International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:3 Volume No.10 May - June – 2026
DOI:10.46647/ijetms.2026.v10i03.028 ISSN: 2581-4621

Partner in crime: Boosting Targeted Poisoning Attacks against Federated Learning

[1]Dr. K. Pavan kumar, [2]Boya Surendra
[1]*Associate Professor, Dr.K.V.Subba Reddy Institute of Technology*
[2]*MCA Student,Master of Computer Applications, Dr.K.V.Subba Reddy Institute of Technology*

ABSTRACT
This paper presents a novel adversarial framework, Partner in Crime, designed to amplify the effectiveness of targeted poisoning attacks against Federated Learning (FL) systems by exploiting collaborative malicious behavior among compromised clients. While traditional targeted poisoning attacks in FL typically rely on a single adversarial participant to manipulate the global model toward attacker-chosen misclassifications, their impact is often constrained by aggregation defenses, limited local influence, and the distributed nature of FL training. To overcome these limitations, the proposed approach introduces a coordinated multi-client attack strategy in which malicious participants act as "partners in crime" to jointly optimize poisoning objectives while remaining stealthy under standard FL aggregation mechanisms. The framework enables compromised clients to collaboratively inject carefully crafted poisoned updates that reinforce one another, thereby increasing attack success rates on specific target classes without significantly degrading overall model utility. By synchronizing gradient manipulation, local model drift, and poisoning intensity across multiple adversaries, the attack achieves stronger targeted misclassification while evading common anomaly detection and robust aggregation defenses. The proposed method is evaluated under diverse FL settings, including varying numbers of malicious clients, data heterogeneity levels, and defense strategies, demonstrating that coordinated poisoning substantially outperforms isolated attacks in both stealth and effectiveness. Experimental results show that Partner in Crime significantly increases targeted attack success rates while maintaining competitive clean-task performance, exposing critical vulnerabilities in collaborative learning systems. This work highlights the urgent need for stronger adversary-aware defenses in Federated Learning and provides important insights into the security risks posed by coordinated malicious participants in decentralized model training.
Keywords: Federated Learning, Targeted Poisoning Attack, Data Poisoning, Model Poisoning, Adversarial Machine Learning, Collaborative Attack, Malicious Clients, Byzantine Attack, Secure Aggregation, Robust Federated Optimization, Distributed Learning Security, Privacy-Preserving Machine Learning.

I. INTRODUCTION

Federated Learning (FL) has emerged as a powerful distributed machine learning paradigm that enables multiple clients to collaboratively train a global model without sharing raw local data, thereby preserving privacy and reducing centralized data risks. Its decentralized nature has made FL highly suitable for privacy-sensitive applications such as healthcare, finance, mobile intelligence, and edge computing. However, despite its privacy-preserving advantages, FL remains vulnerable to security threats due to its reliance on client-provided model updates. Among these threats, poisoning attacks are particularly dangerous because malicious clients can manipulate local training to corrupt the global model. Targeted poisoning attacks are especially concerning, as they aim to force the model to misclassify specific attacker-chosen inputs while preserving overall model accuracy, making them difficult to detect and highly effective in practical deployments.

Traditional targeted poisoning attacks in FL often assume a single malicious client, but such attacks are frequently limited by update averaging, client heterogeneity, and robust aggregation defenses. To overcome these limitations, the Partner in Crime framework introduces a coordinated multi-client attack strategy in which multiple compromised participants collaboratively inject malicious updates to strengthen targeted poisoning outcomes. By synchronizing gradient manipulation, poisoning intensity, and model drift, these malicious clients reinforce one another's influence while remaining

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:3 Volume No.10 May - June – 2026
DOI:10.46647/ijetms.2026.v10i03.028 ISSN: 2581-4621

stealthy under standard detection mechanisms. This coordinated attack significantly improves target misclassification success without severely affecting clean-task performance, exposing a critical vulnerability in Federated Learning systems and emphasizing the need for stronger defenses against collusive adversarial behavior.

