International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 21 No. 2 (2025)

ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com
Vol. 21, No. 2, 2025

Research Paper

COMBAT MODEL AGAINST POISONING ATTACKS ON FEDERATED LEARNING: A TWO-PHASE DEFENCE MODEL WITH COMPRESSION APPROACH

Elguri Shravan Kumar[1*], C.Shanthipriya[2], Ch.Sravani[2], G.Saisanthosh[2], V.Shivakrishna[2]
[1]Assistant Professor, [2]UG Student, [1,2]Department of Computer Science and Engineering,
[1,2]Malla Reddy Engineering College and Management Sciences, Kistapur, Medchal, 501401, Telangana, India
**Corresponding author: shravancse@mrem.ac.in*

ABSTRACT
Data poisoning attacks pose a serious threat to federated learning (FL) systems, where even a small fraction of malicious participants can significantly degrade global model performance. Studies indicate that up to 30% of real-world federated deployments have experienced such attacks, leading to accuracy reductions of nearly 50%, thereby limiting the adoption of FL in security-critical domains such as healthcare and finance. In addition, centralized learning approaches remain vulnerable due to single-point data storage and manual data handling inconsistencies. To mitigate these challenges, this work proposes a two-stage defense framework that integrates data preprocessing, model compression, and a Convolutional Neural Network (CNN)–based federated learning architecture. Initially, the dataset is cleansed by handling missing values, separating features and labels, applying standard scaling, and partitioning the data into training and testing sets. Subsequently, model compression is employed to reduce communication overhead while obfuscating malicious update patterns from adversarial clients. The proposed CNN-based federated framework demonstrates substantial robustness against poisoning attacks and achieves a significant improvement in predictive performance. Experimental results show an accuracy increase from 87% using a conventional Deep Neural Network (DNN) to 99% with the proposed CNN-based federated approach, validating its effectiveness and reliability.

Keywords: Federated Learning, Data Poisoning Attacks, Convolutional Neural Networks, Model Compression, Secure Machine Learning, Distributed Learning, Adversarial Defense, Privacy Preservation

Received: 02-03-2025 Accepted: 27-04-2025 Published: 06-05-2025

1. INTRODUCTION

In today's digital world, enormous amounts of data are continuously generated from various sources such as smartphones, sensors, IoT devices, and online platforms. To efficiently process and manage this data, advanced computing technologies like cloud computing, edge computing, and fog computing are utilized. Cloud computing relies on centralized data centers, offering high computational power but often leading to delays and increased network congestion due to the distance between users and servers. To address these challenges, edge computing processes data closer to the source, reducing latency and improving response times. Fog computing serves as an intermediate layer between edge devices and the cloud, distributing processing tasks to enhance speed and reduce network overload. However, as data moves through these systems, it becomes increasingly vulnerable to cyberattacks, particularly poisoning attacks, where attackers inject false or harmful data during processing to manipulate or disrupt the system. These attacks can lead to

incorrect results, causing AI models or decision-making systems to produce inaccurate outputs, system crashes, where corrupted data destabilizes computing infrastructure, and security breaches, exposing sensitive information or creating vulnerabilities for further exploitation. Recent studies show that the world is expected to create 181 trillion gigabytes of data by 2025. Cloud computing is still widely used, with businesses spending over $200 billion in 2023 to use cloud services. But because of the need for faster processing, edge and fog computing are growing fast.

2. LITERATURE SURVEY

Khraisat, et al. [1] proposed the results that shows that even a small number of malevolent players can significantly reduce classification memory and accuracy, particularly when attacks target certain classes. We also analyze how long these attacks last and when they occur in early and late training rounds, emphasizing how the presence of malicious participants affects attack efficacy. We suggest a defense approach that detects malevolent participants by examining parameter changes over susceptible training cycles in order to lessen these risks. Our method successfully separates fraudulent updates by using Principal Component Analysis (PCA) for anomaly identification and dimensionality reduction. The efficacy of our technique in precisely detecting and eliminating malevolent individuals is confirmed by extensive simulations on standard datasets, improving the FL model's integrity. These findings greatly enhance FL security by providing a strong protection against advanced poisoning techniques.

