IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.16

Secure Federated Learning Framework for Privacy-Preserving AI Models

Prof.K.Ashok Kumar
Associate Professor, Department of Computer Science and Engineering (Cyber security)
Sengunthar Engineering College (Autonomous),Tiruchengode, India
kashokkumar.cse@scteng.co.in, csecshod@scteng.co.in
https://orcid.org/0009-0008-6399-423X
Sonali J, Ambika Kumari, Abinaya K
UG Student, Department of Computer Science and Engineering (Cyber security)
Sengunthar Engineering College (Autonomous), Tiruchengode, India
sonalisai2021@gmail.com, ambikar021@gmail.com, abiaby307@gmail.com

Publication History
Manuscript Reference: IRJCS/RS/Vol.13/Issue03/CSMR26.MRCS10095
Research Article | Open Access | Double-Blind Peer Reviewed Article ID: IRJCS/RS/Vol.13/Issue03/CSMR26.MRCS10095
Received: 30, January 2026, Revised: 13, February 2026, Accepted: 28 February 2026 Published Online: 25 March 2026
https://www.irjcs.com/volumes/Vol13/iss-03/16.CSMR26.MRCS10095.pdf
Article Citation:Prof.Ashok,Sonali,Ambika,Abinaya(2026),A Secure Federated Learning Framework Incorporating Zero Trust Architecture and Differential Privacy, IRJCS: International Research Journal of Computer Science, Volume 13,Issue 03 of 2026 pages 191-195 **Doi:>** https://doi.org/10.26562/irjcs.2026.v1303.16
BibTeX Key Prof.Ashok@2026Secure
Orcid: https://orcid.org/0009-0004-9398-7488
IRJCS papers should be cited as IRJCS (International Research Journal of Computer Science, AM Publications, India 2026, ISSN 2393-9842, https://doi.org/10.26562/ijrcs.2025.v1303.16 The journal's official abbreviation is IRJCS.

Abstract: Federated Learning (FL) enables decentralized de- vices to collaboratively train machine learning models without sharing raw data. Despite its privacy advantages, FL systems remain vulnerable to adversarial threats such as model poisoning, gradient inversion, and Byzantine attacks. This paper proposes a secure federated learning framework that integrates Zero Trust Architecture (ZTA) and Differential Privacy (DP) to mitigate these vulnerabilities. The proposed system verifies every client update using trust scoring and anomaly detection mechanisms. Differential Privacy is applied through gradient clipping and Gaussian noise injection to prevent information leakage. Experimental evaluation using the MNIST dataset demonstrates that the proposed model achieves strong robustness against label-flipping attacks while maintaining high model accuracy.
Index Terms: Federated Learning, Differential Privacy, Zero Trust Architecture, Cyber Security, Model Poisoning

I. INTRODUCTION

With the rapid growth of edge computing and IoT devices, massive amounts of sensitive data are generated across distributed networks. Traditional centralized machine learning approaches require transferring data to a central server, raising serious privacy concerns. Federated Learning (FL) addresses this challenge by al- lowing clients to train models locally and share only model updates with the server. However, FL introduces new security risks including:

- Model Poisoning Attacks
- Data Poisoning Attacks
- Gradient Inversion Attacks
- Byzantine Failures

To address these threats, this research proposes a secure federated learning framework combining:

- Zero Trust Architecture
- Differential Privacy
- Trust Score Monitoring

II. RELATED WORK

Several studies have explored security enhancements for federated learning. Mc Mahan et al. Introduced the Federated Averaging algorithm which allows decentralized training of neural Networks. Abadi et al. proposed Differential Privacy for deep learning to prevent data leakage during training. Recent works also explore robust aggregation techniques such as Krum and Median aggregation to resist adversarial gradients. However, the integration of Zero Trust security principles with FL remains relatively underexplored.

III. SYSTEM ARCHITECTURE

The proposed system consists of three major components:

1) Federated Clients
2) Central Aggregation Server
3) Security Monitoring Dashboard

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.16

Each client trains the global model locally using private data and sends the update to the central server. Before aggregation, the server performs security validation using trust scoring and anomaly detection.

IV. DIFFERENTIAL PRIVACY MECHANISM

To protect client data, gradient updates are modified using clipping and Gaussian noise.
The clipped gradient is defined as:

$$\| _{C} \| = \max(1, \frac{g_i}{2}) \quad (1)$$

Without defense mechanisms, model accuracy dropped to nearly 45% under attack.
Noise is then injected to the update:

$$\Delta w = \frac{1}{k} \Sigma \tilde{g} + N(0, \sigma^2 C^2 (I)$$

(2) This ensures that the training process satisfies (ϵ, δ) differential privacy.

