iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** XI **Month of publication:** November 2024

DOI: https://doi.org/10.22214/ijraset.2024.65669

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue XI Nov 2024- Available at www.ijraset.com

A Trust-Based Framework for Mitigating Poisoning Attacksin Federated Learning

Yakubu Galadima Ibrahim[1], Armaya'U Zango Umar[2], Abdulmalik Umar[3]
[1, 2, 3]*Al-Qalam University Katsina*

Abstract: ***Federated Learning (FL) is a decentralized machine learning approach that allows collaborative model training across multiple devices while preserving data privacy. However, its decentralized nature exposes it to poisoning attacks, where malicious clients corrupt model updates, compromising the global model's integrity. This study introduces a trust-based framework that combines Krum, Differential Privacy (DP), and adaptive trust measures to counter these threats effectively. The framework dynamically assigns trust scores to clients based on their behavior, mitigating the impact of unreliable or malicious updates. Using the MNIST dataset, the framework's robustness is tested against label flipping and gradient manipulation attacks at varying intensities. Results demonstrate the hybrid approach's superior performance in accuracy, precision, recall, and F1-scores compared to standalone defenses, showcasing its adaptability and resilience. This research underscores the impor- tance of integrating robust, scalable defense mechanisms in FL to ensure secure, reliable, and trustworthy systems in adversarial settings***
Keywords: federated learning, security, privacy and poisoning

I. INTRODUCTION

Introduced by Google, [3] Federated Learning (FL) is a decentralized machine learning approach that allows training of shared model across multiple devices or servers without sharing their in- dividual data. However, its decentralized nature makes it vulnerable to poisoning attack, [8] where malicious clients inject manipulated data to corrupt the global model and compromise its performance. Data poisoning involves tampering with client data to induce errors, while model poisoning targets model gradients directly, altering updates to affect the global model. Studies such as [11, 4, 2] have explored these vulnerabilities in depth. Addressing these attacks necessitates robust and efficient defense mechanisms. An existing study adopted Majority voting to aggregates model updates from clients, discarding outliers of a majority threshold [5]. While simple, this approach is susceptible to collusion attacks where malicious actors collaborate to manipulate the majority vote [5]. Byzantine Fault Tolerance (BFT) protocol such as Practical-BFToffer stronger resilience against collusion [10] but requires significant communication overhead, hindering scalability. Federated averaging with anomaly detection identifies and exclude update deviating significantly from the average model [12]. However, its effectiveness depends on the accuracy of anomaly detection algorithm, which can be challenging in complex scenario [7]. Numerous methods have been proposed to counteract these vulnerabilities such as reputation systems that assign trust scores based on behavior, down weighting or discarding updates from low-trust entities [1] and untrusted clients [1]. Krum, an approach by [14], is designed to identify and ex- clude malicious updates, retaining only the most reliable client updates based on proximity to the majority's median gradient. Although effective, Krum is less robust against sophisticated, coordinated attacks[9]. Differential Privacy (DP) [6]provides another layer of security by injecting noiseinto model updates, thus masking individual client contributions. DP, as discussed by [6], pre- serves privacy but can impact accuracy, especially in data-diverse environments. Hybrid defense mechanisms combine approaches like Krum and DP to enhance robustness [13]. These hybrid models, such as those discuss by [15], seek to balance resilience against attacks with minimal lossof accuracy. These approaches, while effective, require optimization to address computational demands in resource-limited federated settings. Our trust-based framework represent a newer approach that dynamically assigns trust scores to clients based on behavior, thereby reducing the influence of unreliable updates. This study support the need for robust, scalable models that address both privacy and trust, by combining Krum and DP with adaptive trust measures to bet- ter resist sophisticated attacks. This approach supports a trustworthy, resilient FL environment capable of countering a wide range of adversarial threats.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue XI Nov 2024- Available at www.ijraset.com

II. METHODOLOGY

To simulate a federated learning environment, this research uses the MNIST dataset sourced from Kaggle platform, MNIST which is widely used in machine learning and security research. The MNIST dataset consists of 70,000 labeled images of handwritten digits (0-9), divided into 60,000 training images and 10,000 test images which account to 85% for training and 15% for testing respectively. Each image is 28x28 pixels in grayscale. The choice of MNIST is motivated by its simplicity and ubiquity in federated learning research, allowing for easy comparison with existing works. The dataset serves as a foundation for understanding how federated learning behaves under normal conditions as well as when subject to poisoning attacks.