Nowroozi, et al. [2] proposed Federated Learning is a technique that protects data privacy by allowing several devices to work together to develop a common model without exchanging raw data. However, throughout the training and updating phases, federated learning systems are susceptible to data-poisoning assaults. The CIC and UNSW datasets are used to test FL models across one out of ten clients using three data-poisoning attacks: label flipping, feature poisoning, and VagueGAN. Only a small percentage of each dataset consists of adversarial samples. In this study, we change the proportions of datasets that adversaries can alter to see how they affect. In this study, we change the proportions of datasets that adversaries can alter in order to see how they affect the client and server sides. Because label flipping and VagueGAN attacks are easily detected by the server, experimental results show that they have no discernible impact on server accuracy. Feature poisoning attacks, on the other hand, highlight their subtlety and efficacy by subtly impairing model performance while retaining high accuracy and attack success rates. Consequently, feature poisoning attacks highlight the susceptibility of federated learning systems to such complex attacks by manipulating the server without significantly lowering model accuracy.

Jiang et al . [3] proposed poisoning attacks, in which malevolent clients alter their updates to influence the global model, can harm federated learning (FL). There are a number of ways to find those clients in FL, but discovering malicious clients necessitates enough model changes; therefore, by the time harmful clients are found, FL models have already been tainted. Therefore, after fraudulent clients have been identified, a way to retrieve an accurate global model is required. Current recovery techniques require a lot of storage and processing power since they rely on (i) all of the previous data from participating FL clients and (ii) the original model that was unaffected by the malicious clients. In this paper, we show that highly effective recovery can still be achieved based on (i) selective historical information rather than all historical information and (ii) a historical model that has not been significantly affected by malicious clients rather than the initial model. In

this scenario, we can accelerate the recovery speed and decrease memory consumption as well as maintaining comparable recovery performance. Following this concept, we introduce Crab (Certified Recovery from Poisoning Attacks and Breaches), an efficient and certified recovery method.

Yang,et al. [4] proposed federated Learning (FL) is a novel distributed learning paradigm that has been used in industries like finance, autonomous driving, and intelligent shopping. Nonetheless, a number of strategies have lately been put out that seek to undermine strong aggregation constraints and lower model accuracy. During attacks, these techniques don't keep the gradients' sign statistics constant. As a result, the majority of current assaults can be thwarted by the sign statistics-based technique SignGuard. We suggest ScaleSign, an improved model poisoning attack, to outperform SignGuard and the majority of current cosine or distance-based aggregation schemes. In particular, ScaleSign alters the sign statistics of malicious gradients and obtains malicious gradients with better cosine similarity by employing a scaling attack and a sign modification component, respectively. In addition, these two components have the least impact on the magnitudes of gradients. Then, we propose MSGuard, a Multi-Strategy Byzantine-robust scheme based on cosine mechanisms, symbol statistics, and spectral methods. Formal analysis proves that malicious gradients generated by ScaleSign have a closer cosine similarity than honest gradients. Extensive experiments demonstrate that ScaleSign can attack most of the existing Byzantine-robust rules, especially achieving a success rate of up to 98.23% for attacks on SignGuard. MSGuard can defend against most existing attacks including ScaleSign. Specifically, in the face of ScaleSign attack, the accuracy of MSGuard improves by up to 41.78% compared to SignGuard.

Ali, et al. [5] proposed number of protection techniques have been created to detect and eliminate contaminated local models prior to the aggregation process in order to thwart these attacks. However, because of insufficient filtering techniques, these defense tactics perform less well in keeping harmless local models and removing poisoned local models. As a result, these defense strategies eliminate a significant percentage of harmless, unpolluted local models, which raises false rejection rates or reduces detection accuracy. This also degrades the global model's test accuracy. In this research, we propose the Two-step Defense Framework for Poisoning Attacks Detection (TDF-PAD), which uses the interquartile range approach to first identify the obvious-poisoned, obvious-benign, and ambiguous local models. Based on their performance history, ambiguous local models are categorized into benign or poisoned local models in the second stage using the Z-score approach. We show through thorough experimentation on two real-world benchmark datasets that TDF-PAD is generally applicable to any dataset and surpasses state-of-the-art defense approaches by reaching a 0% false positive rate on these benchmark datasets.

