Non-IID Data Heterogeneity and Robustness of Personalized Federated Learning Against Label-Flipping Attacks

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How does varying the degree of data heterogeneity (non-IID) across clients impact the robustness of personalized federated learning models against label-flipping poisoning attacks in intrusion. Federated learning (FL) naturally faces the problem of data heterogeneity in real-world scenarios, but this is often overlooked by studies on FL security and privacy. On the one hand, the effectiveness of backdoor attacks on FL may drop significantly under non-IID scenarios. 13 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: Privacy Inference-Empowered Stealthy Backdoor Attack on Federated Learning under Non-IID Scenarios. Research question: How does varying the degree of data heterogeneity (non-IID) across clients impact the robustness of personalized federated learning models against label-flipping poisoning attacks in intrusion detection tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 16 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

16 papers retrieved. 13 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
Existing work [16] has shown that Federated Learning (FL) accuracy is significantly reduced in non-IID data scenarios.	×	0.14
As the number of specified victim classes increases in the SSBL attack, the Main Task Accuracy (MTA) remains almost unaf	×	0.03
As the number of specified victim classes increases in the SSBL attack, both the Attack Success Rate (ASR) and Backdoor	×	0.05
In the SSBL attack on MNIST, CIFAR10, and YAF datasets, only the specified classes are misclassified as the target when	×	0.07
The Attack Success Rate (ASR) of the SSBL method increases as the degree of data heterogeneity increases.	×	0.06
The increase in ASR with higher heterogeneity is inferred to be due to a decline in the accuracy of non-specified classe	×	0.05
Melis et al. [22] demonstrated that model updates in FL could leak unintended information about clients' training data.	×	0.06
It is possible for server-side attackers to reconstruct original training data from collected model updates due to leaka	×	0.09
Wang et al. [23] devised a framework incorporating a GAN with a multitask discriminator to simultaneously discriminate c	×	0.03
Zhang et al. [9] introduced a generative poisoning attack called PoisonGAN for scenarios where the attacker does not hav	×	0.06
Xie et al. [18] proposed utilizing a composite global trigger formed by several local triggers to conduct a distributed	×	0.07
Gong et al. [21] proposed using model-agnostic triggers to increase the attack success rate of Distributed Backdoor Atta	×	0.06
Zhang et al. [11] proposed a method called Neurotoxin to slightly modify model parameters during FL training to improve	×	0.05

References

- http://arxiv.org/abs/2306.15932v2
- http://arxiv.org/abs/2306.08011v1
- http://arxiv.org/abs/2503.20618v1