Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

A Study of Security and Privacy Challenges in Federated Learning System and Their Defense Mechanisms

Ms. Shital Jadhav, Dr. Reena Shinde
Dept. of Computer Science, Sinhgad College of Science, Pune, India

Abstract - Federated Learning (FL) allows different participants to train a shared model without sharing their raw data. This approach is crucial in fields where privacy is important such as healthcare, finance and the Internet of Things (IoT). However, recent research reveals that federated learning faces several privacy and security risks. These risks include gradient leakage, inference attacks, model poisoning and communication interception. This paper thoroughly examines these issues and reviews the defense strategies proposed in existing literature. It discusses a defense strategy that includes cryptographic protection, differential privacy, strong aggregation, secure communication and monitoring client trust. The study emphasizes that federated learning is not inherently secure and requires coordinated defenses to ensure safe use.

Keywords-Federated learning, model attacks, privacy preservation, secure aggregation, differential privacy.

1. INTRODUCTION:

A new approach called Federated Learning (FL) is proposed in which a number of devices or systems work together but do not share the actual data. Traditionally the data of all the participating users is aggregated at a single point called the server and then the model is trained. A major concern of privacy persists, especially when the domain of interest is health, finance or personal mobile usage. On the contrary in FL the model is trained individually and the changes made are communicated and aggregated at a single point.[1][2]

Even though FL has improved the level of privacy, it is still not fully secure. There have been various studies done to prove that the attackers have the ability to get to know certain aspects of the private information from the shared model updates. For example, even though the raw data has not been shared with gradient leakage attacks and membership inference attacks the attackers get to know certain information regarding the training data[3][4] similarly model inversion attacks can reconstruct data from model updates[5] These issues critically highlights privacy risks in federated learning.

2. OBJECTIVES

1. Identify and Categorize Key security threats in federated learning

2. Examine privacy risk coming from shared updates despite of data remains local.

3. Analyze how attacks exploit model gradients.

4. Evaluate effectiveness of defense mechanisms.

3. SECURITY AND PRIVACY CHALLENGES IN FEDERATED LEARNING

Federated Learning (FL) which enables many users or organizations to collaboratively train a machine learning model without sharing their raw data has recently gained popularity. Specifically, with FL each user shares their model update with a central server such as gradients or weights, rather than their raw data raising the privacy of this model compared to traditional centralized learning. Literature has confirmed though, that FL is vulnerable to various severe security risks.

3.1 Gradient Leakage

Although the actual data is not being shared the attacker could possibly get a chance to look at the gradients or models of data being shared by other users. Sometimes the actual data used for the machine learning model could be leaked using the gradients or models of the data being shared. This type of attack is called a gradient leakage attack.[3]

3.2 Membership Inference Attacks:

In this kind of attacks attackers try to find out whether specific data was used in training process or not by observing model behavior and output. Attackers guess certain data (patient data) was part of training dataset[4]

3.3 Model Inversion Attacks:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

In this attacker uses model outputs to reconstruct sensitive features of specific data. For example it may possible to recreate an image and recover private attributes from model predictions [5]

3.4 Poisoning Attacks:

In Federated Learning it was assumed that all participants will act honestly. However in reality any malicious client could intentionally send manipulated updates with the intent of corrupting the global model. This is what is referred to as a poisoning attack. It can either drop model accuracy or cause it to behave incorrectly for certain inputs.[6]

3.5 Backdoor Attacks:

In this type of attack an adversarial member will train the model in such a way that the model behaves as normal under most circumstances but will perform incorrectly for a specific trigger input. This is an important threat as it cannot be easily spotted.[6]

3.6 Communication Attacks:

Since the model updates sent during the transmission process take place over networks an attacker can intercept and/or monitor these network transmissions if they are not specially protected. This can result in leakage as well as disruption.[7]

3.7 Insecure Aggregation:

In Federated Learning updates from all participants are aggregated on a central server. If this process is not secure malicious attackers may have access to intermediate updates and extract private information or manipulate model training.[8]

The federated learning challenges generally split into two criteria's Security and Privacy Security related to integrity and Privacy related to Confidentiality.

Threat Matrix Privacy Vs. Security

Threat Category	Challenges	Attacker	Goal of Attacker
Privacy	Gradient Leakage	Honest but Curious Server	Reconstruct raw user data from weights
	Membership Inference	Any participant/Server	Identify specific record was in training dataset
	Model Inversion	External User/Server	Reconstruct sensitive features or attributes of data
Security	Poisoning Attacks	Malicious Client	Degrade Model Accuracy
	Backdoor Attacks	Malicious Client	Misclassification
	Communication Attacks	Man-in-Middle	Alter updates during transit

Table 1. Threat Matrix Privacy Vs Security

4. DEFENSE MECHANISM IN FEDERATED LEARNING

Since there exists a leakage of privacy in the Federated Learning (FL) systems, defense mechanisms have been developed to ensure that the learning process is adequately protected, thereby safeguarding the user data. This implies that the FL systems are made more secure, reliable, and of high accuracy, while protecting user privacy.