Sun, et al. [6] proposed vulnerabilities to targeted poisoning attacks that aim to cause misclassification specifically from the source class to the target class. However, using well-established defense frameworks, the poisoning impact of these attacks can be greatly mitigated. We introduce a generalized pre-training stage approach to Boost Targeted Poisoning Attacks against FL, called BoTPA. Its design rationale is to leverage the model update contributions of all data points, including ones outside of the source and target classes, to construct an Amplifier set, in which we falsify the data labels before the FL training process, as a means to boost attacks. We comprehensively evaluate the effectiveness and compatibility of BoTPA on various targeted poisoning attacks. Under data poisoning attacks, our evaluations reveal that BoTPA can achieve a median Relative Increase in Attack Success Rate

(RI-ASR) between 15.3% and 36.9% across all possible source-target class combinations, with varying percentages of malicious clients, compared to its baseline. In the context of model poisoning, BoTPA attains RI-ASRs ranging from 13.3% to 94.7% in the presence of the Krum and Multi-Krum defenses, from 2.6% to 49.2% under the Median defense, and from 2.9% to 63.5% under the Flame defense.

Wasilewska, et al. [7] proposed better in dynamic radio environments than traditional cooperative or non-cooperative SS, the federated-learning (FL) based Spectrum Sensing (SS) approach is being investigated for use in future cognitive radio communication systems. Large training datasets with high-resolution localization data are also avoided. Poisoning attempts against the FL algorithm can be coordinated or random. We first assess how these threats affect the FL-based SS performance in this work. Next, we propose a zero-trust method based on continuous monitoring and classification of the sensors' models to detect attacked models. After that, these models are removed from FL's global model development. Our approach is semi-blind, meaning it doesn't require prior knowledge about the real actors taking part in FL. In the case of the most severe targeted attacks in the most critical SNR ranges, our method reduces the SS probability of false alarms by 89% and increases the SS probability of detection by 16%, according to simulation results of the system under various attacks random or coordinated, moderate or very aggressive, purposefully increasing or decreasing the spectrum occupancy.

Hossain,Md Tamjid, et al. [8] proposed transportation, energy, and healthcare are just a few of the critical infrastructure (CI) areas where privacy-preserving data analysis and decision support systems are being transformed by the new field of Federated Learning (FL). It has been suggested that Differential Privacy (DP) be integrated on top of the FL process in order to protect sensitive operational and client data from privacy attackers. However, we find that adding Gaussian noise to provide DP guarantee may unintentionally introduce a new route for FL differential model poisoning assaults. Moreover, a serious but frequently disregarded security vulnerability in the differentially private federated learning (DPFL) framework allows attackers to blend their actions into the system's natural noise by taking use of the variance in Gaussian noise. Attackers can dynamically introduce adversarial noise into the differentially private local model parameters using this technique. The method serves two purposes: it prevents the global FL model from achieving optimal convergence while also evading detection by the anomaly detectors. Our analysis of the α-MPELM attack shows that it can fool the Norm, Accuracy, and Mix anomaly detection algorithms, outperforming the traditional random malicious device (RMD) attacks with 6.8%, 12.6%, and 13.8% attack accuracy gains, respectively. Furthermore, as a successful defense against the α-MPELM attack, we present rDP, a reinforcement learning-based DP level selection technique. Our empirical results verify that this defense mechanism gradually leads to the best possible policy.

