Backdoor attacks on FL:

Adversary A.

Local model with client size N_A.

Feature space D.

Goal: Output a certain class C_A for a set of input samples (I).

I = Trigger set, $I \in D$.

Success of attack: Accuracy of backdoor task

Attack Strategy:

1) Data poisoning:

A: poisons client's training data

Attack data: Input from trigger set with new class C_A.

→ eg: Replace malware data class with benign class.

Attack Impact and Stealthiness:

D_i : benign dataset of compromised client i

D_i^A : injected attack data

$$PDR\ [\text{poisoned data rate}] = \frac{|D_i^A|}{|D_i'|}$$

[D_i' = poisoned dataset]

- very low PDR may have negligible effect on aggregation.
- Attacker should limit PDR to maintain stealthiness.

→ less complex than model poisoning.

Model poisoning:

- Requires stronger adversary
- A can change model updates before submitting to server.
- A can restrict the L_2-norm of the update upto a certain value S:

→ prevents too much suspicion.