4.1 Differential Privacy:

In Differential Privacy noise is added to the model updates prior to them being sent to the server. The added noise is such that no single record within the data is identifiable. Hence it is extremely difficult for a hostile user to attempt a membership inference attack. DP promises a mathematical guarantee of privacy and the value is represented by epsilon (ε).[9]

4.2 Secure Aggregation:

The secure aggregation also provides the guarantee that the server will not be able to see the individual participant's model updates. The only thing that server will see is the aggregated result of the entire model. This means the server or any malicious user will not be able to collect anything from the specific participant's data from the update.[8]

4.3 Homomorphic Encryption:

Homomorphic encryption enables computations to be carried out directly over the data. In FL, data can be encrypted by the data contributor before sending it to the server. This will keep the data private even during computation on the server side.[10]

4.4 Secure Multiparty Computation (SMPC)

SMPC enables a number of participants to collectively compute a function e.g. model aggregation, without the individual inputs of every participant being revealed to the others. The privacy property holds, even if the other clients or the server are not trusted.[11]

4.5 Robust Aggregation Methods

To defeat the effort of poisoning and backdoor attacks, aggregation methods such as Krum, Median and Trimmed Mean are employed. These methods detect malicious

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

updates to reduce their effects before building the global model.[11]

4.6 Gradient Clipping

It limits the size of model updates sent by participants. This reduces chance of gradient leakage and prevent malicious client by sending extremely large object to disrupt the system [9]

4.7 Authentication

Using secure communication protocol TSL/SSL and client authentication ensure only trusted client join the federated learning process.[7]

The following defenses are frequently employed to make Federated Learning safe and considerate of privacy:

- Differential privacy to hide personal information.
- Safe aggregation to safeguard local updates.
- SMPC and homomorphic encryption for computation that protects privacy.
- These mechanisms together help in building trustworthy FL systems suitable for sensitive. applications like healthcare and finance.

Comparative Analysis of Defense Mechanism

Table 2. Comparative analysis of Defense Mechanism

Defense Mechanism	Threats	Privacy/Securiy Strength	Impacon Model Accuracy	Communication/Computation Overhead
Differential privacy	Gradient Leakage,MembeshipInference	High	High	Low
Secure Aggregation SMPC	Data Reconstruction,Server Side Leakage	High	No	Moderate
Homomorphic Encryption	Insecure Aggregation,ServerSnooping	Very High(Cryptographic)	No	High(High latency and data size)
Robust Aggregation	Poisoningand Backoor Attack	High	Moderate	Low to Moderate
Autentication	Unauthorized acess	Moderat	No	Low

5. METHODOLOGY

This Methodology evaluates privacy and security risk in federated learning and test how different defenses reduce risk.

1. **Federated Learning Simulation Setup**
   - One central server and multiple clients with local dataset.
   - Client train model locally on their dataset and send gradients to server.
   - Implement FedAvg in PyTorch/Tensor flow.
2. **Threat Modeling**
   - Simulate common attacks on model updates
3. **Defense Integration**
   - Apply defenses one by one.
4. **Evaluation Metrics**
   - Model Accuracy.
   - Attack Success rate(ASR).
   - Privacy Budget(€).
   - Communication and Computation overhead
5. **Experimental Procedure**
   - Train FL model without defense measure ASR and Accuracy.
   - Apply each defense individually and re-measure.
   - Apply layered defense together(compare result)
6. **Comparative Analysis**
   - Analyze which defense give good privacy with minimum accuracy loss and acceptable overhead.

6. DISCUSSION

By keeping raw data local federated learning enhances privacy but still the shared gradients and weights introduce a new leakage surface.

Observations:

- Even a sincere but curious server can reconstruct user data from gradients.
- Attackers can use model outputs (membership inference) to assess whether a patient's record was used.
- Malicious clients can secretly alter the model (poisoning/backdoor) without being easily detected.
- Unprotected communication and aggregation allow for the interception or modification of updates.

Trade-off realization:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

No one mechanism is adequate. Robust aggregation does not prevent inference attacks and privacy measures (DP, Secure Aggregation) do not prevent poisoning.

A multi-layered defense is required.

Thus, a feasible secure FL pipeline consists of:

Gradient Clipping → Secure Aggregation/SMPC → Robust Aggregation → TLS/Authentication → Differential Privacy

This multi-layered strategy lowers Attack Success Rate (ASR) while maintaining acceptable model accuracy and controllable overhead, which is crucial for the financial and healthcare industries.

7. CONCLUSION

In Federated Learning the data remains at the user's device. However, the sharing of updates can lead to the leakage of sensitive information and attacks such as gradient leakage, membership inference, poisoning and backdoors. The use of Differential Privacy,Secure Aggregation, robust aggregation, gradient clipping and secure communication can help ensure both privacy and integrity of the model. The takeaway is that there is no single solution that works and we need multiple solutions to ensure the safety of Federated Learning, especially in areas such as healthcare, Finance.

REFERENCES

[1] Konečný, J., et al. (2016). *Federated Learning: Strategies for Improving Communication Efficiency*. arXiv:1610.05492.
[2] McMahan, H. B., et al. (2017). *Communication-Efficient Learning of Deep Networks from Decentralized Data*. AISTATS.
[3] Zhu, L., & Han, S. (2020). *Deep Leakage from Gradients*. NeurIPS.
[4] Melis, L., et al. (2019). *Exploiting Unintended Feature Leakage in Collaborative Learning*. IEEE S&P.
[5] Fredrikson, M., et al. (2015). *Model Inversion Attacks that Exploit Confidence Information*. CCS.
[6] Bagdasaryan, E., et al. (2020). *How To Backdoor Federated Learning*. ICML.
[7] Bhagoji, A. N., et al. (2019). *Analyzing Federated Learning through an Adversarial Lens*. ICML Workshop.
[8] Bonawitz, K., et al. (2017). *Practical Secure Aggregation for Privacy-Preserving Machine Learning*. CCS.
[9] Dwork, C., & Roth, A. (2014). *The Algorithmic Foundations of Differential Privacy*. Foundations and Trends in Theoretical Computer Science.
[10] Aono, Y., et al. (2017). *Privacy-Preserving Deep Learning via Additively Homomorphic Encryption*. IEEE TIFS.
[11] Yin, D., et al. (2018). *Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates*. ICML.