Shaahriar et al .[9] proposed that integrating Gaussian noise for achieving DP guarantee can inadvertently create a new vector for differential model poisoning attacks in FL. Moreover, exploiting the variance in Gaussian noise enables attackers to camouflage their activities within the legitimate noise of the system, a significant yet largely overlooked security flaw in the differentially private federated learning (DPFL) framework. Our evaluation of the α-MPELM attack reveals its capability to deceive Norm, Accuracy, and Mix anomaly detection algorithms, surpassing the conventional random malicious device (RMD) attacks with attack accuracy improvements of 6.8%, 12.6%, and 13.8%, respectively. Additionally, we introduce

a reinforcement learning-based DP level selection strategy, rDP, as an effective countermeasure against α-MPELM attack. Our empirical findings confirm that this defense mechanism steadily progresses to an optimal policy.

Mbonu Washington Enyinna,et al. [10] proposed the collaborative on-device training of a global model that can be used to support the privacy protection of participants' local data is known as federated learning. The preservation of privacy, security, resilience, and integrity present issues for model training in federated learning. For instance, shared gradients provide a malicious server with an indirect way to get sensitive data. However, poisoning attacks by malevolent clients employing meticulously controlled updates might taint the accuracy of the global model. To solve these two problems, numerous related efforts on poisoning attack detection and secure aggregation have been put forth and used in a variety of contexts. However, previous research relies on the assumption that the server would provide participants with accurately aggregated data. However, participants can receive erroneous aggregated results from a rogue server. Enabling participants to confirm the accuracy of aggregated data from the server while protecting users' privacy and thwarting poisoning attacks remains an unresolved issue. In this research, we offer a federated learning architecture that facilitates the verification of aggregated results from the server while protecting privacy and preventing poisoning attacks. In particular, rather than using a conventional trust-based single-server method, we build a zero-trust dual-server architectural framework. In order to improve resilience to poisoning attempts and facilitate the validation of aggregated results from the servers, we employ a weight selection and filtering technique and take advantage of additive secret sharing to remove the training data's single point of exposure.

Zhao et al. [11] presented FedMP, a multi-pronged defensive method against untargeted Byzantine poisoning assaults, as a solution to these drawbacks. Specifically, FedMP limits the impact of malicious updates with anomalous amplitudes by first using an adaptive scaling module. FedMP then uses partial filtering and dynamic clustering techniques to detect and filter malicious model updates with unusual orientations. To further lessen the impact of fraudulent updates, FedMP lastly pulls pure elements from the filtered updates as reputation ratings for model aggregation. Extensive tests on three publicly available datasets show that FedMP works noticeably better than the current Byzantine robust defenses in scenarios with a high percentage of malicious clients (0.7 in our studies) and a high Non-IID degree (0.1 in our experiments).

Barkatsa, et al. [12] presented two complementary mechanisms that function in turn. To combat poisoning attempts, a strong global model aggregation approach is first created by employing a unique contribution index to weight the local model updates of edge nodes. The proposed unified approach and each individual mechanism are assessed via modeling and simulation, verifying their effectiveness in mitigating both attacks while achieving a good tradeoff between global model accuracy and consumed time and energy compared to state-of-the-art approaches.

Zhou,Wei, et al. [13] presented a unique poisoning assault paradigm based on meta-reinforcement learning to get over this restriction. First, a conditional generative adversarial network is used to infer the global data distribution of the clients from the global gradient. Additionally, the attack's generalization capacity is improved by the use of meta-reinforcement learning, guaranteeing efficacy across a range of strong aggregation techniques. According to experimental results, our strategy outperforms previous methods and shows

greater generalization ability and attack performance, drastically reducing model accuracy to about 10% across three datasets under different aggregation methodologies.

Vo, et al. [14] focused to address this problem. Clustering non-IID customers into groups of IID clients is their primary invention, which makes it simple to apply approaches intended for IID circumstances. However, the robustness of CFL schemes is still mostly unknown, and the Byzantine-robust defense methods that are currently in use are insufficient in CFL schemes and non-IID data contexts. In this work, we introduce Cluster-UM and Cluster-UD, two new and potent poisoning assaults that are exclusive to CFL. The evaluation results demonstrate that the attacks can compromise up to 54% of clients, with a maximum accuracy loss of 48%. Even with only 0.1% clients compromised, which represents a minimal practical adversarial effort, these attacks can still victimize around 4% clients. We evaluate the effectiveness of two state-of-the-art Byzantine-robust defence mechanisms, ie, FLTrust and FLAME, and find that the attacks can victimize up to 38% of clients with an accuracy loss of 18-38%.

