INTERNATIONAL MULTIDISCIPLINARY BOOK SERIES
https://ibseries.com/
Security and Privacy in AI Systems, Edited By: Dr. Sunita Chaudhary, Dr. Joydeb Patra

IMBS, Vol. 03, No. 1, (2025)
Open Access. Online & Print
ISSN:

CHAPTER 1:
ADVERSARIAL ATTACKS ON FEDERATED LEARNING MODELS IN HEALTHCARE DATA ECOSYSTEMS

Soumili Kundu
BALLB Programme, Brainware University

Abstract
Federated Learning (FL) has emerged as a transformative paradigm for privacy-preserving machine learning, particularly in healthcare ecosystems where sensitive patient data cannot be centrally aggregated. By enabling decentralized model training across hospitals, diagnostic centers, and wearable devices, FL addresses data privacy and regulatory constraints. However, despite its privacy advantages, FL is highly vulnerable to adversarial attacks due to its distributed and trust-based architecture. This paper provides a comprehensive analysis of adversarial threats in healthcare FL systems, focusing on data poisoning attacks, model poisoning attacks, and model inversion attacks. We examine how malicious participants manipulate local training processes to degrade global model performance or extract sensitive medical information. Furthermore, the paper explores attack vectors specific to healthcare, including medical imaging, electronic health records (EHRs), and IoT-based patient monitoring systems. A detailed comparison of attack mechanisms, impact severity, and defense strategies is presented. We also discuss state-of-the-art mitigation approaches such as robust aggregation, differential privacy, blockchain-based verification, and trust-aware learning frameworks. The study highlights critical research gaps and proposes future directions to enhance the robustness of federated healthcare systems against adversarial threats.

Keywords
Federated Learning, Healthcare AI, Adversarial Attacks, Data Poisoning, Model Poisoning, Model Inversion, Privacy Leakage, Secure Aggregation, Differential Privacy, Medical Data Security

1. Introduction

The increasing digitization of healthcare systems has led to the generation of vast amounts of sensitive medical data. Traditional centralized machine learning approaches are often infeasible due to privacy concerns, legal regulations (e.g., HIPAA, GDPR), and institutional data silos. Federated Learning (FL) addresses these challenges by enabling collaborative model training without sharing raw data.

However, FL introduces new vulnerabilities. Its decentralized structure allows malicious clients to inject manipulated updates, leading to compromised model integrity and privacy leakage. This is particularly critical in healthcare, where incorrect predictions can directly impact patient outcomes.

Recent studies demonstrate that FL systems are susceptible to poisoning attacks and privacy leakage mechanisms such as gradient inversion, which can reconstruct sensitive patient data from shared updates. These risks necessitate

ISBN (978-81-685212-9-2)

a deeper understanding of adversarial threats in medical FL ecosystems.

2. Federated Learning in Healthcare

2.1 Applications

- Medical image analysis (MRI, CT scans)
- Disease prediction using EHRs
- Drug discovery and genomics
- Remote patient monitoring (IoT devices)
- Healthcare FL enables cross-institutional collaboration while preserving patient confidentiality. However, heterogeneity in data distribution (non-IID data) and device reliability introduces additional security challenges.

3. Threat Model in Healthcare FL Systems

In FL, adversaries can act as:

- Malicious clients (insiders injecting poisoned updates)
- Curious servers (attempting to infer private data)
- External attackers (intercepting model updates)
- Attack objectives include:
- Degrading model accuracy
- Introducing backdoors
- Extracting private patient data

4. Poisoning Attacks in Federated Learning

4.1 Data Poisoning Attacks

Data poisoning involves injecting malicious or mislabeled data into local training datasets.
Example: Label flipping in medical images (e.g., benign tumor labeled as malignant)
Impact: Reduced diagnostic accuracy
FL is particularly vulnerable because:
Local datasets are not visible to the central server
Malicious updates are aggregated blindly
Research shows that poisoning attacks can significantly degrade model performance and introduce bias.

4.2 Model Poisoning Attacks

Model poisoning directly manipulates model parameters or gradients.
Attackers modify local updates before sharing
Can introduce stealthy backdoors without affecting overall accuracy
Optimization-based poisoning techniques can achieve high success rates while bypassing defense mechanisms.

