International Journal of
Science, Architecture, Technology and Environment

Adversarial Attacks and Defenses in Federated Learning: A State of the Art Survey on Security Vulnerabilities and Mitigation Approaches

Funminiyi Olagunju[1*]

[1]*Department of Electrical Engineering, North Carolina A & T State University*

[*]*Corrusponding author*

DOI: https://doi.org/10.63680/ijsate0524110.08

Abstract

Federated Learning (FL) enables collaborative model training across decentralized devices while preserving data privacy, making it ideal for applications in IoT, healthcare, and autonomous systems. However, FL's distributed nature exposes it to a wide range of adversarial attacks, including data and model poisoning, privacy inference, and sophisticated collusion strategies, which threaten the integrity and confidentiality of the learning process. This survey comprehensively reviews the current state of the art adversarial threats and defense mechanisms in FL. We discuss robust aggregation techniques, anomaly detection, privacy preserving methods, and trust-based frameworks that enhance FL's resilience. Additionally, we explore evaluation metrics, real world use cases, and open challenges, highlighting future research directions such as adaptive defenses, integration with emerging technologies, and benchmarking standardization. This work aims to provide researchers and practitioners with a detailed understanding of FL's security landscape and guide the development of more secure and trustworthy federated systems.

Keywords: Federated Learning, Adversarial Attacks, Data Poisoning, Model Poisoning, Privacy Attacks

1. Introduction

Federated Learning (FL) has emerged as a transformative machine learning paradigm that enables collaborative model training across multiple decentralized devices without directly sharing raw data, thereby preserving privacy and reducing communication costs (McMahan et al., 2017). As FL gains adoption in privacy sensitive domains such as healthcare, finance, and autonomous systems, its decentralized nature introduces unique security vulnerabilities not present in traditional centralized learning (Kairouz et al., 2021). The distributed and partially trusted environment of FL makes it an attractive target for adversaries seeking to disrupt training or extract sensitive information. Various adversarial strategies have been identified, including data poisoning, model poisoning, and inference attacks, each capable of degrading model performance or leaking private data (Bagdasaryan et al., 2020; Hitaj et al., 2017; Zhao et al., 2020). Traditional security mechanisms, such as secure aggregation and differential privacy, are insufficient against sophisticated or coordinated threats, necessitating new, FL specific defense architectures (Bonawitz et al., 2017; Geyer et al.,

2017). Moreover, the presence of heterogeneous data, intermittent connectivity, and limited device resources further complicates the deployment of effective mitigation strategies in real world federated systems (Li et al., 2020; Wang et al., 2021). This survey seeks to provide a comprehensive review of the current landscape of adversarial attacks and defenses in FL, identifying critical vulnerabilities, summarizing state of the art protection techniques, and highlighting emerging challenges and future research directions for secure and trustworthy federated learning.

2.0 Background and Preliminaries

Federated Learning (FL) is a decentralized machine learning paradigm that enables multiple clients such as smartphones, IoT devices, or edge nodes to collaboratively train a shared global model while retaining their data locally, thereby preserving user privacy and reducing communication overhead (McMahan et al., 2017). In its typical **centralized architecture**, a central server coordinates the training by distributing the initial model to all participating clients, who train the model on their local datasets and return updates (usually gradients or model weights), which are then aggregated (e.g., using FedAvg) to update the global model (Kairouz et al., 2021). In contrast, **decentralized architectures** eliminate the central server and rely on peer to peer communication and aggregation, increasing resilience but complicating coordination (Li et al., 2020). The **FL lifecycle** involves multiple rounds of communication between the server and clients, encompassing model distribution, local training, gradient aggregation, and model updating. Common terminology includes *clients* (data holders), *server* (orchestrator of the learning process), *rounds* (iterations of training and aggregation), and *aggregation* (combining local updates into the global model). From a security perspective, FL systems are vulnerable to various threat models. The **honest but curious** adversary assumes the server or clients follow the protocol but attempt to infer private information from the data or model updates (Geyer et al., 2017). **Malicious clients or servers**, on the other hand, actively try to manipulate the learning process through poisoning or tampering with data or updates (Bagdasaryan et al., 2020). Another significant threat arises from **colluding clients**, which can work together to infer private data or reverse engineer the global model (Nasr et al., 2019). The **adversarial goals** in FL typically include *accuracy degradation* (e.g., lowering model performance), *privacy breaches* (e.g., inferring training data or labels), and *stealth* (executing attacks without being detected). Understanding these foundational concepts is critical to assessing both the vulnerabilities and the defenses within federated learning systems.

3.0 Discussion

Adversarial Attacks in Federated Learning

Federated Learning (FL) has emerged as a promising paradigm for decentralized machine learning, enabling collaborative model training while preserving data privacy. However, its distributed nature introduces unique security challenges, making it susceptible to various adversarial attacks. These attacks can be broadly categorized into data poisoning attacks, model poisoning attacks, inference and privacy attacks, and advanced emerging threats. Each category presents distinct mechanisms and objectives, necessitating a comprehensive understanding to develop effective defense strategies.