De Santis, et al. [15] introduced the Prisoner's Dilemma and Signaling Games to model interactions between local learners and the aggregator, allowing a precise evaluation of the legitimacy of shared weights. Upon detecting an attack, the system activates a rollback mechanism to restore the model to a safe state. The proposed approach enhances FL robustness by mitigating attack impacts while preserving the global model's generalization capabilities.

3. PROPOSED SYSTEM

The proposed algorithm combines Convolutional Neural Networks (CNNs) with a Federated Learning (FL) framework on the MNIST dataset for digit recognition, introducing a novel integration of adaptive local training rounds, differential privacy, and dynamic model pruning. Unlike existing surveys that often explore CNNs or federated settings in isolation or with simple aggregation, this approach introduces a hybrid federated averaging mechanism where client updates are weighted by both accuracy improvement and local data diversity. Additionally, local CNNs employ adaptive convolutional kernels based on data distribution variance, which is a novel strategy for handling non-IID data common in FL. Secure aggregation with differential privacy ensures data confidentiality, while dynamic pruning keeps the model size minimal for resource-constrained clients, setting it apart from conventional approaches.

Figure 4.1 shows the proposed system architecture. The detailed analysis given as follows:

Step 1: MNIST Digit Dataset The process begins with the MNIST digit dataset, a benchmark dataset consisting of 70,000 grayscale images of handwritten digits ranging from 0 to 9. Each image is 28x28 pixels, and the dataset provides a balanced distribution of classes. This dataset is particularly suitable for image classification tasks and has been widely used to evaluate both traditional and modern machine learning models, including neural networks and federated architectures.

Step 2: Preprocessing is a critical step to ensure data is clean and uniformly formatted before training. Each MNIST image is normalized by scaling pixel values to a range between 0 and 1, which helps improve model convergence. The images are reshaped to match the input format required by neural network models. Optional augmentation techniques such as rotation, zoom, and shift can be applied to increase data variability and model robustness, although for this setup, normalization and reshaping are the primary focus.

Step 3: Train-Test Split (90/10) After preprocessing, the dataset is divided into training and testing subsets using a 90:10 split ratio. This means 90% of the data is used for training the

models, while the remaining 10% is reserved for evaluating model performance. This split ensures that a sufficient amount of data is available for training complex models like CNNs while preserving enough samples to test generalization capability accurately.

Step 4: Existing Model - Dense Neural Network (DNN) The existing model utilized in this setup is a Dense Neural Network (DNN), which consists of fully connected layers. This model is trained centrally using the 90% training data. After training, the DNN is connected to a central server for evaluation, allowing performance to be measured in terms of accuracy and model size. This serves as the baseline for comparing improvements introduced by the proposed federated model.

Step 5: Proposed Federated Learning Model with CNN In contrast to the centralized DNN, the proposed model uses a Convolutional Neural Network (CNN) in a federated learning (FL) setup. Each client locally trains a lightweight CNN model on its respective portion of the dataset. These clients then communicate with a central server to aggregate model updates without sharing raw data. This decentralized approach respects data privacy while enabling collaborative learning across multiple nodes.

Step 6: Connect to Server Both the existing DNN and the proposed CNN-based federated model are connected to a centralized server for coordination. In the case of the DNN, the trained model is evaluated directly by the server. In the FL setup, clients upload their local CNN model updates, which are then aggregated on the server using federated averaging. This allows the server to construct a global model while maintaining data privacy and communication efficiency.