ZERO TRUST EVALUATION

The Zero Trust principle assumes that no client should be automatically trusted. Each client is assigned a trust score initially set to: $T_k = 100$ If suspicious activity is detected, the trust score decreases.

An anomaly is detected if: $||\Delta w_k||_2 > \tau \times M$ (3)

where M is the median magnitude of all received gradients. Clients whose trust score falls below a threshold are removed from the training process.

FEDERATED LEARNING ALGORITHM

With the proposed Zero Trust mechanism, malicious clients were detected within five rounds and removed from the system.
The model achieved:

- 98.2% accuracy (baseline)
- 96.5% accuracy (under attack)
- 92.4% accuracy with differential privacy

These results demonstrate that the frame work effectively maintains model performance while improving security.

MONITORING DASHBOARD

A real-time monitoring dashboard was developed to visualize the federated learning process.
The dashboard includes:

- Client Trust Score Table
- Accuracy Progress Graph
- Security Event Logs

Server-Sent Events (SSE) were used to stream real-time updates from the backend to the dashboard.

MATHEMATICAL SECURITY ANALYSIS

To further analyze the robustness of the proposed frame- work, we evaluate the behavior of adversarial gradients in the federated learning environment. Let Δw_k represent the update vector submitted by client k. In a standard federated learning scenario, benign updates fol

Algorithm 1 Secure Federated Learning with Zero Trust Initialize global model w_0
Set trust score $T_k = 100$
For each round t do
Distribute model to clients Clients perform local training Apply differential privacy Send updates Δw_k
Evaluate trust score
If $T_k < T$ critical then
Reject update
endif
Aggregate trusted updates
End for

EXPERIMENTAL SETUP

The proposed framework was implemented using Python and PyTorch. The MNIST dataset was used to evaluate system performance. Ten clients were simulated with both IID and Non-IID data distributions. Twenty percent of the clients were configured as malicious adversaries performing label-flipping attacks. low a distribution centered around the true gradient direction. However, adversarial clients may attempt to manipulate the model by scaling their gradients. A common attack is the gradient scaling attack: $\Delta w_{adv} = \alpha \cdot \Delta w_{benign}$ (4) where α is a scaling factor such that $\alpha > 1$. The magnitude of the adversarial gradient therefore becomes $||\Delta w_{adv}||_2 = \alpha ||\Delta w_{benign}||_2$ (5) To detect such manipulations, the server computes the median magnitude of all received updates in round t: $M = Median(\{||\Delta w_k||_2\})$ (6) If the gradient magnitude of a client significantly deviates from this median value, the update is flagged as suspicious. $||\Delta w_k||_2 > \tau \times M$ (7) where τ represents a configurable anomaly threshold. This detection method allows the system to identify abnormal updates while preserving contributions from honest clients.

TRUSTS CORE ADAPTATION MODEL

The trust scoring mechanism dynamically evaluates client behavior across multiple communication rounds.

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.16

Each client k is assigned an initial trust score: $T_k=100$ (8)

If an anomaly is detected during round t, the trust score is penalized according to $T_k=T_k-\Delta P$ (9)

Where ΔP represents the penalty factor.

If the update is validated as benign, the trust score recovers gradually: $T_k=min(100,T_k+\Delta R)$ (10)

Where ΔR is a reward increment.

Clients who set rusts core falls below the critical threshold $T_{critical}$ are excluded from the aggregation process.

$T_k \leq T_{critical}$ (11)

This mechanism enables the system to continuously evaluate client behavior and prevent repeated malicious contributions.

COMMUNICATION EFFICIENCY

Federated learning systems must maintain efficient communication between clients and the central server.

Assume the model contains P parameters and there are N participating clients.

The communication cost per round is:

$C_{round}= 2NP$ (12) The factor of 2 accounts for:

- Server distributing model weights
- Clients returning gradient updates

By applying differential privacy locally, the proposed frame work avoids additional encryption overhead that may significantly increase communication latency.

COMPUTATIONAL COMPLEXITY ANALYSIS

The computational complexity of the proposed frame work is analyzed for both client and server components.

A. Client Side Complexity

Local model training dominates the client computation cost. If E represents local epochs, B represents batch size, and Pre presents the number of model parameters, the training complexity becomes:

$$O(E\times B\times P) \quad (13)$$

Differential privacy introduces additional overhead for gradient clipping and noise injection, but this overhead is negligible compared to neural network training.