3.1 Data Poisoning Attacks

Data poisoning attacks involve the manipulation of the training data by malicious clients to degrade the performance of the global model. These attacks can be further classified into label flipping attacks, backdoor attacks, feature poisoning, and can be targeted or untargeted in nature.

Label Flipping Attacks

Label flipping attacks are a form of targeted data poisoning where an adversary alters the labels of specific data points in their local dataset. For instance, in a classification task, a malicious client might change the label of an image of a dog from "dog" to "cat." This mislabeling introduces noise into the training process, leading the global model to learn incorrect associations. Such attacks are particularly effective in FL systems due to their simplicity and the limited visibility of the central server into individual client data. The impact of label flipping attacks can be significant, especially when the adversary has control over a substantial portion of the data.

Backdoor Attacks

Backdoor attacks involve embedding a hidden trigger within the training data that causes the model to behave incorrectly when the trigger is present, while maintaining normal performance on other inputs. In FL, an attacker might introduce a specific pattern or feature into their local dataset, ensuring that the global model learns to associate this pattern with a particular output. During inference, when the model encounters the trigger, it produces the attacker's desired outcome. These attacks are stealthy and challenging to detect, as they do not degrade the model's overall performance but can be exploited under specific conditions.

Feature Poisoning

Feature poisoning attacks target the features of the training data rather than the labels. By introducing malicious features into their local dataset, an adversary can influence the global model to learn biased or incorrect representations. For example, adding irrelevant or misleading features can cause the model to focus on spurious correlations, leading to poor generalization. Feature poisoning attacks are subtle and can be difficult to distinguish from legitimate data variations, posing a significant threat to the integrity of FL systems.

Targeted vs. Untargeted Attacks

Data poisoning attacks can be classified as targeted or untargeted based on the attacker's objectives. Targeted attacks aim to degrade the model's performance on specific classes or tasks, while untargeted attacks seek to reduce the overall accuracy of the model indiscriminately. Both types of attacks can be detrimental to the performance of the global model, but targeted attacks are often more challenging to detect and mitigate due to their specificity.

3.2 Model Poisoning Attacks

Model poisoning attacks involve malicious clients submitting compromised model updates to the central server, aiming to corrupt the global model. These attacks can manifest as gradient manipulation, Byzantine behaviors, scaling and sign flipping attacks, and can lead to convergence degradation.

Gradient Manipulation

In gradient manipulation attacks, an adversary alters the gradients computed during local training before sending them to the server. By modifying the gradients, the attacker can influence the direction in which the global model is updated, steering it towards suboptimal solutions. This can result in poor model performance and hinder the convergence process. Gradient manipulation is particularly effective in FL systems where clients have significant autonomy over their local training processes.

Byzantine Behaviors

Byzantine attacks refer to scenarios where malicious clients behave arbitrarily, sending incorrect or malicious updates to the server. These behaviors can include sending random gradients, duplicating updates, or introducing noise into the training process. Byzantine attacks are challenging to defend against because the server cannot distinguish between benign and malicious updates based solely on the content of the updates. Robust aggregation techniques are often employed to mitigate the impact of Byzantine behaviors.

Scaling and Sign Flipping Attacks

Scaling and sign flipping attacks involve manipulating the magnitude and direction of the model updates. By scaling the updates or flipping their signs, an adversary can cause the global model to converge to undesirable solutions. These attacks exploit the aggregation process, where updates from multiple clients are combined, to introduce significant deviations in the global model. Detecting and mitigating such attacks require careful monitoring of update patterns and the implementation of robust aggregation methods.

Convergence Degradation

Convergence degradation occurs when malicious clients introduce updates that slow down or prevent the global model from converging to an optimal solution. This can be achieved through various means, such as introducing noise into the updates or submitting updates that are inconsistent with the majority. Convergence degradation attacks undermine the efficiency of the FL process and can lead to prolonged training times and suboptimal model performance.

3.3 Inference and Privacy Attacks

Inference and privacy attacks aim to extract sensitive information from the model or its updates, compromising the privacy of the participating clients. These attacks include membership inference, model inversion, gradient leakage, and side channel attacks.

Membership Inference

Membership inference attacks involve determining whether a particular data point was included in the training dataset. By analyzing the model's responses to specific inputs, an adversary can infer the presence of a data point in the training set. This poses significant privacy risks, especially in sensitive domains like healthcare, where the inclusion of certain data points can reveal personal information.

Model Inversion

Model inversion attacks aim to reconstruct the training data by exploiting the information encoded in the global model. By querying the model and analyzing its outputs, an adversary can infer details about the data used for training. This can lead to the exposure of sensitive information, such as personal attributes or confidential data, undermining the privacy guarantees of FL systems.

Gradient Leakage

Gradient leakage attacks involve extracting information about the training data by analyzing the gradients shared during the FL process. Since gradients reflect the influence of individual data points on the model, malicious clients can use this information to infer details about the training data. Gradient leakage poses a significant threat to the privacy of clients, as it can lead to the exposure of sensitive information without direct access to the data.