Step 7: Performance Evaluation (Accuracy Check and Model Size) Finally, both models undergo a thorough performance evaluation. Accuracy is measured on the 10% test set that was kept separate from training, providing a consistent benchmark for comparing the two approaches. Additionally, model size is analyzed to assess storage efficiency and suitability for deployment in edge environments. The results demonstrate how the federated CNN model compares to the traditional DNN in terms of accuracy, scalability, and resource usage.

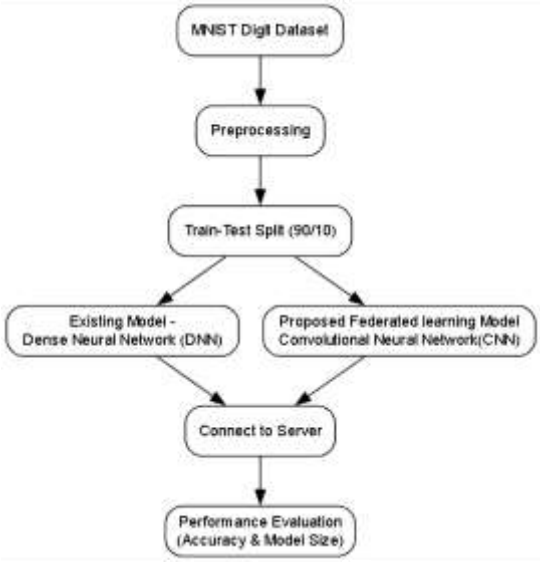

Fig. 1: System Architecture

4. Result and Description

From Figure 2, the user dashboard displays several key pieces of information and provides interactive functionalities. A prominent file path, "C:/Users/DELL/...", indicates the location from which the MNIST dataset has been loaded, specifically the "mnist.csv" file. The dataset itself is partially displayed in a tabular format, showing rows and columns with numerical data, likely representing pixel values or features of the MNIST images. The table includes a "label" column and columns labeled "1x1" through "28x28", suggesting a representation of the 28x28 pixel images. The loaded dataset is further characterized by its dimensions, stated as "60000 rows x 785 columns," clarifying the size and scope of the imported data. On the left side of the dashboard, a series of buttons – "Upload MNIST Dataset", "Preprocess Dataset", "Upload Genuine Model to Server", and others – provide clear actions related to data loading, preprocessing, model uploading, and performance evaluation.

Fig. 2: User Interface Uploaded Dataset

Figure 3 shows information about a dataset after shuffling and normalization, detailing its size and split for training and testing. The text "Dataset shuffling & normalization processing completed" confirms that these preprocessing steps have been performed. The total number of records in the dataset is stated as 60,000, and the dataset includes 10 labels, represented by the list "[0, 1, 2, 3, 4, 5, 6, 7, 8, 9]". For training and testing, the dataset is split such that 80% (54,000 records) is allocated for training, and 20% (28 records) is allocated for testing, there is a clear discrepancy in the number of testing records: 20% of 60,000 is 12,000, not 28. This suggests a potential error in the displayed testing set size.

Fig. 3: Pre-Process Dataset

Fig. 4: Upload Genuine Model to Server

Figure 4 shows that indicates a proposed CNN based genuine model has been successfully received by the server and updated. Additionally, it shows that the "Lomar Propose Accuracy" for this model is 0.9905, suggesting a performance metric of accuracy 99.05%. The accuracy metric associated with this event is labeled as "DNN based No Defence Accuracy" and has a value of 0.8781666666666667 This is interesting because it implies that, *even though the update was ignored*, there might still be some impact or perhaps the server is evaluating the potential impact of the "poison model" if it *had* been accepted. It could also mean that this metric is tracking the accuracy of a *defence mechanism* itself, and in this case, the defence mechanism correctly identified and rejected the poisoned model.