B. Server Side Complexity

The server performs the following operations:

- Gradient magnitude computation
- Median calculation
- Trusts core update
- Aggregation

Computing gradient norms requires $O(NP)$ (14)

Sorting operations for median calculation require $O(N\log N)$ (15)

Thus, the total server complexity per round remains manageable even for large federated networks.

EXTENDED ATTACK EVALUATION

Beyond label-flipping attacks, the system was evaluated against several additional adversarial strategies.

A. Gradient Scaling Attack

In this attack, malicious clients artificially amplify their gradient updates to dominate the aggregation process. The Zero Trust defense mechanism successfully detects such anomalies through gradient magnitude analysis.

B. Random Gradient Attack

Malicious clients may submit completely random gradients in order to destabilize the training process. Because these updates significantly deviate from the median gradient direction, they are quickly detected and discarded.

C. Back door Injection Attack

Attackers may attempt to introduce hidden triggers into the model by manipulating their local datasets. While this attack can be subtle, abnormal gradient patterns eventually reduce the trust score of malicious clients.

ABLATION STUDY

An ablation study was conducted to evaluate the contribution of each component of the proposed system. Four configurations were tested:

- Standard Federated Learning
- Federated Learning + Differential Privacy
- Federated Learning + Zero Trust
- Proposed Secure Framework

The results show that the complete framework achieves the best balance between security and model accuracy.

Table I – Ablation Study Results

Configuration	Accuracy Under Attack
Standard FL	42%
FL+Differential Privacy	55%
FL+ Zero Trust	94%
Proposed Frame work	96.5%

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.16

PRACTICAL DEPLOYMENT CONSIDERATIONS

Deploying secure federated learning systems in real-world environments requires careful consideration of several factors.

A. Edge Device Limitations

Many federated clients are resource-constrained devices such as smart phones or IoT sensors. Efficient model architectures and optimized training procedures are required to ensure compatibility.

B. Network Reliability

Federated networks may experience intermittent connectivity. The framework must support asynchronous training and fault tolerance to handle client dropouts.

C. Regulatory Compliance

Privacy-preserving machine learning techniques such as differential privacy are essential for compliance with data protection regulations including GDPR and HIPAA.

FUTURE RESEARCH DIRECTIONS

Several improvements can further enhance the proposed framework.

- Secure Multi-Party Computation for encrypted aggregation
- Homomorphic Encryption for privacy-preserving gradient updates
- Block chain-based trust management for decentralized verification
- Adaptive differential privacy mechanisms
- Advanced anomaly detection using deep learning models. These extensions will strengthen the security and scalabilityoffederatedlearningsystemsacrosslarge-scaledistributed networks.

REAL-WORLD APPLICATIONS AND CASE STUDIES

Secure federated learning frameworks have significant applications in domains where sensitive data is distributed across multiple organizations or edge devices. The integration of Zero Trust security principles and Differential Privacy ensures that collaborative learning can occur without exposing confidential information.

A. Healthcare Data Collaboration

One of the most promising applications of federated learning is in the healthcare sector. Hospitals and medical institutions generate vast amounts of patient data including medical images, diagnostic records, and electronic health records (EHR). Due to strict regulatory requirements such as HIPAA and GDPR, sharing raw patient data between hospitals is often prohibited. Federated learning allows multiple healthcare institutions to collaboratively train diagnostic models without exchanging sensitive patient records. Each hospital trains the model locally and sends only encrypted updates to the central server. The addition of Differential Privacy further ensures that individual patient records cannot be reconstructed from the transmitted gradients. Meanwhile, the Zero Trust architecture prevents compromised institutions from introducing malicious updates into the global model.

B. Financial Fraud Detection

Financial institutions frequently collaborate to detect fraudulent transactions across multiple banking networks. However, direct sharing of transaction data may violate privacy regulations and expose sensitive financial information. By using federated learning, banks can jointly train fraud detection models while keeping transaction data within their local infrastructure. The proposed framework enhances this process by verifying the integrity of each participating bank through trust scoring. If any participating node behaves abnormally, its updates are rejected, preventing manipulation of fraud detection models.

C. Smart City and IoT Security

Smart city infrastructure relies heavily on distributed IoT devices such as traffic sensors, surveillance cameras, and environmental monitoring systems. These devices continuously generate large volumes of data that can be used to train intelligent systems for traffic prediction, anomaly detection, and public safety monitoring. Federated learning enables these devices to collaboratively improve machine learning models while maintaining data locality. However, IoT networks are often vulnerable to device compromise and malware attacks. By incorporating Zero Trust security principles, the proposed system continuously monitors device behavior and removes suspicious nodes from the training process.