Side Channel Attacks

Side channel attacks exploit indirect information leaks to infer details about the training data or model. These can include timing variations, power consumption patterns, or communication delays that reveal information about the model's operations. Side channel attacks are particularly challenging to defend against because they exploit subtle, often unintended, information leaks inherent in the system's operation.

3.4 Advanced and Emerging Threats

Sybil Attacks

Sybil attacks pose a significant challenge in FL by allowing an adversary to create multiple fake identities, thereby gaining disproportionate influence over the model aggregation process. This is particularly concerning in decentralized FL settings where there is no central authority to validate client identities. The adversary can inject malicious updates from numerous Sybil clients, skewing the global model towards their objectives. To mitigate such attacks, various defense mechanisms have been proposed. For instance, FoolsGold identifies Sybil clients by analyzing the diversity of their model updates, assuming that genuine clients' updates will exhibit more variability than those of Sybil clients. Additionally, SybilWall employs a Sybil resistant aggregation function based on similarity between clients' updates and a probabilistic gossiping mechanism to enhance resilience against Sybil attacks in decentralized FL environments.

Free Rider Clients

Free rider clients are participants who benefit from the global model without contributing to its training. These clients can degrade the performance of the FL system by not providing meaningful updates, thereby reducing the overall utility of the model. To address this issue, incentive mechanisms have been proposed to encourage active participation. For example, reputation based systems can be implemented where clients are rewarded based on the quality and frequency of their updates. Additionally, game theoretic approaches can be employed to model the interactions among clients and design strategies that promote truthful reporting and active participation .

Collusion Among Clients

Collusion among clients occurs when multiple participants collaborate to manipulate the FL process for malicious purposes. This can involve coordinating attacks such as label flipping or backdoor insertion to influence the global model. Detecting and mitigating collusion is challenging due to the distributed nature of FL and the lack of centralized oversight. One approach to counteract collusion is the use of anomaly detection techniques to identify unusual patterns in client updates that may indicate coordinated malicious behavior. Another strategy involves employing robust aggregation methods that can tolerate a certain percentage of malicious updates, thereby reducing the impact of collusion on the global model .

Adaptive Adversaries

Adaptive adversaries are malicious entities that can learn and adjust their attack strategies based on the defenses employed by the FL system. These adversaries can modify their behavior to bypass detection mechanisms, making them particularly difficult to counter. To defend against adaptive adversaries, it is essential to implement dynamic and evolving defense strategies. For instance, the Metric Cascades (MESAS) method employs multiple detection metrics simultaneously to identify poisoned model updates, creating a complex multi objective optimization problem for adaptive attackers. This approach has been shown to be effective in detecting strong adaptive adversaries and distinguishing backdoors from data distribution related distortions .

4.0 Robust Aggregation Techniques

In Federated Learning (FL), the aggregation process combines model updates from multiple clients to form a global model. However, this process is vulnerable to adversarial attacks, such as data poisoning and model poisoning, where malicious clients inject harmful updates to degrade model performance. To mitigate these risks, several robust aggregation techniques have been proposed.

Krum

Krum is a robust aggregation method that selects the client update whose distance to the other updates is minimal, thereby identifying the most representative update among the clients. This approach is particularly effective in scenarios where a small number of clients are compromised, as it minimizes the influence of outliers on the global model.

Multi Krum

Multi Krum extends the Krum algorithm by selecting multiple updates that are closest to the majority, providing a more comprehensive defense against Byzantine failures. This method enhances the robustness of the aggregation process by considering multiple candidate updates, thereby reducing the likelihood of malicious updates affecting the global model.

Trimmed Mean and Median

Trimmed Mean and Median are statistical methods that involve removing a certain percentage of the highest and lowest values before computing the mean or median. These techniques are effective in reducing

the impact of extreme outliers, ensuring that the aggregated model is not unduly influenced by a small number of malicious updates.

Norm Bounding and Clipping

Norm Bounding involves setting a threshold for the L2 norm of client updates, scaling updates that exceed this threshold to the specified limit. Clipping similarly restricts the magnitude of updates to a predefined range. Both methods aim to prevent large, potentially malicious updates from disproportionately affecting the global model.

Adaptive and Weighted Aggregation

Adaptive and Weighted Aggregation techniques assign different weights to client updates based on their reliability or past performance. By giving more weight to trustworthy clients, these methods ensure that the global model is primarily influenced by accurate and consistent updates, enhancing its robustness against adversarial attacks.

4.1 Anomaly Detection and Filtering

Anomaly detection and filtering methods focus on identifying and mitigating suspicious or malicious client behavior during the training process. These techniques aim to detect deviations from expected client behavior and prevent the incorporation of harmful updates into the global model.

Clustering Based Detection (e.g., FoolsGold)

Clustering based detection methods, such as FoolsGold, analyze the similarity of client updates to identify potential adversaries. By grouping similar updates and identifying outliers, these methods can detect clients that deviate from the majority, indicating potential malicious behavior.