II. LITERATURE SURVEY

1. Title: Partner in Crime: Boosting Targeted Poisoning Attacks against Federated Learning
Authors: Shihua Sun, Shridatt Sugrim, Angelos Stavrou, Haining Wang
Abstract: This paper introduces *BoTPA* (Boost Targeted Poisoning Attacks), a generalized pre-training strategy designed to strengthen targeted poisoning attacks in Federated Learning (FL). The method constructs an Amplifier set by leveraging model update contributions from non-target classes and falsifying their labels before FL training begins. This coordinated poisoning mechanism significantly increases attack success while remaining compatible with both data and model poisoning strategies. Experimental evaluation shows substantial gains in attack success rate across multiple datasets and defense settings, demonstrating that coordinated malicious behavior can significantly amplify targeted poisoning effectiveness in FL.

2.Title: Towards Multi-Party Targeted Model Poisoning Attacks against Federated Learning Systems
Authors: Chulin Xie, Minghong Fang, Owen Liu, Neil Zhenqiang Gong
Abstract: This work investigates a coordinated multi-party targeted model poisoning attack in Federated Learning, where multiple compromised clients collaboratively manipulate the global model to misclassify attacker-chosen samples. Unlike traditional single-client attacks, this approach enables adversaries to jointly optimize malicious updates while preserving stealth under common defense mechanisms. The authors propose a boosting strategy to solve the update scaling problem and incorporate stealth metrics such as cosine similarity and clustering accuracy to evade detection. Experimental results show that coordinated malicious clients can achieve highly effective targeted attacks while maintaining model convergence and evading robust aggregation defenses.

3.Title: Data Poisoning Attacks Against Federated Learning Systems
Authors: Vale Tolpegin, Stacey Truex, Mehmet Emre Gursoy, Ling Liu
Abstract: This paper presents one of the earliest systematic studies of targeted data poisoning attacks in Federated Learning. The authors demonstrate that malicious participants can poison local training data through label manipulation and significantly degrade the performance of the global model on attacker-selected target classes. The study shows that even a small fraction of malicious clients can induce substantial drops in classification accuracy while preserving performance on non-targeted classes, making the attack difficult to detect. The paper also analyzes attack timing, persistence, and malicious client participation, and proposes a defense strategy for identifying poisoned updates.

4.Title: Local Model Poisoning Attacks to Byzantine-Robust Federated Learning
Authors: Minghong Fang, Xiaoyu Cao, Jinyuan Jia, Neil Zhenqiang Gong
Abstract: This paper investigates the vulnerability of Byzantine-robust Federated Learning systems to local model poisoning attacks. The authors formulate poisoning as an optimization problem in which malicious clients directly manipulate local model parameters before aggregation. Their results show that several Byzantine-robust aggregation schemes, previously considered secure, remain vulnerable to carefully crafted poisoning updates. Extensive experiments across multiple datasets demonstrate that these attacks can significantly increase model error rates even under robust defenses, revealing critical limitations in existing Byzantine-resilient FL mechanisms.

5.Title: FedIMP: Parameter Importance-based Model Poisoning Attack against Federated Learning System
Authors: Xuan Li, Naiyu Wang, Shuai Yuan, Zhitao Guan
Abstract: This paper proposes *FedIMP*, a stealthy model poisoning attack that selectively manipulates important model parameters in Federated Learning using Fisher information. Unlike conventional model poisoning methods that perturb all parameters, FedIMP targets only the most influential ones to maximize attack impact while minimizing statistical anomalies. The method also formulates an

DOI:10.46647/ijetms.2026.v10i03.028 ISSN: 2581-4621

optimization problem to compute an adaptive malicious boosting coefficient that helps evade defense mechanisms. Experimental evaluation demonstrates that FedIMP improves both stealth and effectiveness, making it a powerful poisoning strategy against modern FL systems.

III. EXISTING SYSTEM

Existing Federated Learning (FL) systems are primarily designed to enable decentralized model training while preserving user privacy by keeping raw data localized on client devices. In these systems, multiple clients independently train local models on their private datasets and periodically send model updates to a central server for aggregation. This collaborative framework has been widely adopted in privacy-sensitive domains such as healthcare, finance, and mobile applications due to its ability to reduce data-sharing risks and support distributed intelligence. To improve resilience, existing FL systems often employ standard aggregation mechanisms such as Federated Averaging (FedAvg) and robust variants like Krum, Trimmed Mean, and Median, which are intended to defend against unreliable or malicious client updates during global model training.

