Taxonomy of Attacks in Federated Learning

Groupe 2 - MLSecOps

November 2025

Abstract

Federated Learning (FL) allows decentralized training of machine learning models without centralizing sensitive data. However, it remains vulnerable to multiple types of attacks. This section presents a taxonomy of known attacks, adapted from Liu et al. (2023) and extended with recent literature.

1 Taxonomy of Attacks in Federated Learning

1.1 Overview

Federated Learning (FL) introduces distributed model training across multiple clients without centralizing data. While this paradigm preserves data locality, it exposes the system to unique security and privacy threats that differ from traditional centralized ML settings. These attacks can be categorized into five principal groups: Integrity, Privacy, Communication, Aggregator/Server, and Adversarial or Robustness attacks (adapted from Lyu et al., 2022; Nguyen et al., 2023).

Figure 1 provides an overview of the main attack surfaces in FL, grouped by layer: Integrity, Privacy, Communication, Aggregation, and Robustness.

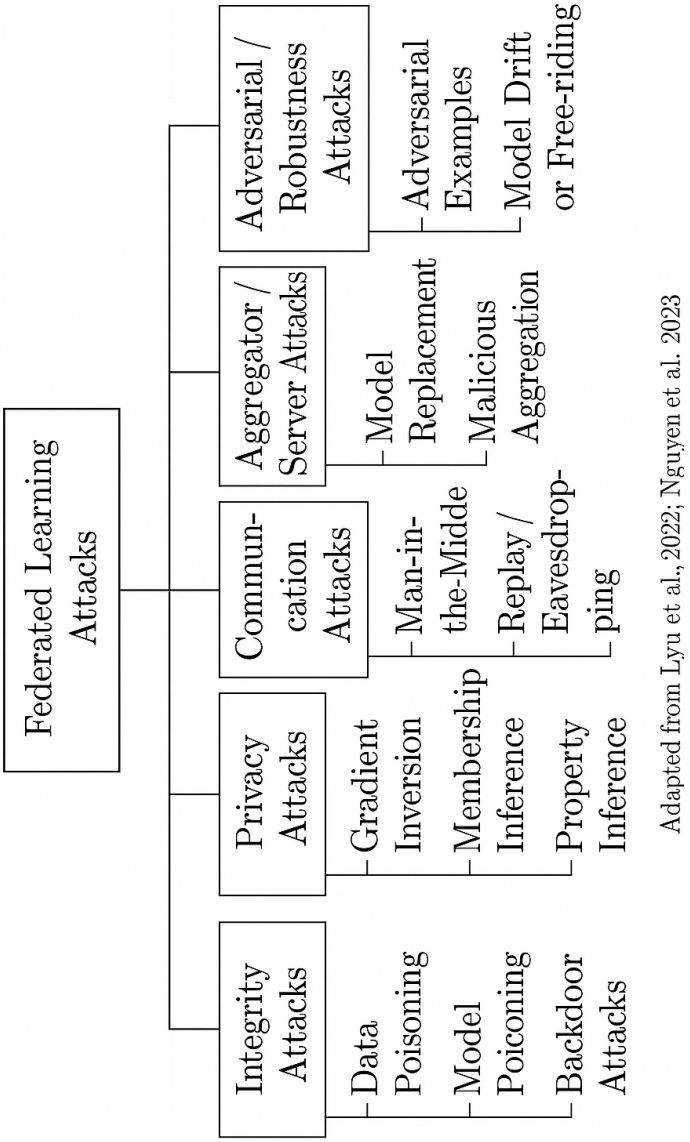

Figure 1: Taxonomy of Attacks and Defenses in Federated Learning

1.2 Integrity Attacks

Integrity attacks aim to compromise the correctness or reliability of the global model. Rather than stealing information, the adversary manipulates local updates to degrade model performance or introduce hidden behaviors. Typical examples include data poisoning—where malicious samples are injected into a client's dataset—and model poisoning, in which attackers directly modify the local model gradients before aggregation. Backdoor attacks constitute a more subtle variant: they embed a hidden trigger (e.g., a specific pixel pattern or phrase) that causes targeted misclassification while preserving overall accuracy. These threats are particularly difficult to detect in FL because of the asynchronous, privacy-preserving nature of local updates. (Bagdasaryan et al., 2020; Fung et al., 2020)