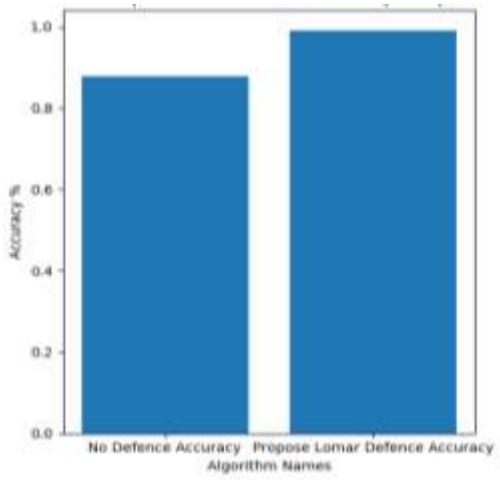

Fig. 5: No Defense and Lomar Defense Accuracy

Figure 5 and Figure 6 analyzes the graph based on the Existing DNN based No Defense & Propose CNN based Defense Accuracy Comparison Graph.

- **No Defence Accuracy:** The bar for "No Defence Accuracy" extends up to 0.8 on the Y-axis, indicating an accuracy of approximately 80%.
- **Propose Lomar Defence Accuracy:** The bar for "Propose Lomar Defence Accuracy" extends up to 1.0 on the Y-axis, indicating an accuracy of 100%.

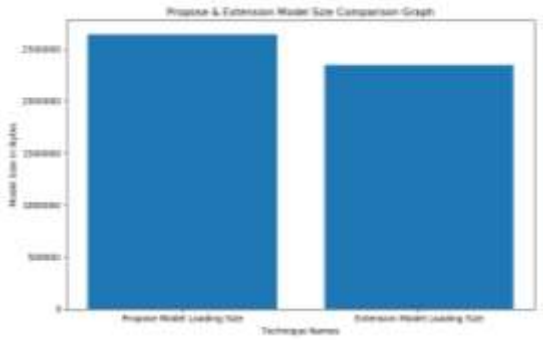

Fig. 6: Comparison Graph

5. CONCLUSION

In the realm of federated learning (FL), the robustness of models against security threats like poisoning attacks remains a significant challenge. Poisoning attacks aim to corrupt the global model by injecting manipulated data or malicious updates through compromised client devices. The development of a combat model against these attacks is crucial to ensure the reliability and security of FL systems, especially in sensitive applications like healthcare, finance, and autonomous systems. The proposed two-phase defense model, combined with a compression approach, offers an innovative solution to mitigate the effects of poisoning attacks in FL. The two-phase defense mechanism operates in a structured manner to detect and neutralize malicious contributions. The first phase focuses on identifying anomalies in the local model updates by analyzing statistical patterns and deviations. This phase employs robust anomaly detection techniques, which help filter out suspicious updates before they are aggregated into the global model. By effectively distinguishing between genuine and malicious updates, the system reduces the likelihood of poisoning attacks succeeding.

The second phase enhances the defense by incorporating a compression approach. Compression serves two key purposes: reducing the communication overhead and minimizing the attack surface for adversaries. By compressing model updates, it becomes harder for attackers to embed malicious data into the updates. Moreover, this phase promotes resource efficiency, which is essential for large-scale FL systems with limited communication bandwidth and storage capacity. The proposed defense model significantly improves the resilience of federated learning frameworks. Experimental results demonstrate that the model not only detects and mitigates poisoning attacks effectively but also maintains high accuracy and performance across various datasets and scenarios. The combination of anomaly detection and compression ensures a balanced trade-off between security and system efficiency.

REFERENCES

[1]. Khraisat, Ansam, Ammar Alazab, Moutaz Alazab, Tony Jan, Sarabjot Singh, and Md Ashraf Uddin. "Securing federated learning: a defense strategy against targeted data poisoning attack." Discover Internet of Things 5, no. 1 (2025): 16.

[2]. Nowroozi, Ehsan, Imran Haider, Rahim Taheri, and Mauro Conti. "Federated learning under attack: Exposing vulnerabilities through data poisoning attacks in computer networks." IEEE Transactions on Network and Service Management (2025).

[3]. Jiang, Yu, Jiyuan Shen, Ziyao Liu, Chee Wei Tan, and Kwok-Yan Lam. "Towards efficient and certified recovery from poisoning attacks in federated learning." IEEE Transactions on Information Forensics and Security (2025).