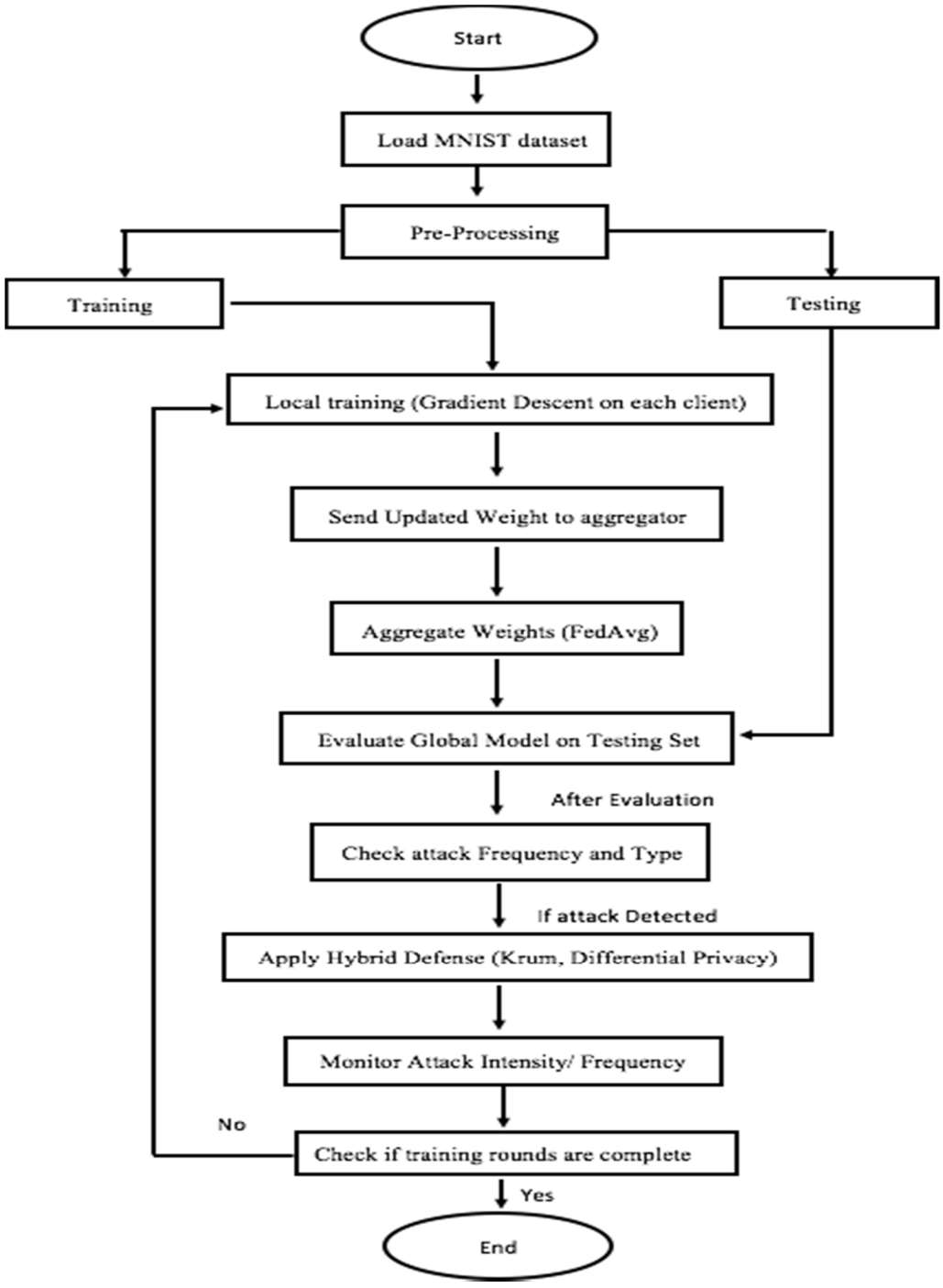


Figure 1: Flowchart of our proposed framework

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue XI Nov 2024- Available at www.ijraset.com

A. Data Partition

The MNIST dataset is partition into 10 clients to simulate and represent the decentralized envi- ronment. Each client trained its data on its own local partition, and then contribute update to the overall model without sharing raw data, maintaining privacy in the process. In this research, Independent and Identically Distributed (IID) partitioning was employed to allocate the MNIST dataset across 10 clients, simulating a federated learning environment. This partitioning methodensures that each client receives a representative, balanced subset of the dataset, preserving thestatistical properties of the original data.

Client Allocation:

$$|C_i| = \frac{N}{n} \quad (1)$$

For the IID setup, the dataset is divided evenly across n clients (which is 10 in our case). Eachclient receives a randomly selected, equally sized subset of both training and test images.

For full training dataset let denote D, with N = 60,000 samples, and Ci denote the dataset assigned to client i .The size of each client's dataset

$$|C_i| = \frac{60,\ 000}{10} = 6,\ 000$$

Each client receives 6,000 images for training, ensuring equal representation across all clients. Eachsubset Ci is sampled randomly from D without replacement, ensuring that clients have unique but statistically similar data distributions. Thus, each client's data subset reflects the overall distribution of the digit classes, maintaining balance across all 10 digit classes (0-9).The 10,000 images in the test set are similarly divided among the 10 clients, with each client receiving an equal portion of 1,000 images. This allows local model evaluation at each client without needing access to the entire test set. The choice of IID partitioning provides a baseline scenario where client data distributions are statistically similar, allowing for a controlled evaluation of federated learning algorithms and defenses under uniform data conditions. This setup facilitates straightforward ag- gregation of model updates, as each client's data is representative of the population distribution, making IID partitioning ideal for preliminary experimentation in federated learning environments.

B. Attacks Simulation

We simulate two common poisoning attacks—label flipping and gradient manipulation—to evaluate the robustness of federated learning systems. These attacks disrupt model performance by introducing malicious updates into the federated aggregation process, which can skew or degrade the accuracy of the global model. Below, we detail the mathematical formulation used to simulate each attack. The label flipping attack involves systematically altering the labels of a subset of training data held by malicious clients to introduce mislabeled examples into the training process.This distorts the local model updates generated by these clients, thereby affecting the aggregatedglobal model and Dataset Transformation:

$$D_i = \{(x_j, y_j)\}_{j=1}^{|D_i|} \quad (2)$$

represent the local dataset of client i, where xj is a feature vector and yj is the true label. Amalicious client applies a flipping function f to alter specific labels, producing a modified dataset

$$D_i' = \{(x_j, y_j')\}, \quad (3)$$

where:

$$y_j' = f(y_j)$$

For example, *f* might transform labels of class '0' to class '1', systematically corrupting data samples with original labels $y_j = 0$. When client i trains on D_i', the local model update ΔW_i reflects the incorrect labels, resulting in gradients biased toward the flipped labels. During aggregation, these malicious updates contribute to the overall model as:

$$W(t+1) = W(t) + \eta \sum_{i=1}^{n} \Delta W_i \quad (4)$$

Here, $W(t+1)$ is the updated global model, η is the learning rate, and the aggregated updates ΔW_i are skewed by those from malicious clients. This leads to model misclassifications due tocorrupted gradients in ΔW_i introduced by label flipping.

