Federated Learning Malware Detection Robustness Under Aggregation-Targeted Adversarial Attacks

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How robust are federated learning-based malware detection models to adversarial attacks targeting the aggregation process, measured by the degradation in F1-score when subjected to gradient poisoning. This work investigates the possibilities enabled by federated learning concerning IoT malware detection and studies security issues inherent to this new learning paradigm. In this context, a framework that uses federated learning to detect malware affecting IoT devices is. 5 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.1/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Federated Learning for Malware Detection in IoT Devices. Research question: How robust are federated learning-based malware detection models to adversarial attacks targeting the aggregation process, measured by the degradation in F1-score when subjected to gradient poisoning or model inversion attacks on the N-BaIoT dataset?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.1/10.

3 Results

14 papers retrieved. 5 claims extracted; 0 independently verified. Quality review score: 3.1/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Federated Learning enables data privacy by design as data is not shared with any external identity.	×	0.09
The exchange of the model parameters and their aggregation to create a unique and global model can be performed through	×	0.08
Previous works dealing with FL for intrusion detection lack the use of realistic datasets in the FL context.	×	0.06
The proposed framework covers both anomaly detection and classification approaches using multi-device datasets.	×	0.09
The benchmark table shows a 95% accuracy for benign device detection in centralized models.	×	0.03

References

- http://arxiv.org/abs/2104.09994v3
- http://arxiv.org/abs/2006.16545v1
- http://arxiv.org/abs/2210.00584v2