Poisoning Attack Mitigation for Federated-Learning-based Multi-Label Classifiers

Kainoa Aqui, Walt Wilber

Mentors: Muhammad Jahanzeb Khan, Ahmad Alsharif

THE UNIVERSITY OF ALABAMA®

INTRODUCTION

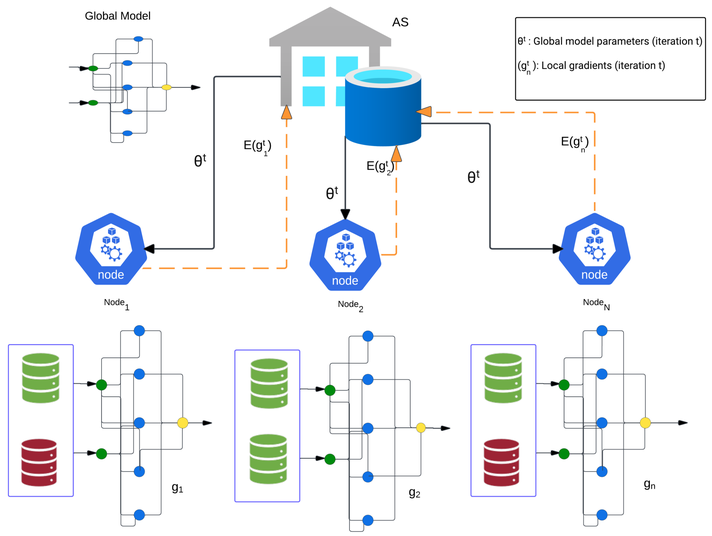
Problem Statement

- Deep Learning (DL)-based approaches in a federated learning setting are vulnerable to breaches of privacy and poisoning attacks
- Poisoning attacks involve maliciously altering input data to influence a model to act in a certain way, potentially introducing backdoors to be exploited later

Research Gap

- Most current Federated Learning (FL) cryptography research deals with security in regards to privacy of contributors, but not poisoning attacks
- Poisoning attacks are difficult to detect without analyzing confidential data from the contributors
- The new system adheres to these design goals:
  - Mitigates harm from poisoned local updates
  - Guarantees consumer privacy
  - Retains high central model accuracy

SYSTEM ARCHITECTURE



- The attack is initiated by contributors to the FL model
- If the attack succeeds, a vulnerability will be added to the model that will be difficult to detect and remove

METHOD

Defense Mechanism

- The system filters out malicious local updates based on the cosine similarity (cs) between each update and the global model, calculated as follows:

$$cs(i) = \frac{\langle W^{t-1},\ g_i^t \rangle}{||W^{t-1}|| \cdot ||g_i^t||}$$

- If a local update falls too far outside a similarity threshold, it will not be included in the current or any subsequent global model
- Functional Encryption is used to encrypt the local updates, allowing the cosine similarity calculation to be performed without ever revealing the real weights of any local model

EXPERIMENT

Dataset Preparation

- The experiment was run using 300 samples of power consumption data from the Irish Smart Energy Trials
- A 60-20-20 train-test-validation split was used, and data was distributed among 30 Detection Nodes

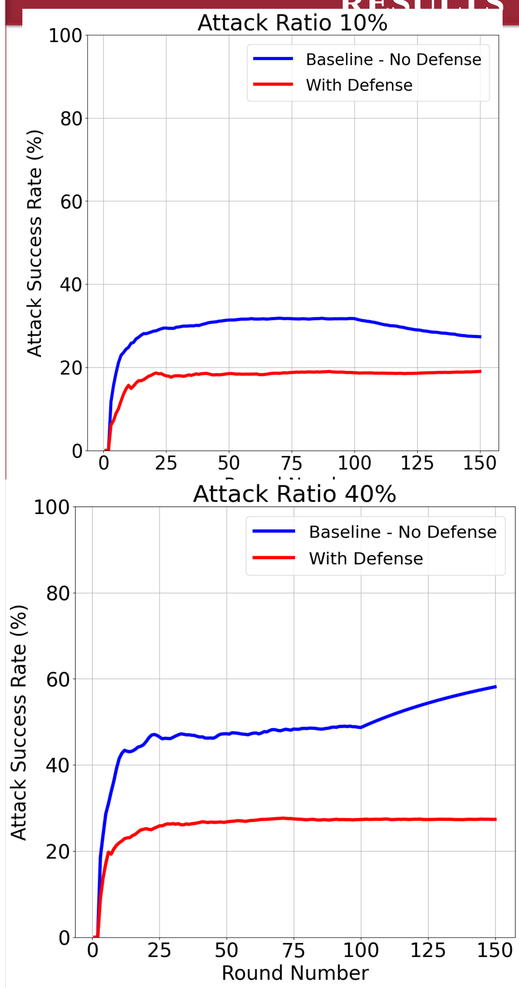
Generation of Results

- A Temporal Convolutional Network (TCN) Model using the new defense mechanism was tasked with classifying consumption measurements as either legitimate or fraudulent
- A certain proportion of the Detection Nodes were provided data that was poisoned with the following methods: Label Flip, Noise, Scaling, and Shift
- The experiment was run with 5 training epochs per client per round for 150 rounds

Performance Metrics

- The effectiveness of the system defense was measured by Attack Success Rate (ASR), which is the proportion of fraudulent data falsely classified as legitimate by the model during validation

RESULTS



- There was a consistent reduction in the ASR across all Attack Ratios

NEXT STEPS

- Fine tuning the model parameters to increase the speed of learning and the final accuracy
- Implementing encryption to the model updates
- Testing effectiveness of the system on cases with more than two labels

College of Engineering Computer Science

College of Engineering Civil, Construction and Environmental Engineering

Alabama Transportation Institute

ACKNOWLEDGMENTS

This material is based on a study supported by the National Science Foundation under Grant No. 2244371. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, and the U.S. Government assumes no liability for the contents or use thereof.