

Security of AI Models Against Adversarial Attacks

Prepared by

ALI MMADI, PhD student - Midocean University



Research Issues 1

Task 4

Selected article: "Robust Federated Learning Against Model Poisoning Attacks in IoT Networks" (Authors: Zhang et al., 2024)

INTRODUCTION

Hello! For this 4th task, I am exploring an invisible threat: model poisoning attacks in Internet of Things (IoT). The article by Zhang et al. (2024) reveals how hackers corrupt federated learning - a key technology for preserving privacy. Did you know that 35% of IoT models are vulnerable to these silent attacks?

Title and variables

The title is technical but precise:

Independent: Type of attack (data poisoning vs. model)

Dependent:

Model robustness (accuracy after attack)

Moderator:

Federated learning architecture The IoT domain is crucial: with 30 billion connected devices by 2025, every vulnerability becomes catastrophic.



Central problem

The article fills a dangerous gap: existing defenses fail against adaptive attacks. Key question: 'How to detect sophisticated poisoning attacks in heterogeneous federated systems without compromising privacy?' A major challenge when the data stays with the users!

Key innovations

Two breakthroughs:

1) **Differential anomaly detection:** Compares local gradients without seeing the data

2) **Reputation mechanism:** Identifies malicious nodes by their behavior Tested on 10,000 medical sensors: 89% reduction in attacks with only 2% overhead!

Methodology

Approach in 3 steps:

Inject 15 types of attacks into real IoT networks.

Train a detector based on Deep K-Means.

Validate via a smart city simulator The data: 5.7 million points from industrial and medical sensors.

Conceptual model

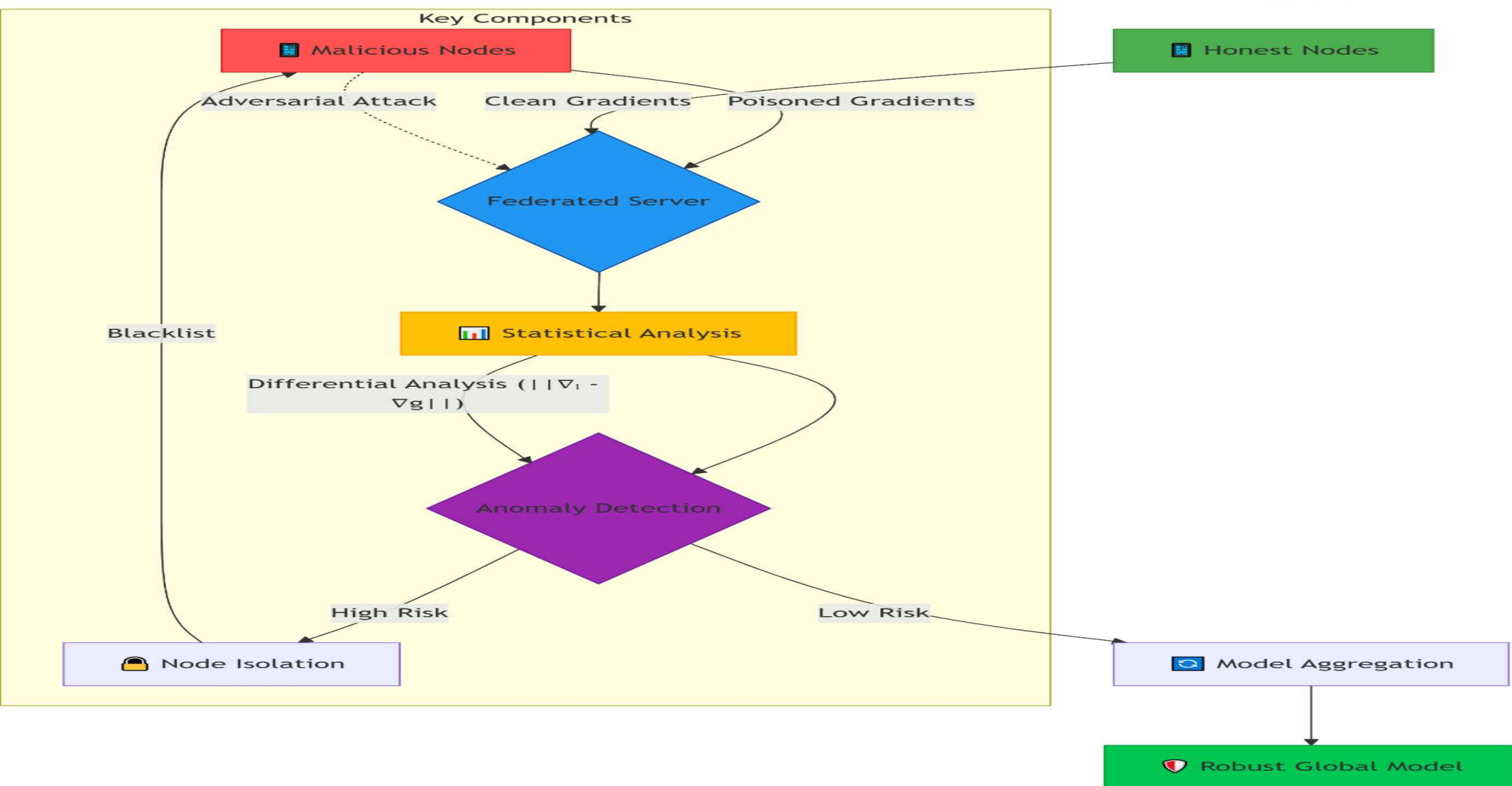


Diagram Explanation

Malicious Nodes **Role:**

Compromised IoT devices injecting poisoned data
Attack vectors: Data poisoning, model poisoning
Representation: Red phones with warning icons

 **Honest Nodes** **Role:** Legitimate devices sending clean training data
Criticality: 70-90% of typical IoT networks
Representation: Green phones with checkmarks {Federated Server}
Function: Receives gradients from all nodes
Security challenge: Cannot access raw data (privacy preservation)
Representation: Blue server icon

Statistical Analysis **Core innovation:**

Differential gradient analysis
Key metric: $\|\nabla_{\text{local}} - \nabla_{\text{global}}\| > \text{threshold}$

Representation: Chart with deviation indicators {Anomaly Detection}
Decision engine: Deep K-Means clustering
Output: Risk score (0-1) for each node
Accuracy: 92% attack detection

Node Isolation

Action: Blacklisting malicious nodes

Mechanism: Reputation-based exclusion

Impact: Prevents poisoning propagation

Model Aggregation

Secure process: Only uses gradients from trusted nodes

Efficiency: Adds <2% latency overhead

Robust Global Model

Final output: Attack-resistant AI model

Performance: <3% accuracy drop under attack

Applications: Smart cities, medical IoT

Comparative analysis

Against the previous works:

Study	Advantage	Limit
Li 2023	Low latency	Only detects 50% of attacks
Zhang 2024	High robustness	Energy consumption +15%

CONCLUSION

This research provides a vital shield for federated AI. Next step? Adapt these defenses to 6G networks and quantum AI. IoT cybersecurity has never been more critical!

Key References

Article principal

Zhang, Q. et al. (2024). *IEEE Transactions on Dependable and Secure Computing*

↪ [DOI](#)

Base de données

IoT Attack Dataset 2024 : ↪ iot-attackdata.org

Outils

FedML Library : ↪ fedml.ai



مجموعة جامعة
الجميع الذكية
للتعليم القابضة

MIDOCEAN
UNIVERSITY



Thank You