Update Similarity Detection

Update similarity detection involves comparing the updates from different clients to identify inconsistencies or anomalies. Significant deviations from the majority can indicate malicious activity, allowing for the exclusion of suspicious updates from the aggregation process.

Outlier Scoring Systems

Outlier scoring systems assign a score to each client update based on its deviation from the expected behavior. Updates with high outlier scores are considered suspicious and can be excluded from the aggregation process, thereby protecting the global model from potential adversarial influences.

4.2 Privacy Preserving Techniques

Privacy preserving techniques are essential in Federated Learning (FL) to ensure that sensitive client data remains confidential during the training process. These methods aim to protect individual privacy while enabling collaborative model training.

Differential Privacy in FL

Differential Privacy (DP) is a mathematical framework that provides strong privacy guarantees by introducing noise into the data or computations. In the context of FL, DP can be applied to model updates to prevent the leakage of sensitive information. By adding carefully calibrated noise, DP ensures that the inclusion or exclusion of a single data point does not significantly affect the output, thereby protecting individual privacy.

A systematic review by Fu et al. (2024) provides an overview of differentially private federated learning, categorizing various DP models and their applications in FL scenarios. The study highlights the importance of selecting appropriate DP mechanisms to balance privacy and model utility.

Secure Multi Party Computation (SMC)

Secure Multi Party Computation (SMC) enables multiple parties to compute a function over their inputs while keeping those inputs private. In FL, SMC can be used to perform model aggregation without revealing individual client updates. This ensures that the server cannot access sensitive information during the aggregation process.

The Danish Sugar Beet Auction, conducted in 2008, was the first large scale practical application of SMC. The auction involved representatives from Denmark's sugar beet processor, the growers' association, and a research group implementing the computation, demonstrating the feasibility of SMC in real world scenarios.

Homomorphic Encryption

Homomorphic Encryption (HE) allows computations to be performed on encrypted data without decrypting it. This property makes HE particularly useful in FL, as it enables secure aggregation of model updates without exposing individual client data.

IBM Research has integrated Fully Homomorphic Encryption (FHE) into its federated learning framework, allowing computations on encrypted data without revealing sensitive information. This approach enhances privacy by ensuring that the server cannot access decrypted model updates.

Trusted Execution Environments (TEEs)

Trusted Execution Environments (TEEs) are secure areas within a processor that execute code in isolation from the rest of the system. In FL, TEEs can be used to perform model aggregation securely, ensuring that the server cannot access sensitive client data during the process. A study published in 2025 explores the use of TEEs in FL, demonstrating their effectiveness in enhancing data security during model aggregation. The research highlights the potential of TEEs to provide a secure environment for FL operations.

4.3 Trust and Reputation Based Systems

Trust and reputation based systems aim to evaluate and ensure the reliability of clients participating in FL. These systems help identify trustworthy clients and mitigate the impact of malicious or unreliable participants.

Reputation Scoring of Clients

Reputation scoring involves assigning a score to each client based on their behavior and contributions to the FL system. Clients with high reputation scores are considered trustworthy and may have more influence in the model aggregation process.

Blockchain for Secure Update Logging and Validation

Blockchain technology provides a decentralized and immutable ledger for recording client updates. In FL, blockchain can be used to log model updates securely, ensuring transparency and accountability in the learning process.

Token Based Incentives to Encourage Honest Participation

Token based incentive mechanisms use digital tokens to reward clients for honest participation in the FL process. These incentives encourage clients to contribute accurate updates and discourage malicious behavior. As a study by Zhang et al. (2023) explores the use of token based incentives in FL, demonstrating their effectiveness in promoting honest participation and improving the overall performance of the learning system.

4.4 Hybrid Defense Strategies

In Federated Learning (FL), hybrid defense strategies integrate multiple security and privacy mechanisms to build robust systems that address diverse adversarial threats. These strategies combine aggregation robustness, privacy preservation, and intelligent client management to optimize both model accuracy and security in resource constrained, adversarial environments.

Combined Aggregation and Privacy Preserving Systems

A key hybrid approach involves blending robust aggregation methods with privacy preserving techniques such as differential privacy, secure multi party computation (SMC), or homomorphic encryption. Robust aggregation techniques like Krum or Trimmed Mean protect the global model from poisoning attacks by mitigating the influence of outlier or malicious client updates. However, these methods alone do not guarantee client data privacy. To address this, privacy preserving schemes are incorporated, which encrypt or obfuscate updates before aggregation, ensuring that sensitive information is not leaked.

For example, Bonawitz et al. (2017) combined secure aggregation protocols with differential privacy in a federated learning system, enabling the aggregation of encrypted client updates while adding noise to preserve privacy. This dual layer defense mechanism enhances both security against adversarial attacks and privacy protection, while still allowing effective global model training.

Similarly, Phong et al. (2018) explored secure aggregation combined with homomorphic encryption, ensuring that server side aggregation is performed on encrypted data without decrypting individual client updates. Such hybrid techniques are pivotal in applications like healthcare and finance, where privacy regulations are stringent, and data integrity is critical.