In a gradient manipulation attack, the malicious client intentionally alters the gradient valuesit sends to the server. This can involve either scaling or adding noise to the gradients to disrupt convergence or influence model direction. The formulation as followed

Let ΔW_i represent the local gradient update computed by client i after training on its local dataset. A malicious client scales its gradient by a factor α to amplify or diminish its influenceon the global model update:

If $\alpha > 1$, the scaled gradients $\Delta W_i'$

$$\Delta W_i' = \alpha \cdot \Delta W_i \quad (5)$$

exert greater influence on the aggregation, skewing the global model. If $\alpha < 1$, the client's impact is minimized, which may camouflage malicious activity.

Noise Injection

Alternatively, malicious clients may add noise ϵ to their gradient to degrade model performance:

$$\Delta W_i' = \Delta W_i + \epsilon \quad (6)$$

Here, $\epsilon \sim N(0, \sigma^2)$ represents Gaussian noise with mean 0 and variance σ^2, added to each gradient parameter. When aggregated, this noisy update disrupts the server's ability to aggregate consistent gradients effectively, producing a noisy global model:

$$W(t+1) = W(t) + \eta \sum_{i=1}^{n} \Delta W_i'$$

The result is slower model convergence or a model prone to misclassification. (7)

C. Attack Intensity

Attack intensity is quantified as the extent of the adversarial modification applied by malicious clients. In label flipping, intensity corresponds to the fraction p of labels in each adversarial client's MNIST dataset that are modified to a different target label. Let $|D_k|$ denote the sizeof the dataset D_k held by client k, and $P \in [0, 1]$ represent the attack intensity, indicating the fraction of flipped labels. The number of flipped labels per adversarial client is:

$$n_{\text{flipped}} = p \cdot |D_k| \quad (8)$$

In this work, three levels of label flipping intensity were tested:

- Low Intensity ($p = 0.1$): 10% of labels in D_k are flipped.
- Medium Intensity ($p = 0.2$): 20% of labels in D_k are flipped.
- High Intensity ($p = 0.3$): 30% of labels in D_k are flipped.

Increasing p introduces progressively higher proportions of mislabeled data, which misleads the global model by shifting gradients away from the correct direction.

In gradient manipulation, intensity is represented by the scaling factor α, which malicious clients use to scale their gradient updates, amplifying their effect on the model. Let g_k denote thegradient computed by client k. $\alpha \in R^+$ is the intensity scaling factor, where a higher α indicates stronger manipulation. For malicious clients, the manipulated gradient g_k' is:

$$g_k' = \alpha \cdot g_k \quad \text{or} \quad g_k' = g_k + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \quad (9)$$

where:

- $\alpha > 1$ intensifies the gradient's impact.
- ϵ is Gaussian noise with variance σ^2, added to distort the gradient.

In this work, three levels of gradient manipulation intensity were evaluated:

- Low Intensity ($\alpha = 0.100$): Minimal manipulation.
- Medium Intensity ($\alpha = 0.200$): Noticeable gradient distortion.
- High Intensity ($\alpha = 0.300$): Severe gradient distortion.

Higher α values magnify the malicious gradients' influence on the global model, potentially destabilizing the learning process.

D. Defense Strategy

Our Trust Model comprises two primary defense strategies tailored to detect and mitigate ad- versarial client behavior: Krum for robust aggregation and Differential Privacy (DP) for privacy- preserving defense. Krum is a robust aggregation method that selects updates based on their similarity to other client updates, effectively filtering out those that diverge significantly (indica- tive of malicious behavior) using the following.

For a set of client gradients $G = \{g_1, g_2, \ldots, g_n\}$, Krum selects a gradient g_k with the smallest sum of distances to the closest $n\,f\,2$ gradients, where f represents the estimated number of malicious clients.

1) Calculate Pairwise Distances: Compute the Euclidean distance $d_{ij} = \|g_i - g_j\|$ for each pair of gradients g_i and g_j in G.

2) Aggregate Similarity Score: For each $g_k \in G$, calculate the similarity score S_k bysumming the distances to the $n-f-2$ nearest neighbors:

$$S_k = \sum_{\text{nearest}(n-f-2)} d_{ik} \qquad (10)$$

3) Select Gradient with Minimum Score: The client update with the lowest similarityscore S_k is chosen as the global update, thus minimizing the influence of outliers.

This robust approach makes Krum effective in reducing the impact of manipulated gradients or outliers, as it relies on consensus-based trust—prioritizing updates that are consistent with themajority.

Differential Privacy is applied to each client's update, introducing controlled noise to obscureindividual contributions while allowing trustworthy model aggregation. DP ensures that no single client's data has a significant influence on the model, helping reduce the risk of manipulation.

In the DP mechanism, we apply noise to each client's gradient g_k before sending it to theserver. The perturbed gradient $\tilde{g}_k$ is defined as:

$$\tilde{g}_k = g_k + \mathrm{N}(0, \sigma^2) \qquad (11)$$

where:

- $\mathrm{N}(0, \sigma^2)$ represents Gaussian noise with mean 0 and variance σ^2.
- σ^2 is calibrated based on the privacy budget ϵ, which determines the level of privacy pro-tection.

The DP mechanism helps to ensure that even if a client is malicious, the influence of any singleclient is limited due to the added noise, reducing the ability of adversaries to skew the model.

