-210.
- [13] Yang, Qiang, et al. "Federated machine learning: Concept and applications." *ACM Transactions on Intelligent Systems and Technology (TIST)* 10.2 (2019): 1-19.
- [14] Huang, Wei, et al. "Fairness and accuracy in horizontal federated learning." *Information Sciences* 589 (2022): 170-185.
- [15] Liu, Yang, et al. "Vertical federated learning." *arXiv preprint arXiv:2211.12814* (2022).
- [16] Ghosh, Avishek, et al. "An efficient framework for clustered federated learning." *Advances in Neural Information Processing Systems* 33 (2020): 19586-19597.
- [17] Briggs, Christopher, Zhong Fan, and Peter Andras. "Federated learning with hierarchical clustering of local updates to improve training on non-IID data." *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020.

- [18] Hsu, Tzu-Ming Harry, Hang Qi, and Matthew Brown. "Federated visual classification with real-world data distribution." *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*. Springer International Publishing, 2020.
- [19] Wahab, Omar Abdel, et al. "Federated machine learning: Survey, multi-level classification, desirable criteria and future directions in communication and networking systems." *IEEE Communications Surveys and Tutorials* 23.2 (2021): 1342-1397.
- [20] Yang, Shengwen, et al. "Parallel distributed logistic regression for vertical federated learning without third-party coordinator." *arXiv preprint arXiv:1911.09824* (2019).
- [21] Deng, Yuyang, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. "Distributionally robust federated averaging." *Advances in neural information processing systems* 33 (2020): 15111-15122.
- [22] Sun, Tao, Dongsheng Li, and Bao Wang. "Decentralized federated averaging." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.4 (2022): 4289-4301.
- [23] Fallah, Alireza, Aryan Mokhtari, and Asuman Ozdaglar. "Personalized federated learning: A meta-learning approach." *arXiv preprint arXiv:2002.07948* (2020).
- [24] Bagdasaryan, Eugene, et al. "How to backdoor federated learning." *International conference on artificial intelligence and statistics*. PMLR, 2020.
- [25] Wang, Hongyi, et al. "Attack of the tails: Yes, you really can backdoor federated learning." *Advances in Neural Information Processing Systems* 33 (2020): 16070-16084.
- [26] Gong, Xueluan, et al. "Backdoor attacks and defenses in federated learning: State-of-the-art, taxonomy, and future directions." *IEEE Wireless Communications* (2022).
- [27] Sun, Ziteng, et al. "Can you really backdoor federated learning?" *arXiv preprint arXiv:1911.07963* (2019).
- [28] Ozdayi, Mustafa Safa, Murat Kantarcioglu, and Yulia R. Gel. "Defending against backdoors in federated learning with robust learning rate." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. No. 10. 2021.
- [29] Fang, Minghong, et al. "Local model poisoning attacks to Byzantine-Robust federated learning." *29th USENIX security symposium (USENIX Security 20)*. 2020.
- [30] Guo, Shangwei, et al. Byzantine-Resilient Decentralized Stochastic Gradient Descent.
- [31] Zizzo, Giulio, et al. "Fat: Federated adversarial training." *arXiv preprint arXiv:2012.01791* (2020).
- [32] Chen, Chen, et al. "Calfat: Calibrated federated adversarial training with label skewness." *Advances in Neural Information Processing Systems* 35 (2022): 3569-3581.
- [33] Li, Xiaoxiao, Zhao Song, and Jiaming Yang. "Federated adversarial learning: A framework with convergence analysis." *International Conference on Machine Learning*. PMLR, 2023.
- [34] Zhang, Jie, et al. "Delving into the adversarial robustness of federated learning." *arXiv preprint arXiv:2302.09479* (2023).
- [35] Prakash, Saurav, and Amir Salman Avestimehr. "Mitigating byzantine attacks in federated learning." *arXiv preprint arXiv:2010.07541* (2020).
- [36] Huang, Hanxun, et al. "Distilling Cognitive Backdoor Patterns within an Image." *arXiv preprint arXiv:2301.10908* (2023).
- [37] Blanchard, Peva, et al. "Machine learning with adversaries: Byzantine tolerant gradient descent." *Advances in neural information processing systems* 30 (2017).
- [38] Lyu, Lingjuan, et al. "Privacy and robustness in federated learning: Attacks and defenses." *IEEE transactions on neural networks and learning systems* (2022).
- [39] Guo, Shangwei, et al. "Robust and privacy-preserving collaborative learning: A comprehensive survey." *arXiv preprint arXiv:2112.10183* (2021).
- [40] Enthoven, David, and Zaid Al-Ars. "An overview of federated deep learning privacy attacks and defensive strategies." *Federated Learning Systems: Towards Next-Generation AI* (2021): 173-196.
- [41] Rodríguez-Barroso, Nuria, et al. "Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges." *Information Fusion* 90 (2023): 148-173.
- [42] Tariq, Asadullah, et al. "Trustworthy Federated Learning: A Survey." *arXiv preprint arXiv:2305.11537* (2023).
- [43] Zhang, Yifei, et al. "A Survey of Trustworthy Federated Learning with Perspectives on Security, Robustness, and Privacy." *arXiv preprint arXiv:2302.10637* (2023).
- [44] Dal Fabbro, Nicolò, Aritra Mitra, and George J. Pappas. "Federated TD Learning over Finite-Rate Erasure Channels: Linear Speedup under Markovian Sampling." *IEEE Control Systems Letters* (2023).
- [45] Miao, Chenglin, et al. "Towards data poisoning attacks in crowd sensing systems." *Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*. 2018.
- [46] Zhang, Hengtong, et al. "Data poisoning attack against knowledge graph embedding." *arXiv preprint arXiv:1904.12052* (2019).
- [47] Barreno, Marco, et al. "Can machine learning be secure?." *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*. 2006.
- [48] Lamport, Leslie, Robert Shostak, and Marshall Pease. "The Byzantine generals problem." *Concurrency: the works of leslie lamport*. 2019. 203-226.
- [49] Xie, Cong, Oluwasanmi Koyejo, and Indranil Gupta. "Fall of empires: Breaking byzantine-tolerant sgd by inner product manipulation." *Uncertainty in Artificial Intelligence*. PMLR, 2020.
- [50] Bernstein, Jeremy, et al. "signSGD with majority vote is communication efficient and fault tolerant." *arXiv preprint arXiv:1810.05291* (2018).
- [51] Damaskinos, Georgios, et al. "Aggregathor: Byzantine machine learning via robust gradient aggregation." *Proceedings of Machine Learning and Systems* 1 (2019): 81-106.
- [52] Geigel, Arturo. "Neural network trojan." *Journal of Computer Security* 21.2 (2013): 191-232.
- [53] Gu, Tianyu, Brendan Dolan-Gavitt, and Siddharth Garg. "Badnets: Identifying vulnerabilities in the machine learning model supply chain." *arXiv preprint arXiv:1708.06733* (2017).