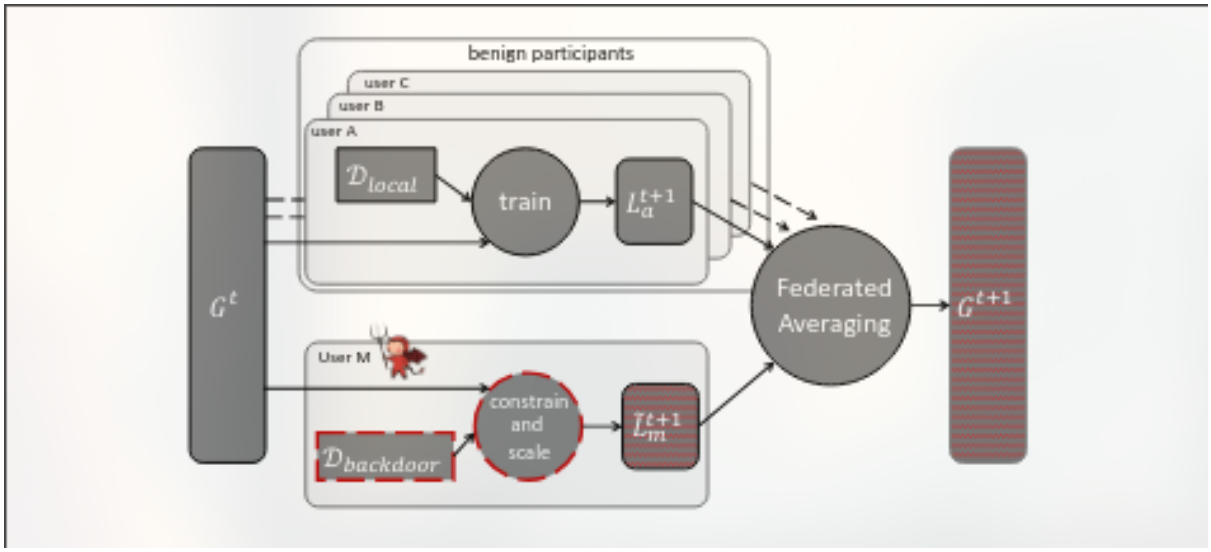

Figure 2: The attacker compromises one or more participants, trains on the backdoor data using the constrain-and-scale technique, and submits the resulting model, which replaces the joint model as the result of federated averaging. (Bagdasaryan et al., 2020; Fung et al., 2020)

1.3 Privacy Attacks

Privacy attacks focus on recovering or inferring information about clients' training data from shared model parameters or gradients. Despite not having access to raw data, attackers can exploit mathematical properties of the gradients to reconstruct private inputs (gradient inversion, Zhu et al., 2019) or determine whether a specific data point was used in training (membership inference). More sophisticated approaches—such as property inference—can deduce aggregate demographic or categorical information about local datasets. These attacks demonstrate that FL alone does not guarantee privacy; without cryptographic protection or differential privacy, model updates remain vulnerable to inversion and inference analysis. (Melis et al., 2019)