4.5 Evaluation Metrics and Benchmarking

Evaluation metrics and benchmarking play a crucial role in assessing the effectiveness of adversarial attacks and the robustness of defense mechanisms in Federated Learning (FL). For measuring **attack effectiveness**, several quantitative metrics are widely employed. One primary metric is the **accuracy drop**, which quantifies the decrease in global model performance caused by an adversarial attack, serving as an indicator of the attack's overall impact on the model's utility (Bagdasaryan, Veit, Hua, Estrin, & Shmatikov, 2020). Another important metric is the **Attack Success Rate (ASR)**, especially critical in backdoor attacks, which measures the proportion of inputs on which the attacker successfully manipulates the model's predictions without detection (Gu, Dolan Gavitt, & Garg, 2019). Moreover, **clean label accuracy** assesses the model's performance on benign, non poisoned data, and is especially relevant for stealthy backdoor attacks that aim to preserve normal classification accuracy while embedding malicious behavior (Chen, Liu, Li, Lu, & Song, 2017).

On the defense side, metrics such as **robust accuracy** have become standard, representing the accuracy of the global model on clean test data despite adversarial attempts to degrade it (Blanchard, El Mhamdi, Guerraoui, & Stainer, 2017). Additionally, **false positive and false negative rates** in anomaly or attack detection systems provide insight into the precision and recall of defense strategies, highlighting their reliability and potential operational trade offs (Fang, Cao, & Gong, 2020). For instance, high false positive rates may lead to unnecessary exclusion of honest clients, adversely affecting model convergence, while high false negatives allow attacks to persist undetected.

Benchmarking in FL research relies heavily on **standardized datasets and frameworks** to ensure reproducibility and comparability across studies. Popular datasets include **LEAF (A Dataset for Federated Settings)**, which offers a suite of realistic FL benchmarks across various domains such as image recognition and natural language processing (Caldas et al., 2018). Other widely used datasets include **FEMNIST**, a federated extension of the MNIST dataset tailored for non IID data distributions, and **CIFAR 10**, which is frequently used in image classification tasks (McMahan et al., 2017). Frameworks such as **TensorFlow Federated (TFF)** and **PySyft** support experimentation with FL algorithms and adversarial scenarios, providing standardized tools for evaluation (Google, 2023; OpenMined, 2023).

Despite the progress, **fair benchmarking remains an open challenge** in the field. Differences in experimental setups, data heterogeneity, and varying assumptions about adversarial capabilities complicate direct comparisons of attack and defense effectiveness (Li, He, Song, & Li, 2021). Moreover, most benchmarks focus on specific attack types or threat models, lacking comprehensive coverage of real world, adaptive adversaries. As highlighted by Kairouz et al. (2021), establishing standardized, widely accepted benchmarks encompassing diverse attacks, defense methods, and practical deployment constraints is critical for advancing the field. Such benchmarks should include metrics beyond accuracy, incorporating communication costs, computational overhead, and energy consumption to reflect the realities of IoT and edge environments.

5.0 Real World Scenarios and Use Cases

Federated Learning (FL) is increasingly being deployed in real world applications across diverse sectors, from healthcare to autonomous vehicles, industrial IoT, and smart homes. Each of these domains presents unique security risks and challenges, highlighting the critical need to understand and mitigate adversarial threats effectively.

Security Risks in Healthcare FL

Healthcare represents one of the most promising yet vulnerable domains for FL, especially for collaborative cross hospital training where data privacy is paramount. FL enables multiple hospitals or medical institutions to jointly train models on sensitive patient data without sharing raw data, thereby enhancing diagnosis, prognosis, and treatment personalization (Sheller et al., 2020). However, this distributed setting introduces significant security risks. Adversarial attacks such as **data poisoning** or **backdoor insertion** can lead to misdiagnosis or biased outcomes by manipulating the training process at one or more institutions (Bagdasaryan et al., 2020). Moreover, **model inversion and membership inference attacks** pose severe privacy concerns as attackers may extract sensitive patient information from shared model updates (Melis, Song, De Cristofaro, & Shmatikov, 2019). For instance, the work by Li et al. (2020) demonstrated how backdoors in FL could cause healthcare models to incorrectly classify medical images, potentially endangering patient safety. Securing healthcare FL demands rigorous defense mechanisms that balance privacy preservation with robustness against adversarial behaviors.

Autonomous Vehicles and Real Time Poisoning Risks

Autonomous vehicles (AVs) rely heavily on real time data exchange and collaborative learning to improve perception, navigation, and decision making capabilities. FL is emerging as a powerful framework to enable vehicles to learn from decentralized sensor data without compromising privacy (Lu et al., 2020). However, AVs are highly susceptible to **real time poisoning attacks**, where malicious actors inject corrupted updates or trigger backdoors during the continuous training process. Such attacks could degrade the vehicle's object detection accuracy or cause erroneous control commands, leading to catastrophic accidents (Shen et al., 2021). The high mobility and dynamic nature of vehicular networks complicate the detection of adversarial updates and synchronization of learning rounds, increasing the attack surface (Zhao, Liu, & Song, 2022). Research by Hou et al. (2021) proposes defense aware client selection and anomaly detection tailored for AV environments to mitigate these risks. Nonetheless, real time responsiveness and safety critical requirements demand ongoing advancements in FL security for autonomous transportation.