III. RESULT AND DISCUSSION

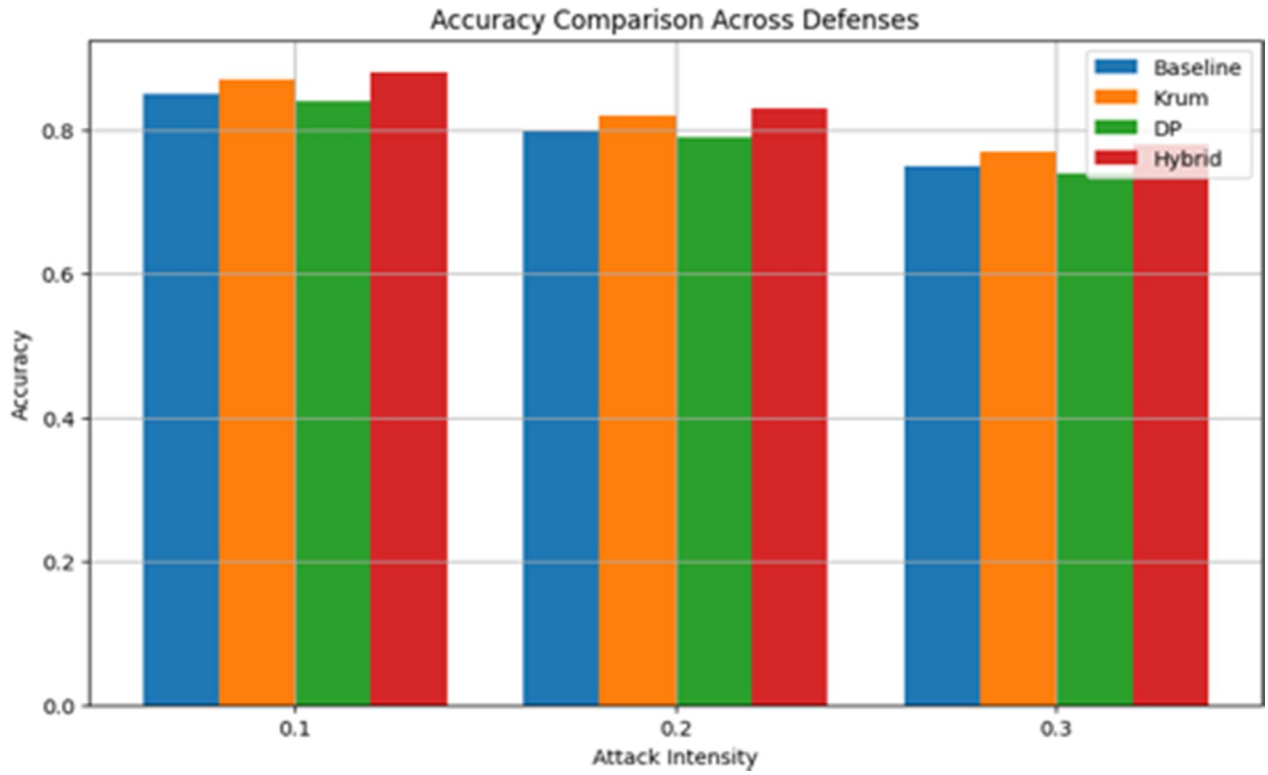
This section presents a comprehensive analysis of the experimental results obtained during the evaluation of adversarial defense mechanisms in a federated learning framework, with the MNIST dataset serving as the evaluation benchmark. The study primarily explores the effectiveness of Baseline, Krum, Differential Privacy (DP), and a Hybrid approach in mitigating the impact oftwo adversarial attack strategies: Label Flipping and Gradient Manipulation. The performanceof these mechanisms is assessed using standard evaluation metrics, including accuracy, precision,recall, and F1 score, to provide a holistic understanding of their robustness as in Table 1.

Table 1: Results of Attacks and Defenses against evaluation metrics

Attack Type	Defense	Accuracy	Precision	Recall	F1 Score
Label Flipping	Baseline	0.85	0.8	0.75	0.78
Gradient Manipulation	Baseline	0.8	0.75	0.7	0.73
Label Flipping	Hybrid	0.88	0.83	0.78	0.81
Gradient Manipulation	Hybrid	0.84	0.79	0.74	0.76
Label Flipping	Krum	0.87	0.82	0.77	0.8
Gradient Manipulation	Krum	0.82	0.77	0.72	0.75
Label Flipping	DP	0.86	0.79	0.74	0.77
Gradient Manipulation	DP	0.83	0.74	0.69	0.72