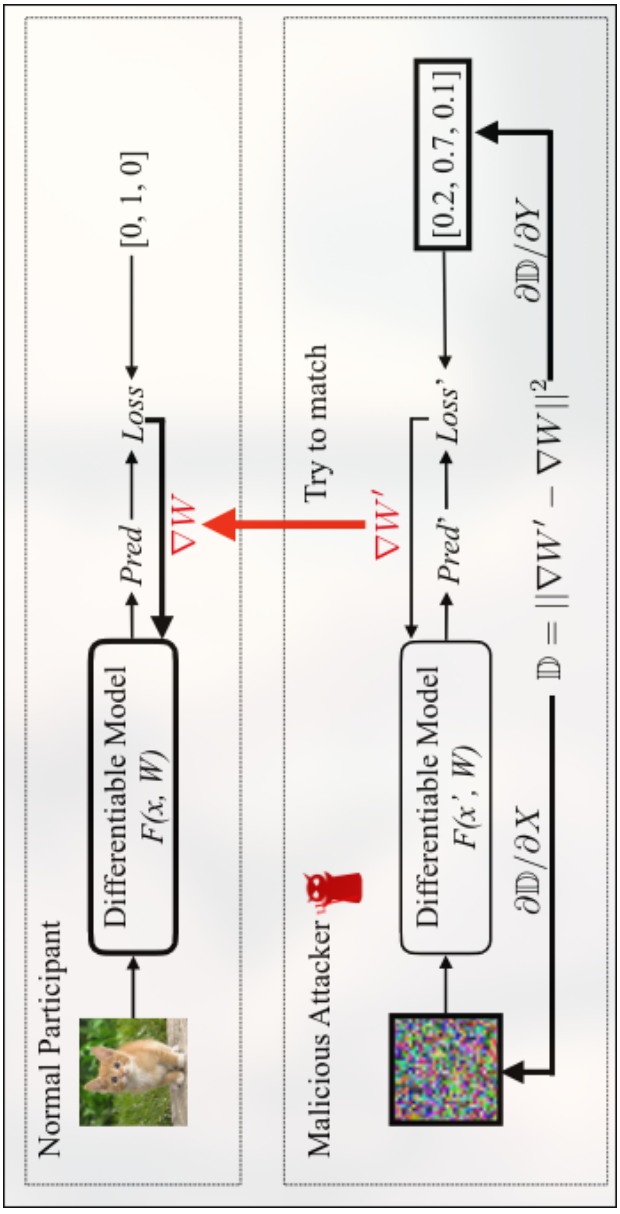

Figure 3: The overview of the DLG algorithm. Variables to be updated are marked with a bold border. While normal participants calculate deltaW to update parameter using its private training data, the malicious attacker updates its dummy inputs and labels to minimize the gradients distance. When the optimization finishes, the evil user is able to steal the training data from honest participants. (Zhu et al. 2019)

1.4 Communication Attacks

In communication attacks, the adversary targets the exchange channel between clients and the central server. Since FL involves frequent transmission of model updates, a man-in-the-middle or replay attacker can intercept, modify, or resend parameter messages. Eavesdropping on updates may reveal sensitive information about gradients or hyperparameters. Securing FL communication requires authenticated and encrypted channels, but even then, timing or metadata side-channels may leak information (Lyu et al., 2022).

1.5 Aggregator / Server Attacks

In a standard FL setting, the server is assumed to be semi-honest—but this assumption may not hold in adversarial contexts. A malicious or compromised server can perform model replacement (substituting the aggregated model with a poisoned version) or malicious aggregation (selectively weighting updates to bias outcomes). Conversely, even honest servers can unintentionally enable these attacks if they rely on naive averaging (FedAvg) without anomaly detection or robust aggregation. Hence, secure aggregation protocols and verifiable model updates are crucial to maintaining integrity at the central level. (Bhagoji et al., 2019)

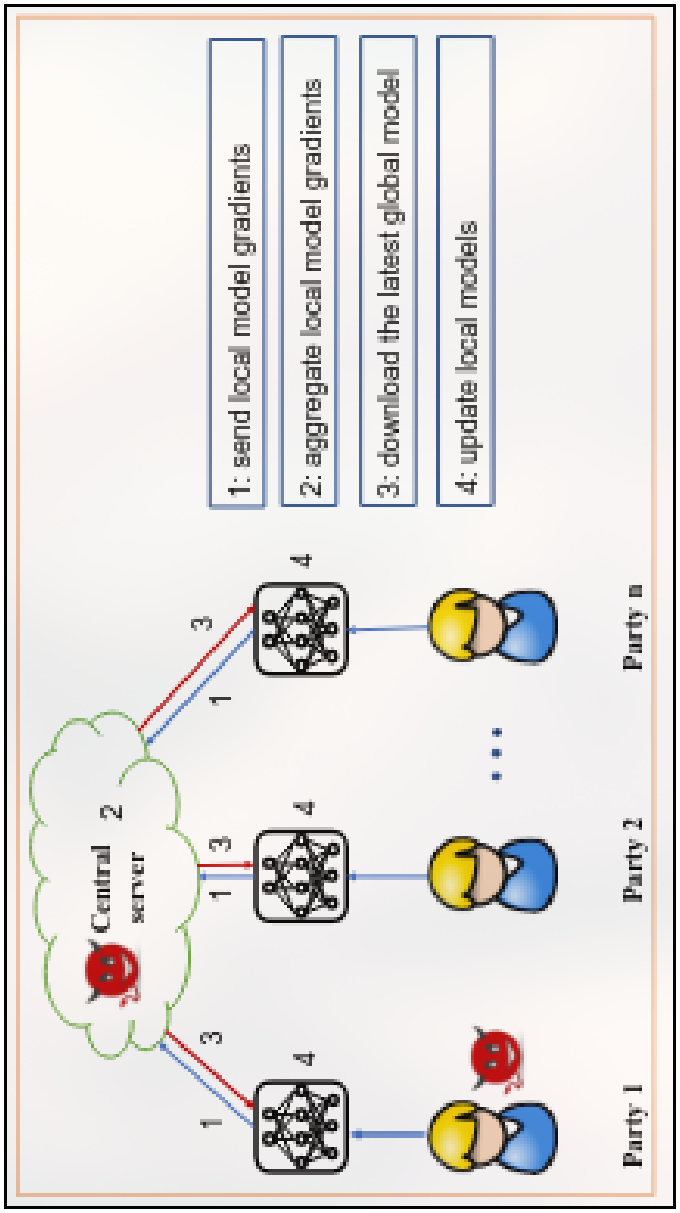

Figure 4: A typical FL training process, in which both the (po- tentially malicious) FL server/aggregator and malicious participants may compromise the FL system. (Threats to Federated Learning: A Survey, Lyu et al.)

