Using Generative Adversarial Networks to Detect Model Poisoning Attacks in Federated Machine Learning

Usama Zafar [1] Salman Toor [1] André Teixeira [1]

[1]Department of Information Technology, Uppsala University, Sweden

Introduction

Federated machine learning (FedML) is a promising approach that enables multiple participants to collaboratively train a shared model without requiring them to share their data. FedML, being decentralized in nature, is vulnerable to various security threats.

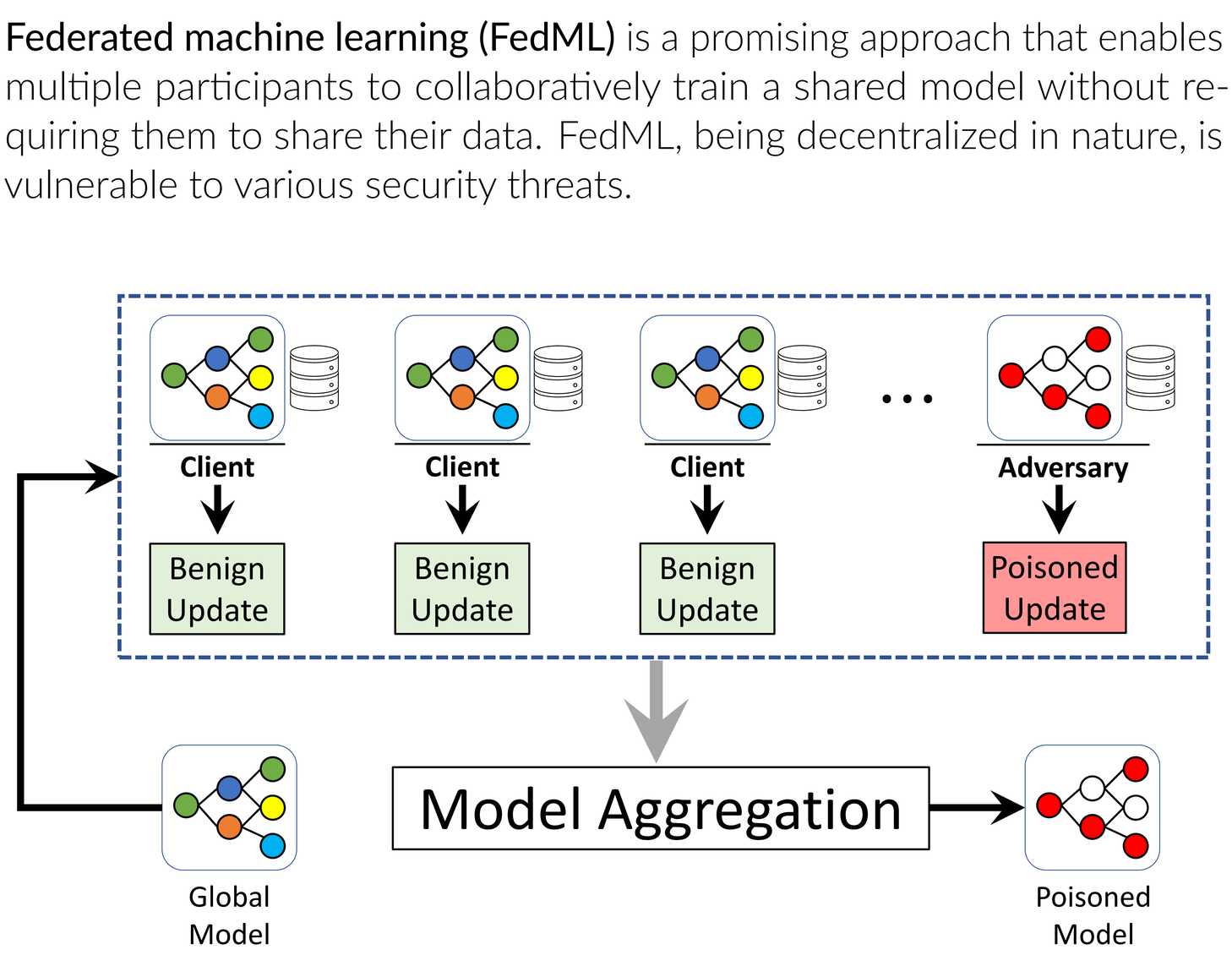
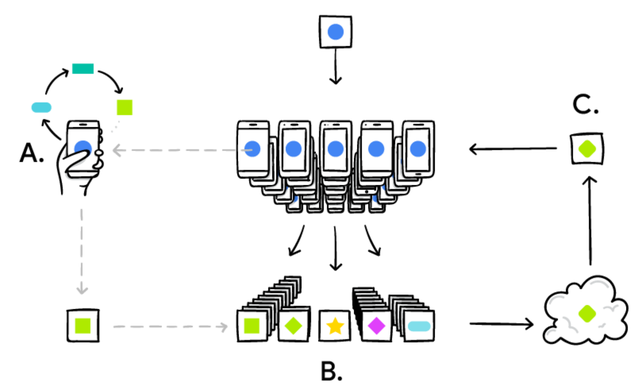