Industrial IoT and Cyber Physical Attacks

The Industrial Internet of Things (IIoT) integrates sensors, actuators, and control systems within manufacturing plants, energy grids, and supply chains, often with stringent reliability and safety requirements. FL enables these distributed devices to collaboratively optimize operations, such as predictive maintenance and anomaly detection, while safeguarding sensitive industrial data (Lu et al., 2020). However, IIoT systems are prime targets for **cyber physical attacks**, where adversaries manipulate sensor data or inject malicious updates to disrupt physical processes (Weber et al., 2019). For example, a compromised FL client might cause false alarms or suppress fault detection, leading to equipment failures or safety hazards (Kang et al., 2020). Moreover, resource constraints and heterogeneity of IIoT devices pose challenges for deploying complex defense mechanisms (Zhang et al., 2021). Researchers emphasize combining lightweight anomaly detection with secure aggregation and trusted hardware to safeguard FL in industrial contexts (Ren et al., 2022). These multi layered approaches aim to maintain operational integrity while preserving data confidentiality.

Smart Home Devices and Model Leakage Threats

The proliferation of smart home devices, including voice assistants, security cameras, and IoT appliances, generates vast amounts of user data that can benefit from FL based collaborative learning to improve personalized services (Hard et al., 2018). Nonetheless, the security of FL in smart homes is challenged by risks such as **model leakage and inference attacks**, where adversaries exploit model updates to extract private user information (Hitaj, Ateniese, & Perez Cruz, 2017). Furthermore, smart home devices are often resource constrained and connected via unreliable networks, complicating the deployment of sophisticated security measures (Mohri, Sivek, & Suresh, 2019). Adversaries might also launch **sybil attacks**, creating multiple fake clients to influence model training and degrade overall system performance (Fung, Yoon, & Beschastnikh, 2021). Studies suggest combining differential privacy, secure aggregation, and robust client selection to mitigate these threats while maintaining service quality (Truex et al., 2019). Securing FL in smart homes remains a key research frontier as these devices become more pervasive and interconnected.

5.1 Open Challenges and Future Research Directions

Despite significant progress in securing Federated Learning (FL) systems, numerous open challenges remain that hinder the deployment of truly robust, scalable, and privacy preserving FL frameworks. Addressing these gaps is essential for enabling FL to realize its full potential, particularly in sensitive and resource constrained applications such as IoT, healthcare, and autonomous systems.

Designing Adaptive, Self-Healing Défense Systems

Traditional defense mechanisms in FL often assume static attack models or fixed threat scenarios. However, adversaries continuously evolve their strategies, employing adaptive and stealthy tactics to bypass existing protections (Sun et al., 2021). Therefore, designing **adaptive, self healing defense systems** capable of detecting and mitigating previously unseen attack vectors in real time remains a critical research frontier. Such systems should incorporate continuous monitoring, anomaly detection, and automated recovery techniques to ensure sustained robustness without human intervention (Li et al., 2022). Leveraging reinforcement learning and online learning approaches to dynamically tune defense parameters could offer promising solutions to this challenge (Nguyen et al., 2022).

Defending Against Coordinated or Stealthy Attacks

Coordinated attacks involving multiple malicious clients colluding to poison the global model pose severe threats to FL security. These **stealthy attacks** are often designed to evade anomaly detection by distributing malicious updates subtly across participants (Fang et al., 2020). Defending against such sophisticated adversaries requires advanced detection techniques that consider the collective behavior of clients over time rather than isolated update anomalies. Graph based trust evaluation, cross client consistency checks, and cryptographic proofs are emerging strategies that could improve resilience against coordinated adversaries (Xie et al., 2020). However, designing scalable and computationally efficient methods for large scale FL systems remains an open challenge.

Privacy Utility Security Trade Offs

Balancing **privacy, utility, and security** in FL is a delicate and ongoing research problem. Privacy preserving mechanisms such as differential privacy and secure multi party computation often degrade model

accuracy or increase communication and computational overhead (Truex et al., 2021). Conversely, stronger security defenses might conflict with privacy guarantees or limit model expressiveness. Exploring new **trade off paradigms** that optimize these conflicting goals without compromising practical feasibility is critical (Wei et al., 2022). Multi objective optimization frameworks and adaptive privacy budgets are promising directions to explore more flexible and context aware protections.

FL Under Constrained Device Capabilities (IoT Settings)

FL deployments on IoT devices face unique challenges due to limited computational power, battery life, and intermittent network connectivity (Zhao et al., 2020). Many existing defense mechanisms, such as robust aggregation and cryptographic protocols, incur significant overheads unsuitable for resource constrained environments. Research into **lightweight, energy efficient security methods** tailored for IoT based FL is urgently needed. This includes designing compressed model updates, partial training schemes, and hardware assisted trusted execution environments that balance security with device limitations (Wang et al., 2021).