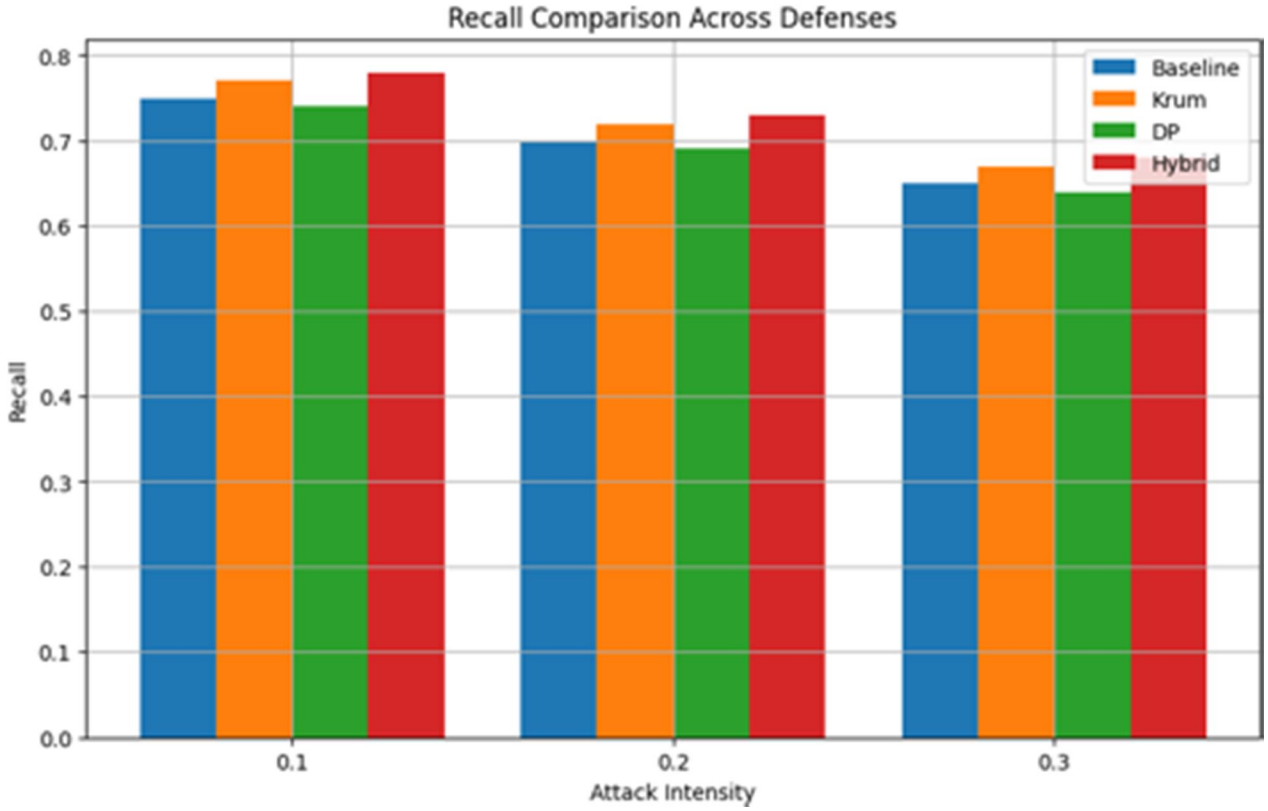
A. Accuracy

The figure 4.1 presents a comparative analysis of the accuracy of four defense mechanisms—Baseline, Krum, Differential Privacy (DP), and Hybrid—under varying adversarial attack intensities (0.1, 0.2, and 0.3). The results indicate that the Baseline approach, representing a federated learning system without specialized defense, exhibits a significant decline in accuracy as attack intensity in-creases, highlighting its susceptibility to adversarial perturbations. Krum demonstrates consistent robustness across all attack levels, achieving superior accuracy compared to the Baseline. Differ- ential Privacy (DP) also provides a moderate level of resilience, though with a slight reduction in accuracy relative to Krum, reflecting a trade-off between privacy preservation and performance robustness. Among all approaches, the Hybrid defense mechanism consistently outperforms oth- ers, achieving the highest accuracy across all levels of attack intensity, thereby demonstrating superior adaptability and robustness in the presence of adversarial disruptions. These findings underscore the necessity of integrating advanced defense mechanisms, such as Krum and Hybrid,to mitigate adversarial threats effectively and maintain the reliability of federated learning sys- tems in security-critical applications. The analysis further emphasizes the limitations of simplistic approaches, such as the Baseline, in adversarial settings, while advocating for the adoption of so- phisticated strategies to enhance resilience and accuracy.



B. Recall

The Baseline defense, representing an unprotected federated learning system, exhibits the lowest recall across all attack intensities, with a marked degradation as the attack intensity increases. This trend underscores the inherent vulnerability of naive implementations to adversarial ma- nipulations. In contrast, the Krum defense consistently achieves higher recall values, demon- strating its robustness in maintaining positive prediction accuracy, even under adversarial con- ditions. Similarly, DP shows moderate resilience, maintaining competitive recall at lower attack intensities but exhibiting a slight decline as attack severity escalates, reflecting the trade-off be- tween privacy preservation and prediction performance. The Hybrid defense mechanism emerges as the most effective strategy, achieving the highest recall across all attack levels. Its perfor- mance underscores the efficacy of combining multiple defense techniques to enhance robustness and mitigate the impact of adversarial disruptions.