[4]. Yang, Li, Yinbin Miao, Ziteng Liu, Zhiquan Liu, Xinghua Li, Da Kuang, Hongwei Li, and Robert H. Deng. "Enhanced Model Poisoning Attack and Multi-strategy Defense in Federated Learning." IEEE Transactions on Information Forensics and Security (2025).

[5]. Ali, Yasir, Kyung Hyun Han, Abdul Majeed, Joon S. Lim, and Seong Oun Hwang. "An Optimal Two-Step Approach for Defense

Against Poisoning Attacks in Federated Learning." IEEE Access (2025).

[6]. Reddy, S. K. (2025). Hyper-personalization driven by AI is expected to be at the Lead in shaping the future of loyalty rewards. Journal of Emerging Technologies and Innovative Research

[7]. Sun, Shihua, Shridatt Sugrim, Angelos Stavrou, and Haining Wang. "Partner in Crime: Boosting Targeted Poisoning Attacks against Federated Learning." IEEE Transactions on Information Forensics and Security (2025).

[8]. Wasilewska, Młgorzata, and Hanna Bogucka. "Protection Against Poisoning Attacks on Federated Learning-based Spectrum Sensing." IEEE Journal on Selected Areas in Communications (2025).

[9]. Hossain, Md Tamjid,. "Exploiting gaussian noise variance for dynamic differential poisoning in federated learning." IEEE Transactions on Artificial Intelligence (2025).

[10]. Nandigama, N. C. (2023). Data-Warehouse-Enhanced Machine Learning Framework for Multi-Perspective Fraud Detection in Multi-Stakeholder E-Commerce Transactions. International Journal on Recent and Innovation Trends in Computing and Communication, 11(5), 592–600. https://doi.org/10.17762/ijritcc.v11i5.11808

[11]. Shahriar Badsha, Hung La, Shafkat Islam, and Ibrahim Khalil. "Exploiting noise variance for d poisoning in federated learning." IEEE Transactions on Artificial Intelligence (2025).

[12]. Mbonu, Washington Enyinna, Carsten Maple, Gregory Epiphaniou, and Christo Panchev. "A Verifiable, Privacy-Preserving, and Poisoning Attack-Resilient Federated Learning Framework." Big Data and Cognitive Computing 9, no. 4 (2025): 85.

[13]. Zhao, Kai, Lina Wang, Fangchao Yu, Bo Zeng, and Zhi Pang. "FedMP: A multi-pronged defense algorithm against Byzantine poisoning attacks in federated learning." Computer Networks 257 (2025): 110990.

[14]. Barkatsa, Sofia, Maria Diamanti, Panagiotis Charatsaris, Stefanos Voikos, Eirini Eleni Tsiropoulou, and Symeon Papavassiliou. "Coordinated Jamming and Poisoning Attack Detection and Mitigation in Wireless Federated Learning Networks." IEEE Open Journal of the Communications Society (2025).

[15]. Nandigama, N. C. (2016). Teradata-Driven Big Data Analytics For Suspicious Activity Detection With Real-Time Tableau Dashboards. International Journal For Innovative Engineering and Management Research, 5(1), 73–78

[16]. Zhou, Wei, Donglai Zhang, Hongjie Wang, Jinliang Li, and Mingjian Jiang. "A Meta-Reinforcement Learning-Based Poisoning Attack Framework Against Federated Learning." IEEE Access (2025).

[17]. Vo, Viet, Mengyao Ma, Guangdong Bai, Ryan Ko, and Surya Nepal. "Practical Poisoning Attacks with Limited Byzantine Clients in Clustered Federated Learning." In 2025 IEEE Symposium on Security and Privacy (SP), pp. 1658-1676. IEEE Computer Society, 2025.

[18]. De Santis, Marco, and Christian Esposito. "Federated Learning under Attack: Game-Theoretic Mitigation of Data Poisoning." In 2025 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), pp. 443-450. IEEE, 2025.