D. Autonomous Vehicle Networks

Autonomous vehicles rely on machine learning models for object detection, navigation, and decision-making. Training these models requires massive datasets collected from vehicle sensors including cameras, LiDAR, and radar. Sharing raw sensor data between manufacturers or vehicles raises privacy and proprietary concerns. Federated learning allows vehicles to share model updates instead of raw driving data. The proposed framework ensures that malicious vehicles cannot inject poisoned updates into the shared model. This approach enables safer collaborative learning across autonomous vehicle fleets.

E. Industrial AI and Manufacturing

Modern manufacturing environments use machine learning to monitor production lines, detect equipment failures, and optimize operational efficiency. Factories located in different geographical regions may wish to collaborate in improving predictive maintenance models. Using secure federated learning, each factory can train models on its local sensor data while sharing only model updates with a central coordinator. The integration of Zero Trust verification ensures that compromised industrial nodes cannot manipulate the learning process. These applications demonstrate that the proposed framework is suitable for large-scale deployment in privacy-sensitive and security-critical domains.

ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.16

CONCLUSION

This paper presented a secure federated learning framework that integrates Zero Trust Architecture and Differential Privacy. The system successfully detects malicious clients while protecting private training data. Experimental results demonstrate strong resilience against poisoning attacks with minimal loss in model accuracy. Future work will explore secure multi-party computation and block chain-based trust management to further enhance federated learning security.

REFERENCES

1. H.B.McMahan,E.Moore,D.Ramage,S.Hampson,andB.A.yArcas,"Communication efficient learning of deep networks from decentralized data," in Proc. Artificial Intelligence and Statistics (AISTATS), 2017.
2. M.Abadi,A.Chu,I.Good fellow, H. B. McMahan, I. Mironov, K.Talwar,andL.Zhang, "Deep learning with differential privacy," in Proc.ACM SIGSAC Conference on Computer and Communications Security,2016.
3. J.Geiping, H.Bauermeister, H. Dro¨ge, and M. Moeller, "Inverting gradients: How easy is it to break privacy in federated learning?," in Advances in Neural Information Processing Systems, 2020.
4. P.Blanchard, E.M.El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in Advances in Neural Information Processing Systems, 2017.
5. S.Rose,O.Borchert,S.Mitchell,andS.Connelly,"Zero Trust Architecture," NIST Special Publication 800-207, National Institute of Standards and Technology, 2020.
6. B.Hitaj,G.Ateniese,andF.Perez-Cruz, "Deep models under the GAN: Information leakage from collaborative deep learning," in Proc. ACMSIGSAC Conference on Computer and Communications Security, 2017.
7. K.Bonawitz et al., "Practical secure aggregation for privacy-preserving machine learning," in Proc .ACMSIGSAC Conference on Computer and Communications Security, 2017.
8. L.Melis, C.Song, E.De Cristofaro, and V. Shmatikov, "Exploiting un-intended feature leakage in collaborative learning," in IEEE Symposium on Security and Privacy, 2019.
9. N.Papernot et al.,"Semi-supervised knowledge transfer for deep learning from private training data,"in International Conference on Learning Representations (ICLR), 2017.
10. R.Shokri and V.Shmatikov, "Privacy-preserving deep learning," in Proc.ACM SIGSAC Conference on Computer and Communications Security,2015.
11. Y. Zhao et al., "Federated learning with non-IID data," arXiv preprintarXiv:1806.00582, 2018.
12. T.Li, A.K. Sahu, A.Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 50–60, 2020.
13. Q.Yang,Y.Liu,T.Chen,and Y.Tong, "Federated machine learning: Concept and applications, "ACM Transactions on Intelligent Systems and Technology, vol. 10, no. 2, 2019.
14. C.Dwork and A. Roth, "The algorithmic foundations of differential privacy," Foundations and Trends in Theoretical Computer Science,2014.
15. K.Gai, M.Qiu, and H.Zhao, "Privacy-preserving data aggregation in edge computing," IEEE Internet of Things Journal, 2019.
16. S.Kairouz et al.,"Advances and open problems in federated learning," Foundations and Trends in Machine Learning, vol.14, no.1–2, 2021.
17. A.G.Dimakis et al., "A survey on federated learning systems: Vision, hype and reality, "IEEE Communications Surveys and Tutorials, 2021.
18. Y.Liu,X.Chen,and Q.Yang,"Secure federated learning: A survey," IEEE Intelligent Systems,2020. Z.Linetal., "Defense against adversarial attacks in federated learning," IEEE Transactions on Neural Networks and Learning Systems,2021.
19. X.Wang,Y.Han,and V.C.Leung, "Convergence of federated learning in distributed networks," IEEE Transactions on Wireless Communications, 2020.