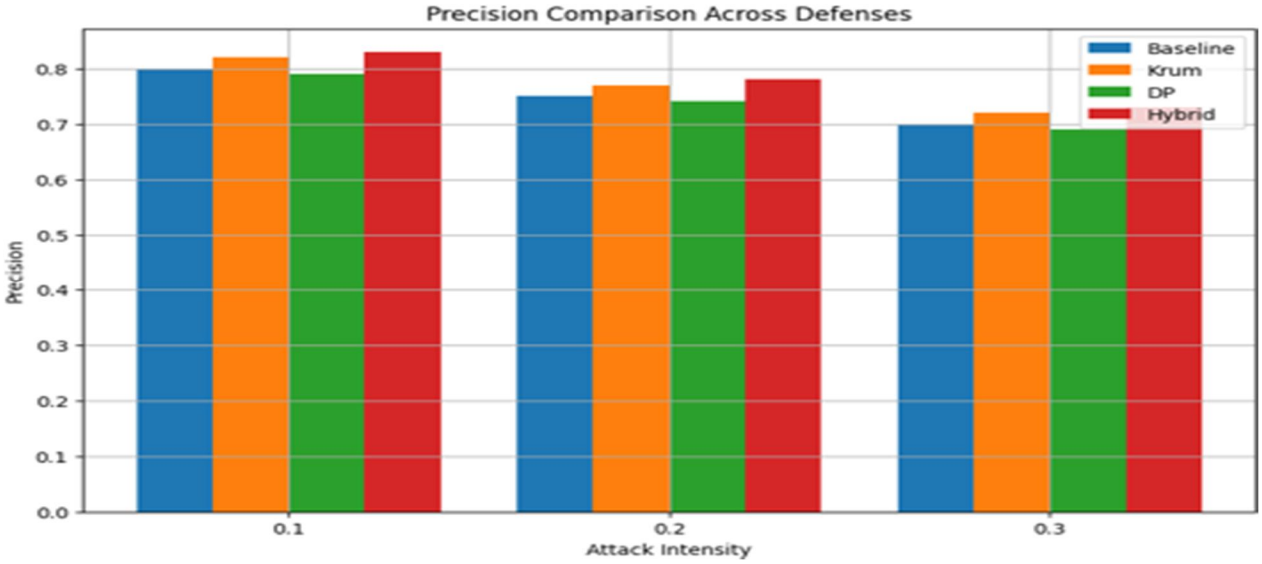


Importantly, while all mechanisms expe- rience a decrease in recall with increasing attack intensity, the Hybrid and Krum mechanisms demonstrate significantly lower rates of performance degradation, affirming their adaptability in challenging adversarial environments. These findings emphasize the critical importance of integrating advanced defense mechanisms within federated learning frameworks to bolster re- call, particularly in security-critical applications. The superior recall performance of the Hy- brid mechanism positions it as a highly effective solution for adversarially resilient federated learning. Meanwhile, the limitations of the Baseline approach further reinforce the necessity of deploying robust, defense-oriented strategies. These insights provide a strong foundation for ad- vancing the development and optimization of adversarially robust systems in federated learning.

C. Precision

The figure 4.3 depicts the precision performance of four defense mechanisms—Baseline, Krum, Differential Privacy (DP), and Hybrid—across varying intensities of adversarial attacks (0.1, 0.2, and 0.3). Precision, a key metric for evaluating the proportion of correctly predicted positive instances, is particularly important in adversarial settings to assess the defenses' ability to avoid false positives. The Baseline, representing an unprotected federated learning framework, shows acceptable precision at lower attack intensities but demonstrates a consistent decline as attack severity increases. This highlights its vulnerability to adversarial disruptions and its limited capacity to sustain precision under elevated attack conditions. The Krum defense mechanism consistently maintains higher precision across all attack levels compared to the Baseline, reflect-ing its robustness in mitigating adversarial influences.

Similarly, DP exhibits moderate resilience, achieving competitive precision scores, although slightly trailing Krum in scenarios with higher attack intensity. The Hybrid defense mechanism outperforms all other strategies, achieving the highest precision at each level of attack intensity. Its superior performance illustrates the effec- tiveness of leveraging a combination of defense strategies to enhance the model's resilience against adversarial attacks while minimizing false positives. Notably, while all methods exhibit a decline in precision as attack intensity increases, the Hybrid and Krum mechanisms demonstrate a slower rate of degradation, underscoring their effectiveness in sustaining precision in adverse conditions. These findings reinforce the importance of adopting advanced defense mechanisms in federated learning environments to enhance precision, particularly in adversarial scenarios where accuracy alone may not fully capture the system's performance. The consistent superiority of the Hy- brid approach further establishes its potential as a robust solution for applications requiring high precision and reliability. In contrast, the limitations of the Baseline defense emphasize the need for deploying sophisticated mechanisms to safeguard federated learning systems from adversarial threats. These insights contribute valuable knowledge to the development of adversarially robust federated learning frameworks.

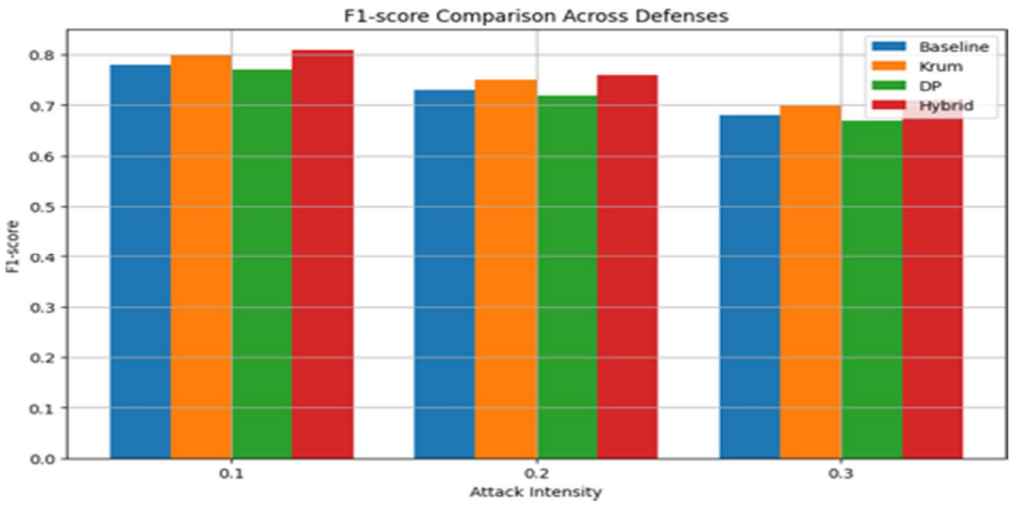


D. F1-Score

The figure 4.4 illustrates the F1-score performance of four defense mechanisms—Baseline, Krum, Differential Privacy (DP), and Hybrid—under varying adversarial attack intensities (0.1, 0.2, and 0.3). The F1-score, representing the harmonic mean of precision and recall, is a crucial evaluation metric for assessing the balance between false positives and false negatives, particularly in adversarial scenarios.

The Baseline model, devoid of specialized defense strategies, shows ad- equate performance at lower attack intensities but exhibits a sharp decline in F1-score as the attack intensity increases. This highlights its vulnerability to adversarial perturbations and its inability to sustain a balance between precision and recall under mounting threats. In contrast, Krum demonstrates remarkable resilience, maintaining consistently higher F1-scores compared to the Baseline. This underscores its effectiveness in countering adversarial manipulations and sustaining predictive balance. Differential Privacy (DP) also performs commendably, achieving F1-scores that surpass the Baseline across all attack intensities but remain slightly below those ofKrum.