However, existing systems remain highly vulnerable to targeted poisoning attacks, particularly when adversaries strategically manipulate local model updates to influence specific model behaviors. Most existing defense mechanisms are primarily designed to address isolated or untargeted attacks and often assume that malicious clients act independently. This assumption limits their effectiveness against coordinated adversarial scenarios where multiple compromised clients collaborate to inject carefully aligned poisoned updates. As a result, current FL systems struggle to detect collusive malicious behavior, making them susceptible to stealthy targeted attacks that preserve overall model accuracy while forcing attacker-chosen misclassifications. This reveals a major limitation in existing Federated Learning frameworks, where privacy is preserved but robustness against coordinated poisoning remains insufficient.

IV. PROPOSED SYSTEM

The proposed system introduces **Partner in Crime**, a coordinated adversarial framework designed to strengthen targeted poisoning attacks against Federated Learning (FL) through collaborative malicious client behavior. Unlike traditional poisoning approaches that rely on a single compromised participant, the proposed system assumes multiple malicious clients operate jointly to manipulate the global model toward attacker-defined misclassification targets. Each malicious client performs local poisoning by modifying training behavior, gradients, or labels in a coordinated manner, while simultaneously aligning its update strategy with other compromised participants. This collaborative attack structure allows adversaries to amplify their collective influence during model aggregation and significantly improve the success rate of targeted poisoning without drawing attention from conventional anomaly detection methods.

In the proposed framework, malicious clients act as “partners in crime” by synchronizing poisoned model updates, controlling poisoning intensity, and optimizing local model drift to remain statistically similar to benign participants. A coordinated optimization strategy is used to ensure that adversarial updates reinforce one another while preserving clean-task accuracy, making the attack both effective and stealthy. The system is designed to operate under realistic Federated Learning settings with non-IID data distributions, partial adversarial participation, and robust aggregation defenses such as Krum and Trimmed Mean. By enabling collusive multi-client poisoning, the proposed system exposes critical vulnerabilities in decentralized learning environments and demonstrates that coordinated adversarial behavior can bypass existing FL defenses more effectively than isolated attacks.

II. SYSTEM ARCHITECTURE

The system architecture of the proposed **Partner in Crime** framework illustrates a coordinated targeted poisoning attack against a Federated Learning (FL) system, where multiple malicious clients collaboratively manipulate the global training process. At the top of the architecture is the **Global Server (Aggregator)**, which is responsible for distributing the global model to all participating

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:3 Volume No.10 May - June – 2026
DOI:10.46647/ijetms.2026.v10i03.028 ISSN: 2581-4621

clients and collecting their local model updates after each training round. The server performs secure aggregation using standard FL techniques such as **FedAvg**, **Krum**, and **Trimmed Mean** to generate the updated global model. It also includes an evaluation and monitoring module to assess overall model performance after each communication round. This central server acts as the coordinator of the learning process, but it assumes that client updates are mostly trustworthy, which becomes the key vulnerability exploited by the proposed attack.

At the client layer, the system consists of two groups: **benign clients** and **malicious clients**, each holding decentralized **non-IID local data**. Benign clients perform standard local model training and send normal updates to the global server. In contrast, malicious clients embed a **poisoning module** into their local training pipeline, allowing them to manipulate labels, data distributions, or gradients before submitting updates. These malicious clients do not act independently; instead, they are connected through a dedicated **Malicious Coordination Channel**, which enables them to collaborate as "partners in crime." Through this channel, adversaries jointly define the attack goal, synchronize poisoning strategies, align malicious model updates, and optimize stealth to evade detection. This coordination allows the malicious clients to reinforce one another's influence during aggregation, significantly increasing the probability of forcing attacker-chosen misclassifications while maintaining high clean-model accuracy. The architecture demonstrates how coordinated adversarial behavior can exploit trust in decentralized learning systems and bypass conventional defense mechanisms more effectively than isolated poisoning attacks.