1.6 Adversarial / Robustness Attacks

Beyond the training process itself, adversarial or robustness attacks target the learned model's behavior during inference. Attackers craft adversarial examples—inputs with imper-

ceptible perturbations that cause misclassification or exploit model drift and free-riding phenomena, where participants benefit from others' updates without contributing valid gradients. These attacks highlight that FL systems must combine robustness evaluation, adversarial training, and continuous monitoring to maintain resilience against both internal and external threats. (Nguyen et al., 2023)

Table 1: Comprehensive Mapping of Federated Learning Attacks

Category	Subtype	Description	Key References
Poisoning Attacks	Data Poisoning	Manipulation of local datasets to bias global model updates	[1, 2, 3]
	Model Poisoning	Direct modification of gradients or model weights to introduce hidden behavior or degradation	[4, 5]
Backdoor Attacks	Trigger-based Backdoors	Introduction of hidden malicious patterns (triggers) that cause misclassification when activated	[1]
Inference Attacks	Membership Inference	Determining whether specific data points were used in training	[6, 7]
	Property Inference	Inferring sensitive attributes of users' data from gradients or model parameters	[7]
	Gradient Inversion	Reconstructing training samples from shared gradients	[8, 9]
Communication Attacks	Man-in-the-Middle (MITM)	Intercepting or modifying updates in transit between clients and server	[10]
	Sybil Attacks	A single adversary simulates multiple fake clients to bias aggregation	[11]
Free-rider Attacks	Model Theft	Participants submit fake updates but still benefit from global model	[12]
Model Replacement	Model Overwrite	Substituting the global model with a maliciously crafted one	[1]

1.7 Representative Works

Table 2 summarizes key representative works, their publication venues, and their classification in the CORE database.

Attack Type	**Representative Paper**	**Venue**	**CORE Rank**	**Description**
Data Poisoning	Bagdasaryan et al., 2020	AISTATS	A	Demonstrates model-replacement and backdoor attacks targeting the global model aggregation in FL.
Gradient Leakage / Inference	Zhu et al., 2019	NeurIPS	A*	Shows that private training data can be reconstructed from shared gradients.
Property / Membership Inference	Melis et al., 2019	IEEE S&P	A*	Reveals feature leakage through model updates during collaborative training.
Privacy Analysis	Nasr et al., 2019	IEEE S&P	A*	Formalizes white-box inference attacks against centralized and federated models.
Byzantine / Robustness	Blanchard et al., 2017	NeurIPS	A*	Introduces Byzantine-tolerant gradient aggregation methods for adversarial clients.
Sybil / Collusion Attacks	Fung et al., 2020	RAID	A	Demonstrates Sybil-based poisoning and proposes FoolsGold defense.
Distributed Backdoors	Xie et al., 2020	ICLR	A*	Introduces DBA attacks — coordinated small-scale poisoning to evade detection.
GAN-based Reconstruction	Hitaj et al., 2017	ACM CCS	A*	Early work using GANs to extract data from collaborative learning setups.

Table 2: Mapping of Attacks in Federated Learning: representative works and venues.

References

[1] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "Backdoor attacks on federated learning," *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

[2] M. Fang, X. Cao, J. Jia, and N. Z. Gong, "Local model poisoning attacks to byzantine-robust federated learning," *USENIX Security Symposium*, 2020.

[3] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[4] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Analyzing federated learning through an adversarial lens," *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

[5] G. Baruch, M. Baruch, and Y. Goldberg, "A little is enough: Circumventing defenses for distributed learning," *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[6] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," *IEEE Symposium on Security and Privacy (S&P)*, 2019.

[7] L. Melis, C. Song, E. D. Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," *IEEE Symposium on Security and Privacy (S&P)*, 2019.

[8] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[9] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients - how easy is it to break privacy in federated learning?" *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[10] L. Lyu, H. Yu, and Q. Yang, "Threats to federated learning: A survey," *ACM Computing Surveys*, 2020.

[11] C. Fung, C. J. Yoon, and I. Beschastnikh, "Mitigating sybils in federated learning poisoning," *Proceedings of the 21st International Symposium on Research in Attacks, Intrusions and Defenses (RAID)*, 2018.

[12] Y. Wang, Y. Hu, G. Li, and Z. Lin, "Attack of the tails: Yes, you really can backdoor federated learning," *USENIX Security Symposium*, 2021.