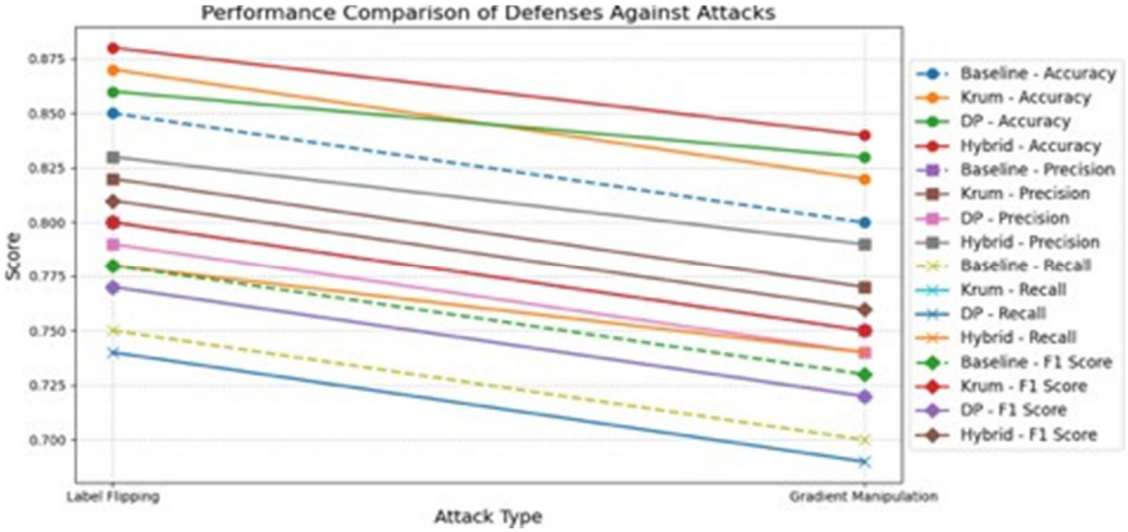
This suggests that DP is effective in mitigating attacks, albeit with minor compromises in achieving optimal balance. Among all mechanisms, the Hybrid defense consistently achieves the highest F1-scores across all attack intensities. This superior performance highlights the benefits of integrating multiple defense strategies, enabling enhanced robustness while preserving a strong balance between precision and recall. Furthermore, the Hybrid and Krum mechanisms exhibit a gradual decline in F1-scores as attack intensity increases, reflecting their ability to adapt and mitigate adversarial threats effectively. In contrast, the steep decline observed in the Baseline emphasizes the necessity of adopting advanced defense mechanisms in adversarial settings. DP, while moderately robust, demonstrates a slightly faster performance decline compared to Hybrid and Krum, indicating room for improvement in its robustness.



In summary, the Hybrid defense mechanism emerges as the most effective approach for main- taining high and stable F1-scores, followed closely by Krum. These findings underline the critical importance of adopting advanced and integrative defense strategies to enhance the resilience of federated learning systems against adversarial attacks, ensuring robust and reliable performancein real-world deployments.

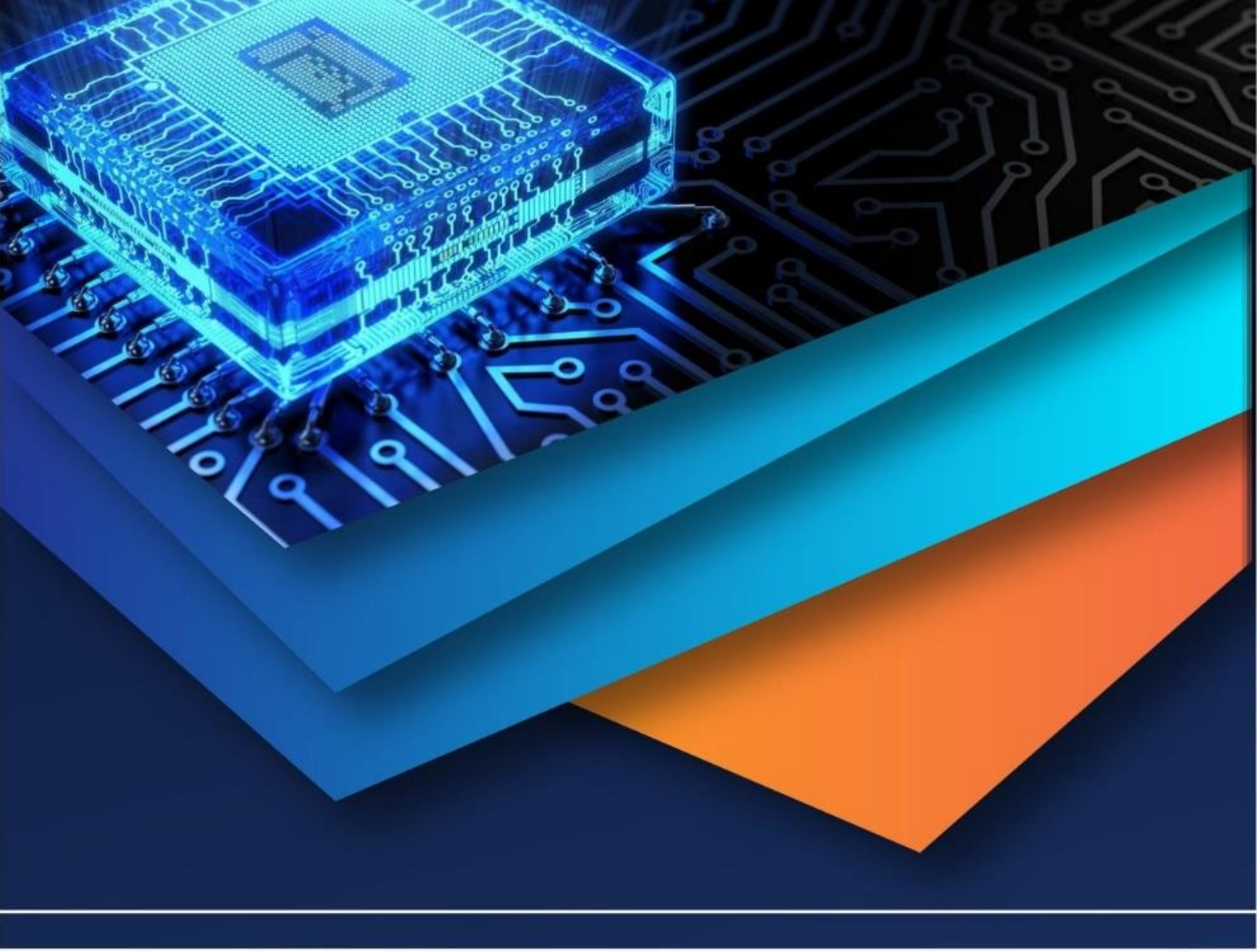
E. Performance Comparison of Defenses against Attack