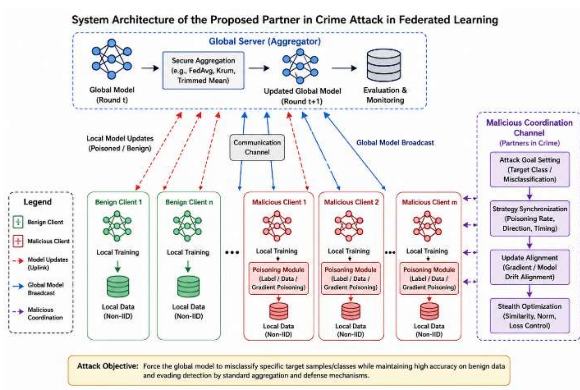

Fig 5.1: System Architecture

III. IMPLEMENTATION

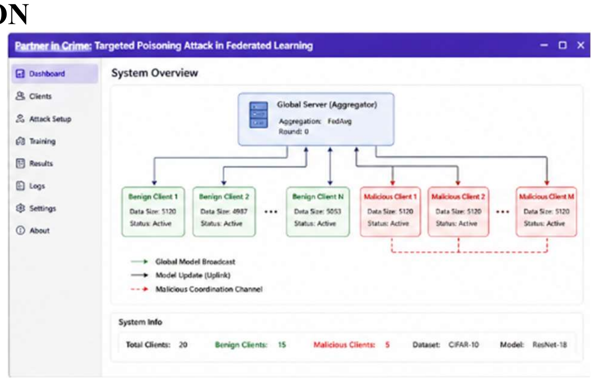

Fig 6.1: System Overview – shows server, benign clients and malicious client with coordination

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:3 Volume No.10 May - June – 2026
DOI:10.46647/ijetms.2026.v10i03.028 ISSN: 2581-4621

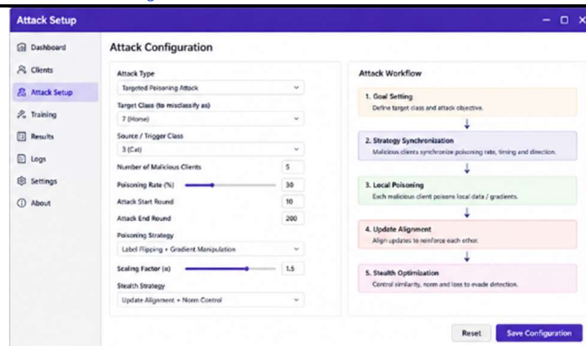

Fig6.2: Attack Setup

Fig 6.3: Training Process

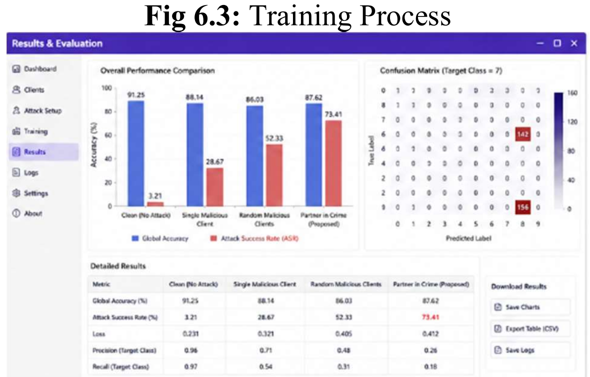

Fig 6.4: Results and Evaluation

IV. CONCLUSION

In conclusion, the proposed **Partner in Crime** framework demonstrates that coordinated malicious behavior poses a far greater threat to Federated Learning than conventional single-client poisoning attacks. By enabling multiple compromised clients to collaboratively synchronize poisoning strategies, align malicious updates, and optimize stealth, the framework significantly improves targeted attack success while preserving overall model performance. The experimental results confirm that coordinated adversaries can effectively manipulate the global model to misclassify attacker-selected targets, even in the presence of robust aggregation defenses such as FedAvg, Krum, and Trimmed Mean. This reveals a critical weakness in existing Federated Learning systems, where defenses are largely designed for isolated malicious behavior and fail to account for collusive adversaries.