Co Design with Blockchain, 6G, and Quantum Resilient Methods

Integrating FL with emerging technologies like **blockchain**, **6G networks**, and **quantum resistant cryptography** opens exciting avenues for enhancing security and trust. Blockchain can provide immutable audit trails and decentralized trust management to prevent malicious updates (Zheng et al., 2020). Meanwhile, 6G promises ultra low latency and massive connectivity, facilitating more secure and timely federated aggregation (Wang et al., 2022). However, these integrations require co designing protocols that align with FL's communication patterns and security requirements. Additionally, the looming threat of quantum computing necessitates developing **quantum resilient cryptographic methods** for FL to ensure long term security (Chen et al., 2021).

Certification and Verification of Defense Protocols

A fundamental challenge lies in the **certification and formal verification** of defense mechanisms deployed in FL. Given the complexity and heterogeneity of FL systems, guaranteeing that security protocols function correctly under all threat models is difficult. Developing **standardized benchmarking frameworks** and formal verification tools that rigorously evaluate the effectiveness and correctness of defense protocols is crucial for building trust and facilitating widespread adoption (Bagdasaryan & Shmatikov, 2020). Such efforts should also consider practical deployment constraints and real world adversarial behavior.

6.0 Conclusion

Federated Learning offers a promising approach to collaborative model training while preserving data privacy. However, its decentralized nature introduces numerous security vulnerabilities, including data and model poisoning, privacy attacks, and sophisticated adversarial behaviors. This survey reviewed the key adversarial threats and defense mechanisms, highlighting robust aggregation methods, anomaly detection, privacy preserving techniques, and trust-based systems. Despite progress, challenges remain in developing adaptive, efficient, and scalable defenses, especially for resource constrained IoT environments. Future research must focus on integrating FL with emerging technologies like blockchain and 6G, improving benchmarking standards, and designing self-healing systems. Securing FL is essential for its safe deployment across critical applications, and ongoing efforts will be key to achieving resilient and trustworthy federated systems.

Declaration of Conflicting Interests

The authors declare no potential conflicts of interest with respect to the research, authorship and publication of this article.

Funding

The author received no financial support for the research, authorship and publication of this article.

References

[1] Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2020). How to backdoor federated learning. Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, 2938–2948.
[2] Blanchard, P., El Mhamdi, E. M., Guerraoui, R., & Stainer, J. (2017). Machine learning with adversaries: Byzantine tolerant gradient descent. Advances in Neural Information Processing Systems, 31, 119–129.
[3] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., ... & Seth, K. (2017). Practical secure aggregation for federated learning on user held data. Advances in Neural Information Processing Systems, 30.
[4] Caldas, S., Wu, P., Li, T., Konečný, J., McMahan, H. B., Smith, V., & Talwalkar, A. (2018). LEAF: A benchmark for federated settings. arXiv preprint arXiv:1812.01097.
[5] Chen, M., Cheng, M., Liu, Y., & Xiong, N. (2021). Quantum secure federated learning: A survey. IEEE Access, 9, 135624–135642.
[6] Chen, X., Liu, C., Li, B., Lu, K., & Song, D. (2017). Targeted backdoor attacks on deep learning systems using data poisoning. arXiv preprint arXiv:1712.05526.
[7] Fang, M., Cao, X., & Gong, N. Z. (2020). Local model poisoning attacks to Byzantine robust federated learning. arXiv preprint arXiv:1805.09818.
[8] Fu, J., Hong, Y., Ling, X., Wang, L., Ran, X., Sun, Z., Wang, W. H., Chen, Z., & Cao, Y. (2024). Differentially Private Federated Learning: A Systematic Review. arXiv preprint arXiv:2405.08299.
[9] Fung, C., Yoon, C. J. M., & Beschastnikh, I. (2020). The limitations of federated learning in sybil settings. 23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID).
[10] Geyer, R. C., Klein, T., & Nabi, M. (2017). Differentially private federated learning: A client level perspective.
[11] Gu, T., Dolan Gavitt, B., & Garg, S. (2019). BadNets: Identifying vulnerabilities in the machine learning model supply chain.
[12] Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., ... & Ramage, D. (2018). Federated learning for mobile keyboard prediction.
[13] Hitaj, B., Ateniese, G., & Perez Cruz, F. (2017). Deep models under the GAN: Information leakage from collaborative deep learning. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 603–618.
[14] Hou, Z., Zhao, M., Shi, W., Xu, Y., Wang, Z., & Chen, X. (2021). Adaptive client selection for secure federated learning in vehicular networks. IEEE Transactions on Intelligent Transportation Systems, 22(9), 5713–5724.
[15] IBM Research. (2023). Federated Learning Meets Homomorphic Encryption. Retrieved from https://research.ibm.com/blog/federated learning homomorphic encryption
[16] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends® in Machine Learning, 14(1–2), 1–210.
[17] Kang, J., Xiong, Z., Niyato, D., Yu, R., & Zhang, D. (2020). Reliable federated learning for industrial IoT with power control and incentive mechanism. IEEE Internet of Things Journal, 7(7), 6480–6492.
[18] Li, Q., He, B., Song, D., & Li, H. (2021). Practical evaluation of backdoor attacks against federated learning. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 1813–1823.
[19] Li, T., He, X., Song, J., Guo, H., & Wang, X. (2022). Adaptive defense for federated learning: A reinforcement learning approach. IEEE Transactions on Neural Networks and Learning Systems, 33(8), 3910–3923.