Figure 1. Poisoning Attack in FedML



Poisoning Attack

Poison attacks are a serious security threat that can lead to unreliable results. Broadly categorized as either:

- **Data Poisoning Attacks** based on fake data injection.
- **Model Poisoning Attacks** based on fake update injection.

Goal: undermine model's performance.

Challenges

- **Filtering** poisoned model updates from genuine updated.
- **Ensuring** privacy is not violated during the filtration process.

Proposed Solution

Using **Generative Adversarial Networks (GANs)** in conjunction with last model state to robustly differentiate poisoned updates from benign updates.

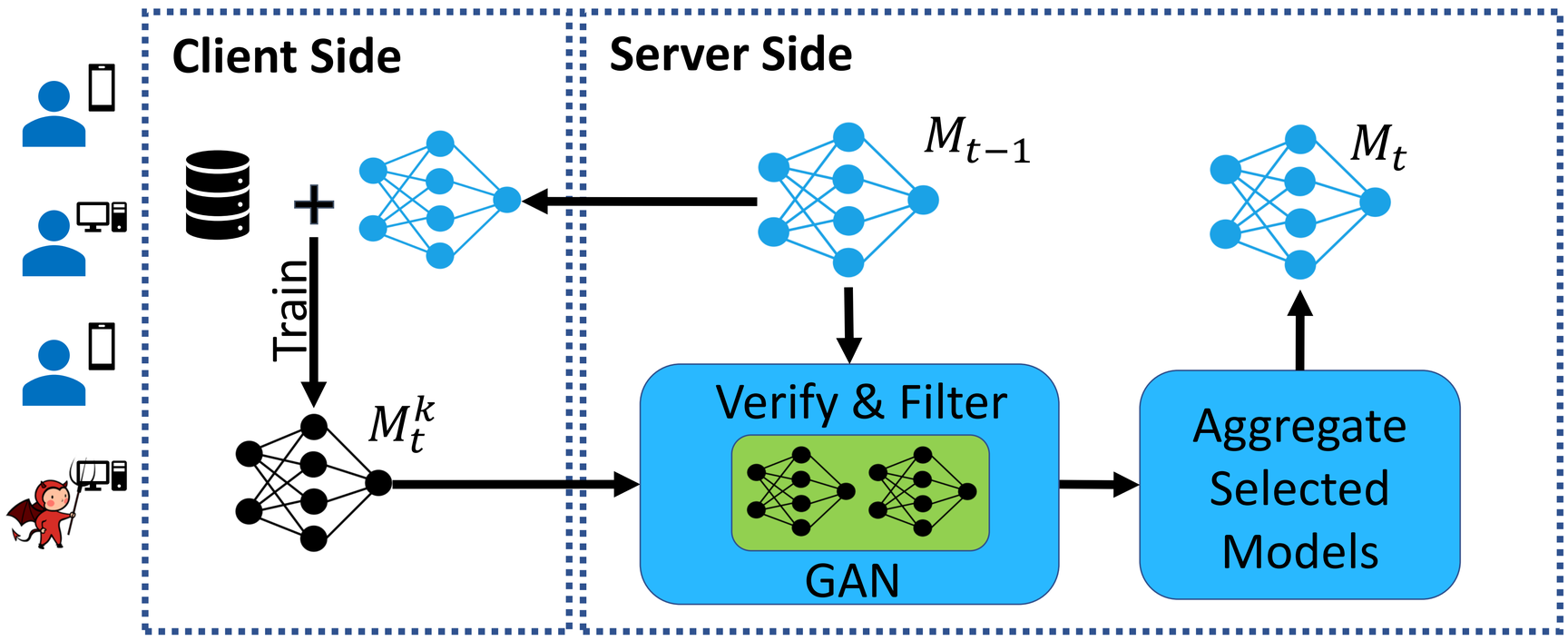


Figure 2. Using GANs to filter poisoning attacks in FedML

References

[1] Federated learning: Collaborative machine learning without centralized training data. https://ai.googleblog.com/2017/04/federated-learning-collaborative.html. Accessed: 2023-05-20.

Acknowledgments

This project is funded by *Graduate School in Cybersecurity, Uppsala University.*

UPPSALA UNIVERSITET