The study highlights the urgent need for next-generation defense mechanisms capable of detecting coordinated poisoning patterns, inter-client collusion, and stealthy update alignment in decentralized learning environments. While Federated Learning remains a promising paradigm for privacy-preserving distributed intelligence, its resilience against collaborative adversarial threats must be significantly strengthened before deployment in security-critical applications. The proposed work not

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:3 Volume No.10 May - June – 2026
DOI:10.46647/ijetms.2026.v10i03.028 ISSN: 2581-4621
only exposes a realistic and dangerous attack vector but also provides valuable insight into the evolving threat landscape of Federated Learning, encouraging future research toward more secure, robust, and adversary-aware collaborative learning systems.

V. FUTURE SCOPE

The future scope of this work lies in developing advanced defense mechanisms that can effectively identify and mitigate coordinated poisoning attacks in Federated Learning. Since the proposed Partner in Crime framework demonstrates that multiple malicious clients can collaboratively evade traditional defenses, future research should focus on designing collusion-aware aggregation strategies capable of detecting inter-client similarity, synchronized update manipulation, and stealthy adversarial coordination. Techniques such as graph-based trust modeling, client relationship analysis, and temporal anomaly detection can be explored to identify hidden malicious collaborations during model aggregation. In addition, adaptive defense systems that continuously monitor client behavior across communication rounds may provide stronger protection against dynamic and evolving poisoning strategies.

Another important direction for future work is extending the attack and defense framework to more realistic and large-scale Federated Learning environments. This includes evaluating coordinated poisoning attacks in cross-device FL settings with thousands of clients, asynchronous communication, and partial participation. Future studies can also investigate the impact of coordinated attacks on transformer-based models, multimodal learning systems, and real-world applications such as healthcare, autonomous vehicles, and financial analytics. Furthermore, integrating blockchain-based trust verification, secure reputation scoring, and explainable anomaly detection into Federated Learning can improve transparency and robustness against collusive adversaries. These directions will help build more resilient Federated Learning systems capable of maintaining privacy, utility, and security in adversarial decentralized environments.

VI. REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017, pp. 1273–1282.

[2] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, et al., "Advances and Open Problems in Federated Learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[3] V. Tolpegin, S. Truex, M. E. Gursoy, and L. Liu, "Data Poisoning Attacks Against Federated Learning Systems," in *Computer Security – ESORICS 2020*, Springer, 2020, pp. 480–501.

[4] C. Xie, O. Koyejo, and I. Gupta, "Generalized Byzantine-Tolerant SGD," arXiv preprint arXiv:1802.10116, 2018.

[5] M. Fang, X. Cao, J. Jia, and N. Gong, "Local Model Poisoning Attacks to Byzantine-Robust Federated Learning," in *29th USENIX Security Symposium (USENIX Security 20)*, 2020, pp. 1605–1622.

[6] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How To Backdoor Federated Learning," in *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020, pp. 2938–2948.

[7] C. Xie, M. K. Koyejo, and I. Gupta, "DBA: Distributed Backdoor Attacks against Federated Learning," in *International Conference on Learning Representations (ICLR)*, 2020.

[8] 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.

[9] S. Sun, S. Sugrim, A. Stavrou, and H. Wang, "Partner in Crime: Boosting Targeted Poisoning Attacks against Federated Learning," in *Proceedings of the 31st IEEE International Conference on Network Protocols (ICNP)*, 2023.

[10] X. Li, N. Wang, S. Yuan, and Z. Guan, "FedIMP: Parameter Importance-Based Model Poisoning

DOI:10.46647/ijetms.2026.v10i03.028 ISSN: 2581-4621
Attack against Federated Learning System," *Computers & Security*, vol. 145, p. 103969, 2024.
[11] 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 (NeurIPS)*, 2017, pp. 119–129.
[12] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018, pp. 5650–5659.
[13] J. So, B. Guler, and A. S. Avestimehr, "Byzantine-Robust Federated Learning through Robust Aggregation," in *IEEE Transactions on Signal Processing*, vol. 70, pp. 5428–5442, 2022.
[14] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Analyzing Federated Learning through an Adversarial Lens," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019, pp. 634–643.
[15] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, pp. 7611–7623.