4.3 Advanced Poisoning Techniques

Attack Type	Mechanism	Impact
Label Flipping	Mislabel training data	Accuracy degradation
Backdoor Attack	Embed hidden triggers	Targeted misclassification
Gradient Manipulation	Alter gradients	Model divergence
GAN-based Poisoning	Generate adversarial samples	High stealth attacks

Advanced methods such as hyperdimensional data poisoning can increase attack impact by up to 5–10× compared to traditional techniques.

5. Model Inversion Attacks in Healthcare FL

Model inversion attacks aim to reconstruct sensitive input data from shared gradients or model updates.

5.1 Gradient Inversion

Attackers reconstruct patient data (e.g., medical images) from gradients
Exploits information leakage in model updates
Studies show that gradient-based attacks can recover private data even without direct access to datasets.

5.2 Privacy Risks in Healthcare

Reconstruction of patient medical images
Exposure of genetic or diagnostic information
Violation of confidentiality laws

5.3 Attack Workflow

Collect gradients from FL updates
Optimize synthetic inputs to match gradients
Reconstruct original data

6. Adversarial Attacks in Medical FL Systems

Healthcare-specific vulnerabilities include:
Medical Imaging Systems: Sensitive to adversarial perturbations due to complex textures.
EHR Systems: Structured data susceptible to inference attacks
IoT Healthcare Devices: Limited security and computational power

7. Defense Mechanisms

7.1 Robust Aggregation Techniques

Krum, Trimmed Mean, Median-based aggregation
Detect and remove malicious updates

7.2 Differential Privacy (DP)

Adds noise to gradients
Limits data leakage but may reduce accuracy

7.3 Secure Multi-Party Computation (SMPC)

Encrypts model updates
Prevents direct access to gradients

7.4 Trust-Based Learning Frameworks

Assign trust scores to clients
Example: weighted aggregation models improve robustness against poisoning and inversion attacks (SSRN)

7.5 Blockchain-Based FL

Ensures transparency and tamper-proof updates
Useful for healthcare audit trails

8. Comparative Analysis of Attacks

Attack Type	Target	Goal	Severity	Detection Difficulty
Data Poisoning	Training Data	Reduce accuracy	Medium	Moderate
Model Poisoning	Model Updates	Backdoor insertion	High	High
Model Inversion	Gradients	Data reconstruction	Critical	Very High

9. Challenges in Securing Healthcare FL

Non-IID and heterogeneous data distribution
Limited computational resources in medical IoT devices
Trade-off between privacy and model accuracy
Lack of standardized security benchmarks

10. Future Research Directions

Hybrid defense mechanisms combining DP + blockchain
AI-driven anomaly detection for malicious clients
Privacy-preserving explainable AI in healthcare FL
Secure hardware integration (Trusted Execution Environments)

11. Conclusion

Federated Learning offers a promising solution for privacy-preserving healthcare analytics, but its decentralized nature introduces significant adversarial vulnerabilities. Poisoning attacks and model inversion attacks pose severe risks, including compromised model integrity and leakage of sensitive patient data. While various defense mechanisms exist, no single solution provides complete protection. Therefore, a multi-layered security approach is essential for deploying robust and trustworthy federated learning systems in healthcare environments.

References

1. Ma, W., Zhao, Q., & Tian, W. (2025). Defense against multi-label poisoning attacks in federated learning. Scientific Reports. (Nature)
2. Zhou, X., Xu, M., & Wu, Y. (2021). Deep model poisoning attack on federated learning. Future Internet. (MDPI)
3. Kasyap, H., & Tripathy, S. (2023). Hyperdimensional data poisoning attacks in FL. Expert Systems with Applications. (ScienceDirect)
4. Gupta, P., et al. (2023). Inverted loss function-based poisoning attack. Computers & Security. (ScienceDirect)
5. Singh, A. K., et al. (2023). Detection of poisoning attacks in FL. Data Mining and Knowledge Discovery. (PMC)
6. Al-Matari, M. R., et al. (2025). FedDefend++ framework for FL security. SSRN. (SSRN)
7. Kalapaaking, A. P., et al. (2023). Blockchain-based FL for healthcare security. arXiv. (arXiv)
8. Information Sciences (2023). Gradient inversion and privacy leakage in FL. (ScienceDirect)
9. Journal of Electrical Systems (2024). Model poisoning challenges in FL. (Journal of Electrical Systems)
10. ScienceDirect (2024). Adversarial risks in medical image-based FL systems. (ScienceDirect)