[20] Li, X., Gu, Y., Li, K., Wang, Y., & Li, J. (2020). Backdoor attacks on deep learning based medical image analysis systems: A survey. arXiv preprint

[21] Lo, S. K., Liu, Y., Lu, Q., Wang, C., Xu, X., Paik, H. Y., & Zhu, L. (2021). Blockchain based trustworthy federated learning architecture. arXiv preprint arXiv:2108.06912.

[22] Lu, Y., Huang, X., Dai, Y., & Wang, Y. (2020). Edge cloud collaborative federated learning for industrial Internet of Things. IEEE Network, 34(5), 92–99.

[23] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 1273–1282.

[24] Melis, L., Song, C., De Cristofaro, E., & Shmatikov, V. (2019). Exploiting unintended feature leakage in collaborative learning. Proceedings of the 2019 IEEE Symposium on Security and Privacy, 691–706.

[25] Mohri, M., Sivek, G., & Suresh, A. T. (2019). Agnostic federated learning. Proceedings of the 36th International Conference on Machine Learning (ICML), 4615–4625.

[26] Nasr, M., Shokri, R., & Houmansadr, A. (2019). Comprehensive privacy analysis of deep learning: Passive and active white box inference attacks against centralized and federated learning. IEEE Symposium on Security and Privacy, 739–753.

[27] Nguyen, H., Tran, N., & Thai, M. (2022). Self adaptive defense mechanisms for federated learning with reinforcement learning. IEEE Internet of Things Journal, 9(10), 7229–7241.

[28] Phong, L. T., Aono, Y., Hayashi, T., Wang, L., & Moriai, S. (2018). Privacy preserving deep learning via additively homomorphic encryption. IEEE Transactions on Information Forensics and Security, 13(5), 1333–1345.

[29] Ren, Y., He, D., Huang, X., & Guizani, M. (2022). Robust federated learning for industrial IoT networks: Threats, solutions, and future directions. IEEE Communications Magazine, 60(1), 98–103.

[30] Sheller, M. J., Reina, G. A., Edwards, B., Martin, J., & Bakas, S. (2020). Multi institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, 92–104.

[31] Shen, X., Li, C., Zhou, Y., & Yu, Y. (2021). Real time federated learning against poisoning attacks in autonomous vehicles. IEEE Transactions on Vehicular Technology, 70(8), 7997–8009.

[32] Sun, Z., Xu, J., Lin, X., Zhang, R., & Liu, Q. (2022). Free riders in federated learning: Attacks and defenses. IEEE Transactions on Dependable and Secure Computing, 19(3), 1797–1812.

[33] Truex, S., Baracaldo, N., Anwar, A., Steinke, T., Ludwig, H., Zhang, R., & Zhou, Y. (2021). A hybrid approach to privacy preserving federated learning. Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security, 1–11.

[34] Wang, H., Yurochkin, M., Sun, Y., Papailiopoulos, D., & Khazaeni, Y. (2021). Federated learning with matched averaging. International Conference on Learning Representations.

[35] Wang, J., Liu, Q., Liang, H., Joshi, G., & Poor, H. V. (2021). A field guide to federated optimization. Proceedings of the IEEE, 109(5), 756–776.

[36] Wang, Y., Guo, S., Yu, R., & Leung, V. C. (2022). Federated learning with 6G: Opportunities and challenges. IEEE Network, 36(1), 30–36.

[37] Xie, C., Koyejo, S., & Gupta, I. (2019). Zeno: Distributed stochastic gradient descent with suspicion based fault tolerance. International Conference on Machine Learning (ICML), 6893–6901.

[38] Zhang, X., Hua, Y., & Qian, C. (2023). Secure decentralized learning with blockchain. arXiv preprint arXiv:2310.07079.

[39] Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2020). Federated learning with non IID data. arXiv preprint arXiv:1806.00582.

[40] Zheng, Z., Xie, S., Dai, H., Chen, X., & Wang, H. (2020). Blockchain challenges and opportunities: A survey. International Journal of Web and Grid Services, 14(4), 352–375.

[41] Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. Advances in Neural Information Processing Systems, 32.

[42] Zhu, X., Wu, S., Shi, Y., & Song, M. (2021). Sybil based poisoning attack against federated learning systems. Journal of Computer Security, 29(6), 879–906.