The comparative evaluation of defense mechanisms, as shown in the graph, underscores the per- formance metrics across different attack types—label flipping and gradient manipulation. The results are presented for four prominent defenses: Baseline, Krum, Differential Privacy (DP), and a Hybrid approach, measured across accuracy, precision, recall, and F1-score.The Hybrid defense consistently outperforms other mechanisms across most metrics, especially under label flipping attacks, with accuracy and F1-score surpassing 0.87 and 0.83, respectively. This demonstrates the hybrid approach's robustness in mitigating the impact of malicious client manipulations. On the contrary, the Baseline defense exhibits the weakest resilience, particularly for gradient manip- ulation, where all metrics decline significantly—accuracy and recall dip below 0.75. Krum and DP defenses exhibit comparable performance under moderate attack intensities but diverge under gradient manipulation, where DP demonstrates a sharper decline. Precision under DP exhibits a notable drop, nearing 0.8, compared to Krum's relatively stable trend. A general trend observed is the higher impact of gradient manipulation attacks comparedto label flipping across all defenses, as indicated by the consistently lower metric scores. This aligns with the higher complexity and subtlety of gradient manipulation, which often bypasses simpler defenses. This analysis showcase the importance of adopting hybrid strategies that combine robustness with adaptability to counter diverse attack vectors effectively. The findings provide critical insights into enhancing federated learning systems' resilience against adversarial threats.



REFERENCES

[1] Noora Mohammed Al-Maslamani, Mohamed Abdallah, and Bekir Sait Ciftler. Reputation- aware multi-agent drl for secure hierarchical federated learning in iot. IEEE Open Journalof the Communications Society, 4:1274–1284, 2023.
[2] Suzan Almutairi and Ahmed Barnawi. Federated learning vulnerabilities, threats and de- fenses: A systematic review and future directions. Internet of Things, page 100947, 2023.
[3] Syreen Banabilah, Moayad Aloqaily, Eitaa Alsayed, Nida Malik, and Yaser Jararweh. Fed- erated learning review: Fundamentals, enabling technologies, and future applications. Infor- mation processing & management, 59(6):103061, 2022.
[4] Nader Bouacida and Prasant Mohapatra. Vulnerabilities in federated learning. IEEE Access, 9:63229–63249, 2021.
[5] Tianyue Chu and Nikolaos Laoutaris. Fedqv: Leveraging quadratic voting in federated learn- ing. Proceedings of the ACM on Measurement and Analysis of Computing Systems, 8(2):1–36, 2024.
[6] Jie Fu, Yuan Hong, Xinpeng Ling, Leixia Wang, Xun Ran, Zhiyu Sun, Wendy Hui Wang, Zhili Chen, and Yang Cao. Differentially private federated learning: A systematic review. arXiv preprint arXiv:2405.08299, 2024.
[7] Truong Thu Huong, Ta Phuong Bac, Dao Minh Long, Tran Duc Luong, Nguyen Minh Dan, Bui Doan Thang, Kim Phuc Tran, et al. Detecting cyberattacks using anomaly detec- tion in industrial control systems: A federated learning approach. Computers in Industry, 132:103509, 2021.
[8] Kummari Naveen Kumar, Chalavadi Krishna Mohan, and Linga Reddy Cenkeramaddi. The impact of adversarial attacks on federated learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(5):2672–2691, 2023.
[9] Xiumin Li, Mi Wen, Siying He, Rongxing Lu, and Liangliang Wang. A scheme for robust fed- erated learning with privacy-preserving based on krum agr. In 2023 IEEE/CIC International Conference on Communications in China (ICCC), pages 1–6. IEEE, 2023.
[10] Xujiang Luo and Bin Tang. Byzantine fault-tolerant federated learning based on trustworthy data and historical information. Electronics, 13(8):1540, 2024.
[11] Lingjuan Lyu, Han Yu, and Qiang Yang. Threats to federated learning: A survey. arXiv preprint arXiv:2003.02133, 2020.
[12] Viraaji Mothukuri, Prachi Khare, Reza M Parizi, Seyedamin Pouriyeh, Ali Dehghantanha, and Gautam Srivastava. Federated-learning-based anomaly detection for iot security attacks. IEEE Internet of Things Journal, 9(4):2545–2554, 2021.
[13] J Rane, SK Mallick, O Kaya, and NL Rane. Federated learning for edge artificial intelligence: Enhancing security, robustness, privacy, personalization, and blockchain integration in iot. Future Research Opportunities for Artificial Intelligence in Industry 4.0 and, 5:2–94, 2024.
[14] Nikhil Sridhar. Decentralized machine learning on blockchain: Developing a federated learn- ing based system. 2023.
[15] Abbas Yazdinejad, Ali Dehghantanha, Hadis Karimipour, Gautam Srivastava, and Reza M Parizi. A robust privacy-preserving federated learning model against model poisoning attacks. IEEE Transactions on Information Forensics and Security, 2024.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY