WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

The Formalisms of Privacy Leakage: Reviewing the Mathematical Foundations of Server-Side Inference Attack Strategies in Federated Learning

David Odera, Joshua Agolla & Loice Agong'

Jaramogi Oginga Odinga University of Science and Technology,210-40601, Bondo ,Kenya

ABSTRACT

Federated Learning (FL) preserve privacy by training models on decentralized devices and sharing only model updates instead of sharing the raw data to a central server. The focus is on learning data patterns in a heterogeneous environment to train a stable model capable of making precise predictions. Federated Learning (FL) applications spans numerous areas including autonomous self-driven cars, fraud detection, natural language processing and in most internet based organizations. Despite these privacy assurances, inference attacks present a viable threat by deciphering sensitive information from the shared model parameters. This investigation delineates the operational mechanisms of Membership Inference (MIA) and Property Inference Attacks (PIA) within FL, elucidating their mathematical foundations rooted in gradient analysis and statistical inference. A central thesis of this work is that the paramount threat emanates from a malicious central server, an adversary whose privileged access to all updates facilitates potent, large-scale privacy breaches. Through a synthesis of cutting-edge attacks and a critical evaluation of defenses, this work exposes a critical vulnerability in prevailing FL security paradigms. Our position negates strong assumptions which poses central server as honest and non-malicious. We conclude that future research must prioritize developing defenses effective against server-side adversaries to fulfill FL's promise of privacy-preserving collaborative learning.

Keywords: Membership Inferencing Attack, Property Inferencing, Generative Adversarial Attack, Bayesian Inferencing, Deep Leakage, Gradient Matching, Regularization, Homomorphic Encryption, Multiparty Computation, Source Inferencing

1. Introduction

To tackle security challenges in AI, distributed learning methods like federated learning are now widely used (Albshaier et al., 2025; Feng et al., 2025). This methodology entails multiple rounds of local model training on decentralized data subsets. The fundamental privacy mechanism is the transmission of only learned model updates (such as gradients or weights) to a central server, circumventing the need to exchange raw data (McMahan et al., 2016). This server then combines all the received updates to create a single, refined global model (McMahan et al., 2016). Given global model $\boldsymbol{\theta_0}$, with $\boldsymbol{n}$ number of clients, each containing a data $\boldsymbol{D_i}$ where $\boldsymbol{i}$ represent the index of a client. The updated global model is calculated as;

Equation 1 Federated Averaging

$$\boldsymbol{\theta_0 = \frac{1}{n}\sum_{i=1}^{n} \theta_i{}^{D}} \tag{1}$$

The global update $\boldsymbol{\theta_0}$ is based on a learning rate $\boldsymbol{\eta_g}$ which may also be considered as a step-size (Lin et al., 2025). If $\boldsymbol{\theta_0 = G}$, then global model is calculated using the expression, $\boldsymbol{G_t = G + \eta_g \theta_0}$, with $\boldsymbol{G_t}$ becoming a stable model after $\boldsymbol{t}$ rounds of training. Each client model is also trained in stochastic gradient descent (SGD) (Mills et al., 2023) and an updated local model is calculated using the formula $\boldsymbol{\theta_i{}^{r} = \theta_i - \eta_l \frac{\partial L}{\partial \theta_i{}^{D}}}$ where L is the loss objective function (Jadbabaie et al., 2023), $\boldsymbol{\eta_l}$ the local step-size and $\boldsymbol{r}$ is the number of epochs.

Even though baseline framework was designed for privacy preservation, recent studies suggests intentional violation of privacy through leakage leading to threats such as inference attacks (Bai, Hu, et al., 2025; Carlini et al., 2022; Hasan, 2023; Rao et al., 2025). Membership inferencing attack (MIA) and property inferencing attack (PIA) are two major categories of inferencing that adversaries may use to infer private information from participating devices in a distributed machine learning such as FL systems (Bai, Hu, et al., 2025; Wang et al., 2023). MIA is where an adversary may determine whether a specific record was used to train a model and reveal information about clients (Bai, Hu, et al., 2025). PIA tends to perform analysis on the states of local datasets and correlate it with possible information that is not present in training process (M. Li et al., 2025). Experimental studies (Bai, Hu, et al., 2025; Carlini et al., 2022; Rao et al., 2025; Wan et al., 2024) proved that baseline FL protocol may not withstand success of these attacks.

According to Das et al. (2025), the most notable state-of-art MIAs include WikiMIA (Shi et al., 2023), BookMIA (Y. Liu et al., 2022) and Temporal Wiki. Das et al. (2025) assessed the LAION-MI attack (Schuhmann et al., 2022) on a temporally partitioned arXiv dataset and performance was quantified using metrics including the Area Under the ROC Curve (AUC ROC) and the True Positive Rate at low False Positive Rates (such as TPR@1% FPR), emphasizing the attack's capability to identify members with high confidence under stringent privacy constraints. W. Wei et al. (2020), developed a framework to measure how gradient compression in federated learning impacts the effectiveness of client privacy leakage attacks. It also proposed preliminary mitigation strategies to underscore the need for a systematic evaluation method for understanding MI threats and building theoretical defenses against them.

In this work, we discuss the fundamental concepts of inference attacks, including attack mechanisms, theoretical groundings and their implications for the design of secure FL systems. A keen attention is on the attack formulations and vectors, differentiating between a malicious client attack and curious server. We elaborate the understanding of attack strategies in a schematic diagram and tables for clarity. Further this study examines contemporary defense techniques and their limitations with a view to persuade researchers in embracing multiple defense strategies going forward.

1.1 Motivation

The core motivation for federated learning (FL) is to leverage the patterns within distributed datasets while respecting data locality and privacy. This is crucial in domains like healthcare, finance, and personal devices, where data is inherently sensitive and governed by strict regulations. The initial design of FL operated on the assumption that sharing model parameters, rather than raw data, constituted a sufficient privacy safeguard. However, the discovery of inference attacks fundamentally challenges this assumption. The local updates shared to improve the global model serve as a side channel that can be exploited by a curious central server. This leads to a twofold motivation for this study: First, Nasr, Shokri, and Houmansadr (2019) opines that inference attacks are not merely theoretical but are practical and can achieve high accuracy. A successful MIA can reveal that an individual's record was used to train a model for a sensitive condition violating their privacy. A PIA, as described by Ganju et al. (2018), can deduce hidden attributes of an entire user group leading to discrimination and bias exploitation. Second, Zang et al. (2022) argue that many proposed solutions operate under the assumption of a trusted server and focus primarily on defending against attacks from other clients. This leaves the system highly vulnerable to a far more potent adversary (curious central server). The server has unobstructed access to all model updates and can aggregate them to launch powerful, large-scale inference attacks with minimal effort. Developing defenses that are effective against both client-side and server-side adversaries is therefore a critical and urgent challenge. This paper is motivated by the need to comprehensively analyze these threats to build a more realistic and robust foundation for privacy-preserving federated learning.

1.2 Organization

The remainder of this paper is structured as follows: *Section 2* contains related work, describing empirical studies on various kinds of inferencing attacks in federated learning. Section 3, examines inferencing attack, strategies, their mathematical foundations and formulations in federated learning. Section 4 provides summaries of attack techniques and limitations in table and a schematic diagram illustrating attack strategies in various components of federated learning. Finally, in section 5, the paper outlines a roadmap for subsequent investigation geared towards mitigating inferencing attacks in distributed machine learning systems (FL).

2. Related Work

Our work builds upon and synthesizes a growing body of literature focused on privacy vulnerabilities in Federated Learning. We explored the following as related work of this study. A comprehensive survey by Bai et al. (2025) examined Membership Inference Attacks (MIAs) in Federated Learning (FL), categorizing them into novel update-based and trend-based approaches that exploits the protocol's unique multi-round collaborative nature. Their work highlighted FL-specific attacks and how they differ from centralized learning by leveraging internal model details and historical data trajectories. It also discussed associated defenses to address these enhanced privacy risks. Lee et al. (2021) conducted empirical study using Digestive Neural Network (DNN) defense to protect against inferencing. According to experiment, there was significant improvement of accuracy upto 16.17% compared to differential privacy (DP) and a reduction of attack accuracy of 9% in gradient sharing. To measure MI attack efficacy, Carlini et al. (2022) proposed using the TPR at a low FPR (e.g., 0.1%). As reported by Das et al. (2025), following an empirical study, TPR which is similar to FPR meant the attack was no better than guessing local data points. However, a TPR that is ten times higher (e.g., 1%) showed the adversary could accurately identify a small fraction of members with high confidence (Das et al., 2025). Source Inference Attack (SIA) demonstrated a critical privacy risk beyond standard membership inference as described in (Hu et al., 2021). Lyu et al. (2020), revealed that Vertical Federated Learning (VFL) is vulnerable to novel label inference attacks, where a malicious participant infers private labels owned by another, even for data outside the training set. Xia et al. (2023), conducted a survey and provided a comprehensive classification and analysis of both poisoning attacks and their corresponding defense strategies in federated learning.

3. Inferencing Attack and Foundations

Inferencing of private data through training updates from local models is widely categorized by various studies (Bai, Hu, et al., 2025; M. Li et al., 2025; Lyu et al., 2020; Rao et al., 2025; Xia et al., 2023) into two major categories as described below

3.1. Membership Inference Attacks (MIA)

The foundational work by Shokri et al. (2017) demonstrated that ML models are vulnerable to MIAs due to their different behavior on training versus non-training data. The work of Nasr et al. (2019) categorized FL attacks into two methods: a non-disruptive approach using a shadow model for inference (passive) and an aggressive method that injects malicious parameters into the training process to force increased data exposure (active). With reference to MIA, the key distinction in the threat model is the role of the adversary (Nasr et al., 2019). Most early defenses assumed a malicious client, as noted by Zhang et al. (2022). However, a server-based MIA is arguably a more significant threat, as the server has direct access to every client's model update, making it trivial to query the global model and perform inference on any data point of interest (Bai, Hu, et al., 2025). Our analysis emphasizes that any robust FL system must be designed to mitigate this server-level threat.

3.2. Property Inference Attacks (PIA)

Whereas MIA targets specific data records, Property Inference Attacks (PIA) aim to deduce global properties of a client's dataset (Bai, Zhang, et al., 2025). Building on the foundational work of Ganju et al. (2018), adversaries may train a meta-model using synthetically generated data that mimics a target client's dataset, enabling the extraction of hidden properties by analyzing gradient updates. Shen et al. (2022) further demonstrated that even aggregated global models can inadvertently reveal statistical properties of client data, highlighting the persistent risk of information leakage in FL. These PIAs can deduce unlabeled sensitive attributes such as demographic or financial information solely from model training behavior (Yadav et al., 2023).

3.3. Mathematical Preliminaries and Formulation Strategies of Attacks

The following mathematical formulations may be used by an adversary to launch attack by inferencing data points from training distribution.

Equation 2 Probability Density Function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \qquad (2)$$

Where x is the data point within data distribution μ is mean and σ is standard deviation (Hollands & Huget, 1983; Ren et al., 2019).

Equation 3 Continuous Density Function (Yamato, 1971)

$$F(x) = P(a \leq x \leq b) = \int_a^b f(x)d(x) \qquad (3)$$

Equation 4 Bayes' Theorem (Schulman, 1984)

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A).P(B|A)}{P(B)} \qquad (4)$$

where P(A) and P(B) probabilities of A and B respectively, with P(A|B) being probability of A given B and P(B|A) is probability of B given A, while probability of both A and B occurring remains as $P(A \cap B)$

Equation 5 Neyman-Pearson Lemma Function (Ji & Zhou, 2010)

$$\lambda = \frac{L(\theta_0)}{L(\theta_1)} = \frac{P(x|H_0)}{P(x|H_1)} \leq k \qquad (5)$$

Equation 6 GAN value Function (Goodfellow et al., 2020)

$$V(G, D) = \text{℮}_{x \sim P_{data}}[\ln(D(x)] + \text{℮}_{z \sim P_z}[\ln(1 - D(G(z))] \qquad (6)$$

Equation 7 Gradient Leakage Function

$$(x^*, y^*) = arg \min_{x', y'} ||\nabla_\theta L(f_\theta(x'), y') - \nabla_\theta L(f_\theta(x), y)||||_2^2 \qquad (7)$$

$$(x'^*, y'^*) = arg \min_{x', y'} || \frac{\partial l(F(x', W), y')}{\partial W} - \nabla W||^2 \text{ where } \nabla W' = \frac{l(F(x', W), y')}{W} \qquad (8)$$

Equation 8 L2 Regularization (Cortes et al., 2012)

$$\min_\theta \, [\textstyle\sum_{z_i \in D_{train}} \ell(\theta: z_i) + \frac{\lambda}{2}||\theta||^2] \qquad (9)$$

Equation 9 Cosine Similarity (Ye, 2011)

$$sc_i = \cos\theta_i = \frac{\langle y_i, x_0 \rangle}{\|y_i\|.\|x_0\|} \qquad (10)$$

Equation 10 Gradient Matching Formula (W. Wei & Liu, 2022)

$$(x^*, y^*) = arg \min_{x', y'} ||\nabla_\theta L(f_\theta(x'), y') - \nabla_\theta L(f_\theta(x), y)||||_2^2 \qquad (11)$$

3.4 Membership Inferencing Using Probability Density Function and Bayes Theorem

Eq. 2 is known as Gaussian equation which is a foundational probabilistic framework that allows attacker to model differences in model behavior and construct optimal statistical tests to infer private information about the clients' training data from the FL updates (X. Yang & Wu, 2023). Gaussian-based ratio test provides a statistically rigorous method for launching an inference attack (Ma et al., 2024).

Given the described FL process in Eq. 1 a stable global model G_t is derived from aggregating clients' updates: $\theta_0 = \frac{1}{n}\sum_{i=1}^{n} \theta_i{}^D$, and each client perform local SGD given by $\theta_i{}^r = \theta_i - \eta_l \frac{\partial L}{\partial \theta_i{}^D}$, where L is the loss function, η_l is the local learning rate, and r is the number of local epochs. An attacker can leverage Bayesian reasoning based on Eq. 4 above to perform a Membership Inference Attack (MIA) against a target client.

The attacker's goal is to determine if a specific data point z = (x, y) was present in client $i's$ private dataset D_i, based on the information revealed from FL training, primarily the client's model update $\theta_i{}^D$. The optimal approach is to calculate the posterior probability using Bayes' Theorem in Eq. 4, the concept is also applied in (Campbell & Gustafson, 2023):

$$P(Member \mid \theta_i{}^D, z) \;=\; [\, f(\theta_i{}^D \mid z, Member) \;*\; P(Member)\,] \,/\, f(\theta_i{}^D \mid z)$$

Where

$P(Member \mid \theta_i{}^D, z)$ is *Posterior Probability*, defines the probability that z is a member of D_i, given the observed client update $\theta_i{}^D$

$f(\theta_i{}^D \mid z, Member)$ is *Likelihood*. This is the probability density of observing the specific model update $\theta_i{}^D$ *given that* the point z is a member of the client's dataset.

$P(Member)$ is the *Prior Probability* which explains attacker's initial belief about the probability that z is a member (given by 0.5).

$f(\theta_i{}^D \mid z)$ is *the Marginal Likelihood* which is the total probability of observing the update $\theta_i{}^D$ under all hypotheses (Member and Non-member). It used to normalize the equation.

The optimal decision rule supported by Eq. 5 implies the attacker will infer membership if the posterior probability for "Member" is greater than for "Non-member":

Infer "Member" if: $P(Member \mid \theta_i{}^D, z) > P(Non - Member \mid \theta_i{}^D, z)$

The malicious server rains multiple dummy client models and in every case it determines whether a specific point z was included in the training data or not. The server also compute the model update updates to estimate the two critical likelihood distributions ($f(\theta_i{}^D \mid z, Member)$ and $f\left(\theta_i{}^D \middle| z, Non - Member\right)$). To execute the attack, the server calculates the likelihood of this update under both PDFs.

3.5. Deep Leakage Gradients

The Deep Leakage from Gradients (DLG) (W. Wei & Liu, 2022) method minimizes the distance between the observed gradient and a dummy gradient. Eq. 10 involves optimizing a dummy data point and its label such that the gradient computed from this dummy pair (x', y') closely approximates the real gradient derived from the original data (x, y) (H. Yang et al., 2024). Employing a batch size of 8, P. Liu et al. (2022) evaluated that both the training image and its corresponding label can be successfully reconstructed. The formula presented above (Eq. 10) optimizes the similarity between the dummy gradients and the gradients of the original data, as follows:

$$(x'^*, y'^*) = arg \min_{x', y'} || \frac{\partial l(F(x', W), y')}{\partial W} - \nabla W ||^2$$

$$\text{where } \nabla W' = \frac{l(F(x', W), y')}{W}$$

Wainakh et al. (2021) leveraged inherent properties of gradients to uncover a significant relationship between the gradient and the model's output. Specifically, they demonstrated that gradients contain highly sensitive information about the training data, including features that can be extracted through careful analysis. Typically, these gradient properties reveal a direct correlation between the input data (and its label) and the resulting gradient. Their research showed that even without full data reconstruction, attributes like class representatives and feature correlations can be extracted from gradient updates (Wainakh et al., 2021). These exploitable properties are particularly pronounced in the final layer of a neural network, where gradients directly correspond to output predictions and contain the most discriminative information. Wainakh et al., (2021), described properties that were utilized to develop the attack methodology as gradient sparsity patterns that reveal feature importance, gradient magnitude distributions that correlate with class separation, and gradient direction similarities that expose data relationships within and across classes.

3.6. Source Inferencing

Source Inference Attacks (Hu et al., 2023) employ mathematical formulations to identify the specific client responsible for contributing a particular data point within a federated learning system. This attack methodology utilizes vector similarity measures (Elhussein & Gursoy, 2024) to trace data provenance through gradient analysis. Given the gradient of loss function $g = \nabla_i \ell(\theta, z)$ computed on the target data point **z** and model update or gradient vector

$\Delta\theta_i$ submitted by client $\boldsymbol{i}$. The attack employs cosine similarity in Eq. 9 as the core mathematical measure to quantify the alignment between the target data point's gradient and each client's model update:

$$sim(i) = (\Delta\theta_i \cdot g) / (||\Delta\theta_i|| \times ||g||)$$

This formulation computes the cosine of the angle between the two vectors (weights), providing a normalized measure of directional similarity that is invariant to vector magnitude. The optimization objective is then formulized as

$$i^* = \arg\max_i(i)$$

where the optimal client identification i^* is determined by finding the maximum value of the similarity function across all clients. The foundation of this attack rests on the mathematical principle that a client's model update represents the aggregation of gradients from all data points in its local dataset. The gradient of an individual data point **g** will demonstrate maximal directional alignment with the aggregate update $\Delta\theta_i$ from the client that originally contributed that data point, due to the additive nature of gradient computation.

This mathematical approach transforms the source identification problem into a measurable optimization task, leveraging geometric properties of vector spaces to infer data provenance within collaborative learning systems. This body of research delineates a key distinction where Membership Inference Attacks (MIAs) determine whether a specific sample was used during training, PIAs deduce underlying characteristics or labels within client datasets. Although countermeasures like Differential Privacy (DP) are commonly employed, they frequently necessitate a compromise between privacy protection and model performance. Our analysis synthesizes these insights and advocates for an updated threat model that prioritizes defense against sophisticated server-side adversaries a critical step toward designing robust, privacy-conscious federated learning frameworks.

4. Applications of Inferencing and Defense Strategies

This study summarizes areas where the attack techniques formulations have been applied using a table format. Specific gaps of every attack strategy is also defined in order to inform mitigation and secure strategies. This table catalogs how centralized server exploit the shared updates to infer sensitive information about the clients' private training data.

Table 1 Existing Implementations of Inferencing Attack Strategies

Author(s)	**Attack Strategy**	**Technique Name**	**Adversarial Component**	**Methodology**	**Limitation**
(B. Wang et al., 2019; Shokri et al., 2017b)	Membership Inference (MIA)	Passive MIA via Shadow Models	Server	Train shadow-model-based membership classifiers	Computationally expensive; requires accurate distribution matching
(Nasr et al., 2019)	MIA	Active MIA via Gradient Ascent	Server	Client correction reveals data membership	Highly intrusive, detectable, degrades global model performance
(Nasr et al., 2019)	MIA	KKT-based MIA	curious central server	Formulates MIA via integer programming using KKT conditions.	Limited to simpler models (e.g., logistic regression) where KKT conditions are tractable. Becomes computationally infeasible for large models like deep neural networks.
(Ganju et al., 2018; Melis et al., 2018)	Property Inference (PIA)	Gradient Signature Analysis	Server	The server analyzes client gradient statistics (mean, variance) to train an attack model that infers data membership	Attacker trains on synthetic data, which may not match real distribution, reducing accuracy

Author(s)	Attack Strategy	Technique Name	Adversarial Component	Methodology	Limitation
(Geiping et al., 2020; Huang et al., 2021; Zhu et al., 2019)	Source Inference	Cosine Similarity Attack	Server	Compare client update-to-target gradient cosine similarity to identify the source client	Larger client batch sizes dilute the target's gradient signal, reducing attack effectiveness
(Huang et al., 2021; Zhu et al., 2019)	Gradient Inversion	Deep Leakage from Gradients (DLG)	Server/Client	Optimize *dummy* data to match a client's gradient, reconstructing their private dataset	Only works effectively on very small batch sizes (often 1). Fails with larger batches, modern architectures (BatchNorm), and any form of gradient compression or noise (e.g., DP).
(Kabir et al., 2023)	IoT Volumetric MIA	Traffic Analysis MIA	Server	High classifier confidence in a device type suggests its traffic was in the training data, revealing its network presence	Only detects device type presence, not activity. Mitigated by lowering prediction confidence.

4.1 Schematic Diagram of Attack Inferencing Strategies

The study illustrates how inference attacks may be used by centralized server to infers data from clients by use using updates from clients. Loss optimization is managed through stochastic gradient descent in both client and server. In Figure 1, we intentionally located each inference attack strategy within a specific client to demonstrate how a local update can be computationally manipulated by the server to infer private information.

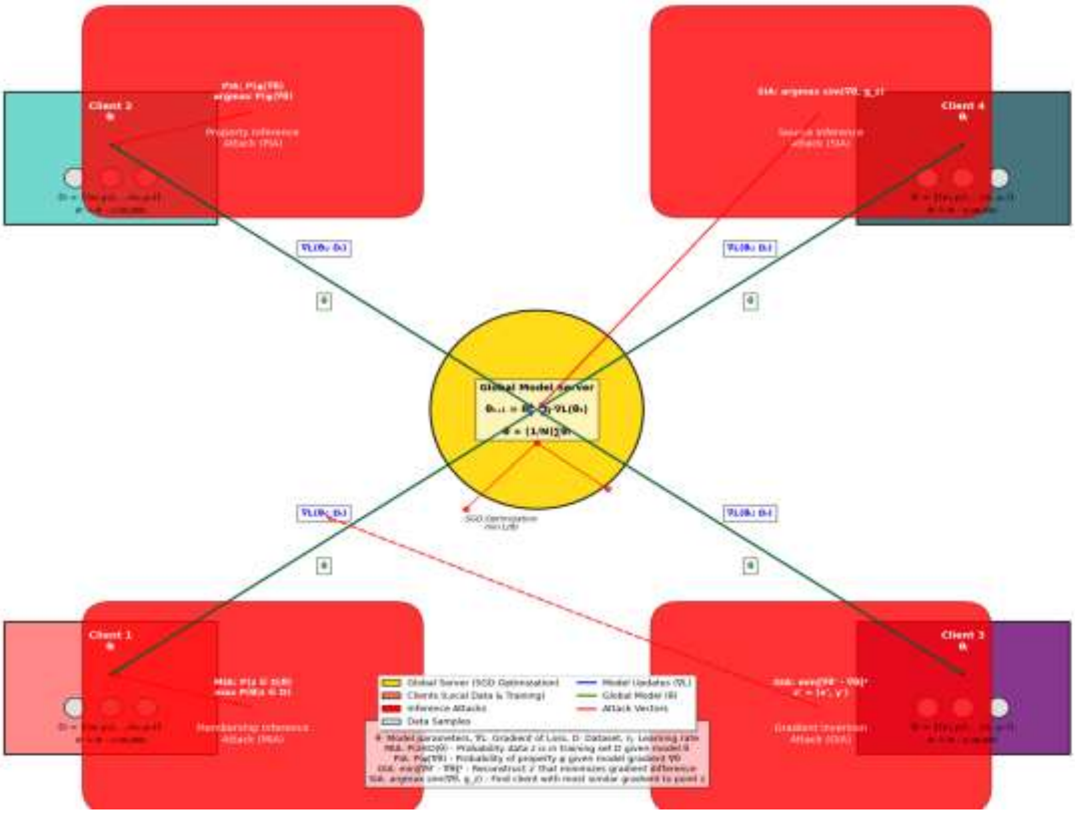

Client 2
Property Inference Attack (PIA)
Client 4
Source Inference Attack (SIA)
Global Model Server
SGD Optimization
Client 1
Membership Inference Attack (MIA)
Client 3
Gradient Inversion Attack (GIA)
Global Server (SGD Optimization)
Clients (Local Data & Training)
Inference Attacks
Data Samples
Model Updates (∇L)
Global Model (θ)
Attack Vectors

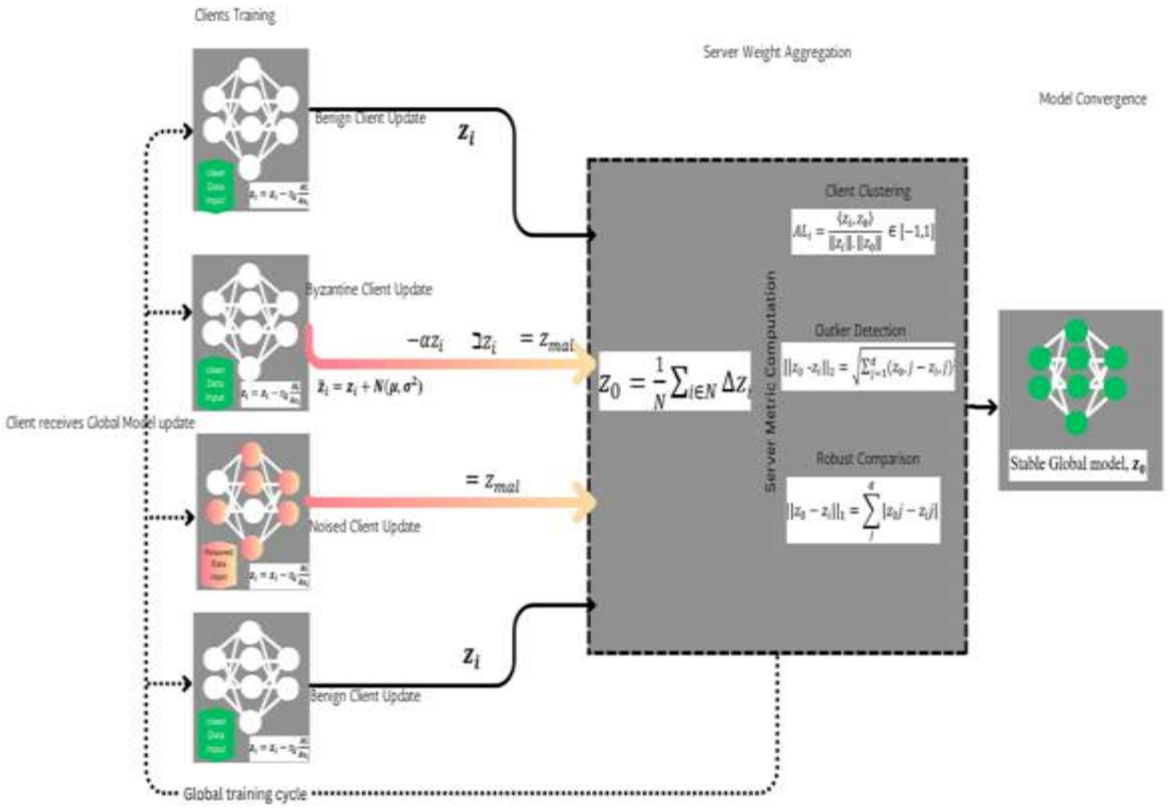
Figure 1 Schematic Diagram of Federated Learning with Inferencing Attacks

4.2 Existing Defence Techniques

The table below illustrates a taxonomy of foundational defence methods designed to neutralize privacy inference threats within the Federated Learning (FL) model. Every defence category contains a corresponding technique and existing vulnerability associated with exchange of model updates. These methods include homomorphic encryption (HE), differential privacy (DP), secure multi-party computation, anonymization and regularization.

Table 2 Defence Against Inferencing Attacks in FL

Proponents	**Defense Classification**	**Objective**	**Common Techniques**	**Trade-off**
(Cui et al., 2023; Z. Li & Zhang, 2020; McMahan et al., 2017; K. Wei et al., 2020)	*Differential Privacy (DP)*	Protect individual data points through perturbation	DP-SGD, Noise injection before aggregation.	Adding noise reduces model utility and performance accuracy. Requires careful tuning of the privacy budget (ε).
(Cheon et al., 2017; Phong et al., 2018; Q. Zhang et al., 2021)	*Homomorphic Encryption (HE)*	Perform computation on encrypted data.	Paillier encryption for secure aggregation. CKKS encryption	Have significant computational overhead and communication costs, making training slow and impractical for large models.
(Mohassel & Rindal, 2018; So et al., 2020; X. Zhang et al., 2023)	*Secure Multi-Party Computation (SMPC)*	Jointly compute a function with private inputs.	Secret sharing of model updates.	Introduces communication overhead between parties and complex coordination protocols.
(Chen et al., 2025; Wu et al., 2018; X. Zhang et al., 2023)	*Anonymization & Mixing*	Break the link between update and client.	Secure Aggregation, Update shuffling.	Secure Aggregation protocols can be broken if the server colludes with a subset of clients.
(Fereidooni et al., 2021; Pillutla et al., 2019; Shejwalkar & Houmansadr, 2021)	*Anomaly Detection*	filter out malicious or anomalous client updates.	Statistical filtering (e.g., median/trimmed mean), Clustering-based methods (e.g., K-Means), Norm-based thresholding, metric distance	Elimination of legitimate updates from clients with non-IID data, potentially biasing the global model
(Abbasi Tadi et al., 2023; Bai, Hu, et al., 2025; Salem et al., 2019)	*Regularization*	Reduce overfitting to lessen MIA leakage.	Increased L2 weight decay, Early stopping.	Introduce underfitting, limiting the final performance (accuracy) of the global model.


Clients Training
Server Weight Aggregation
Model Convergence
Benign Client Update
z_i
Client Clustering
$AL_i = \frac{\langle z_i, z_0 \rangle}{\|z_i\| . \|z_0\|} \in [-1,1]$
Byzantine Client Update
$-\alpha z_i$
$\exists z_i$
$= z_{mal}$
$\tilde{z}_i = z_i + N(\mu, \sigma^2)$
$Z_0 = \frac{1}{N} \sum_{i \in N} \Delta z_i$
Server Metric Computation
Outlier Detection
$\|z_0 - z_i\|_2 = \sqrt{\sum_{j=1}^{d} (z_{0,j} - z_{i,j})^2}$
Client receives Global Model update
Robust Comparison
$\|z_0 - z_i\|_1 = \sum_{j}^{d} |z_0 j - z_i j|$
Stable Global model, z_0
$= z_{mal}$
Noised Client Update
Client Data Input
Poisoned Data Input
z_i
Benign Client Update
Global training cycle

5. Conclusion and Future Work

This paper offers a systematic exploration of the technical and procedural mechanics behind server-side inference attacks in Federated Learning, asserting that the central server is the primary and most formidable malicious entity. The privileged role of the central server in Federated Learning enables it to be a powerful adversary, capable of orchestrating precise, large-scale privacy breaches through various inference attacks. This capability undermines FL's initial promise that sharing gradient updates alone provides sufficient privacy protection. A curious server can can therefore utilize mathematical toolkit to exploit privacy of individual devices in a network.

Techniques spanning theoretical frameworks like Bayesian inference and the Neyman-Pearson lemma to practical optimizations like gradient matching and cosine similarity equip a server with potent means to deduce sensitive information. Our synthesis of these attacks into a unified schematic (Figure 1) and taxonomic classification (Tables 1 & 2) clarifies the attack landscape and exposes a critical vulnerability in prevailing FL security paradigms.

This paper's evaluation demonstrates that contemporary defense mechanisms suffer shortcomings in dealing with inference attacks within federated learning. The study shows that techniques like Differential Privacy, Homomorphic Encryption, and Secure Multi-Party Computation incur prohibitive costs in utility or efficiency and are largely ineffective against a primary adversarial server. By negating the naive assumption of honest central server, this analysis concludes that realizing true privacy-preserving collaborative learning demands a paradigm shift towards defenses explicitly engineered to mitigate server-side adversaries.

Future research should pioneer cryptographic techniques, such as zero-knowledge proofs or verifiable secret sharing, which empower clients to cryptographically confirm that the server has correctly aggregated their updates without manipulation. This capability would directly neutralize the core threat posed by a malicious central server, ensuring the integrity of the aggregation process itself. There is need to theoretically tighten the privacy-utility trade-off through new algorithms that offer stronger formal guarantees against inference attacks without crippling model performance. This endeavor should involve creating adaptive differential privacy mechanisms with dynamic privacy budgets, developing hybrid models that combine defenses like lightweight secure multi-party computation with minimal noise injection, and establishing novel FL-specific privacy metrics and game-theoretic frameworks to move beyond ad-hoc evaluations.

Acknowledgements

We would like to appreciate Dr. Vincent Nyangaresi for his thought provoking critique and invaluable assessment of this work. Our gratitude also goes to the reviewers for their insightful suggestions, which significantly strengthened this work. We would also like to thank members of our department at Jaramogi Oginga Odinga University of Science and Technology for their valuable discussions and feedback on early drafts towards developing thesis.

Lastly, we are eternally grateful to our families for their patience, support, and encouragement during the long hours of research and writing. This work is dedicated to them.

References

Abbasi Tadi, A., Dayal, S., Alhadidi, D., & Mohammed, N. (2023). Comparative Analysis of Membership Inference Attacks in Federated and Centralized Learning. Information, 14(11), 620. https://doi.org/10.3390/info14110620

Albshaier, L., Almarri, S., & Albuali, A. (2025). Federated Learning for Cloud and Edge Security: A Systematic Review of Challenges and AI Opportunities. Electronics, 14(5), 1019. https://doi.org/10.3390/electronics14051019

Bai, L., Hu, H., Ye, Q., Li, H., Wang, L., & Xu, J. (2025). Membership Inference Attacks and Defenses in Federated Learning: A Survey. ACM Computing Surveys, 57(4), 1–35. https://doi.org/10.1145/3704633

Bai, L., Zhang, X., Zhang, S., Ye, Q., & Hu, H. (2025). ProVFL: Property Inference Attacks Against Vertical Federated Learning. IEEE Transactions on Information Forensics and Security, 20, 6529–6543. https://doi.org/10.1109/TIFS.2025.3581743

Campbell, H., & Gustafson, P. (2023). Bayes Factors and Posterior Estimation: Two Sides of the Very Same Coin. The American Statistician, 77(3), 248–258. https://doi.org/10.1080/00031305.2022.2139293

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., & Tramèr, F. (2022). Membership Inference Attacks From First Principles. 2022 IEEE Symposium on Security and Privacy (SP), 1897–1914. https://doi.org/10.1109/SP46214.2022.9833649

Chen, R., Dong, Y., Liu, Y., Fan, T., Li, D., Guan, Z., Liu, J., & Zhou, J. (2025). FLock: Robust and Privacy-Preserving Federated Learning based on Practical Blockchain State Channels. Proceedings of the ACM on Web Conference 2025, 884–895. https://doi.org/10.1145/3696410.3714666

Cheon, J. H., Kim, A., Kim, M., & Song, Y. (2017). Homomorphic Encryption for Arithmetic of Approximate Numbers. In T. Takagi & T. Peyrin (Eds.), Advances in Cryptology – ASIACRYPT 2017 (Vol. 10624, pp. 409–437). Springer International Publishing. https://doi.org/10.1007/978-3-319-70694-8_15

Cortes, C., Mohri, M., & Rostamizadeh, A. (2012). L2 Regularization for Learning Kernels (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1205.2653

Cui, Y., Meerza, S. I. A., Li, Z., Liu, L., Zhang, J., & Liu, J. (2023). RecUP-FL: Reconciling Utility and Privacy in Federated Learning via User-configurable Privacy Defense. https://doi.org/10.48550/ARXIV.2304.05135

Das, D., Zhang, J., & Tramèr, F. (2025). Blind Baselines Beat Membership Inference Attacks for Foundation Models. 2025 IEEE Security and Privacy Workshops (SPW), 118–125. https://doi.org/10.1109/SPW67851.2025.00016

Elhussein, A., & Gursoy, G. (2024). A Universal Metric of Dataset Similarity for Cross-silo Federated Learning (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2404.18773

Feng, Y., Guo, Y., Hou, Y., Wu, Y., Lao, M., Yu, T., & Liu, G. (2025). A survey of security threats in federated learning. Complex & Intelligent Systems, 11(2), 165. https://doi.org/10.1007/s40747-024-01664-0

Fereidooni, H., Marchal, S., Miettinen, M., Mirhoseini, A., Mollering, H., Nguyen, T. D., Rieger, P., Sadeghi, A.-R., Schneider, T., Yalame, H., & Zeitouni, S. (2021). SAFELearn: Secure Aggregation for private FEderated Learning. 2021 IEEE Security and Privacy Workshops (SPW), 56–62. https://doi.org/10.1109/SPW53761.2021.00017

Ganju, K., Wang, Q., Yang, W., Gunter, C. A., & Borisov, N. (2018). Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations. Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, 619–633. https://doi.org/10.1145/3243734.3243834

Geiping, J., Bauermeister, H., Dröge, H., & Moeller, M. (2020). Inverting Gradients—How easy is it to break privacy in federated learning? (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2003.14053

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. Communications of the ACM, 63(11), 139–144. https://doi.org/10.1145/3422622

Hasan, J. (2023). Security and Privacy Issues of Federated Learning (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2307.12181

Hollands, K. G. T., & Huget, R. G. (1983). A probability density function for the clearness index, with applications. Solar Energy, 30(3), 195–209. https://doi.org/10.1016/0038-092X(83)90149-4

Hu, H., Salcic, Z., Sun, L., Dobbie, G., & Zhang, X. (2021). Source Inference Attacks in Federated Learning (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2109.05659

Hu, H., Zhang, X., Salcic, Z., Sun, L., Choo, K.-K. R., & Dobbie, G. (2023). Source Inference Attacks: Beyond Membership Inference Attacks in Federated Learning (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2310.00222

Huang, Y., Gupta, S., Song, Z., Li, K., & Arora, S. (2021). Evaluating Gradient Inversion Attacks and Defenses in Federated Learning. https://doi.org/10.48550/ARXIV.2112.00059

Jadbabaie, A., Makur, A., & Shah, D. (2023). Federated Optimization of Smooth Loss Functions. IEEE Transactions on Information Theory, 69(12), 7836–7866. https://doi.org/10.1109/TIT.2023.3317168

Ji, S., & Zhou, X. Y. (2010). A generalized Neyman–Pearson lemma for g-probabilities. Probability Theory and Related Fields, 148(3–4), 645–669. https://doi.org/10.1007/s00440-009-0244-4

Kabir, E., Song, Z., Rashid, M. R. U., & Mehnaz, S. (2023). FLShield: A Validation Based Federated Learning Framework to Defend Against Poisoning Attacks (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2308.05832

Lee, H., Kim, J., Ahn, S., Hussain, R., Cho, S., & Son, J. (2021). Digestive neural networks: A novel defense strategy against inference attacks in federated learning. Computers & Security, 109, 102378. https://doi.org/10.1016/j.cose.2021.102378

Li, M., Gjoreski, M., Barbiero, P., Slapničar, G., Luštrek, M., Lane, N. D., & Langheinrich, M. (2025). A Survey on Federated Learning in Human Sensing (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2501.04000

Li, Z., & Zhang, Y. (2020). Membership Leakage in Label-Only Exposures (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2007.15528

Lin, Y., Li, W., Song, J., & Li, X. (2025). Online Federated Reproduced Gradient Descent With Time-Varying Global Optima. IEEE Transactions on Signal Processing, 73, 1379–1393. https://doi.org/10.1109/TSP.2025.3549591

Liu, P., Xu, X., & Wang, W. (2022). Threats, attacks and defenses to federated learning: Issues, taxonomy and perspectives. Cybersecurity, 5(1), 4. https://doi.org/10.1186/s42400-021-00105-6

Liu, Y., Zhao, Z., Backes, M., & Zhang, Y. (2022). Membership Inference Attacks by Exploiting Loss Trajectory (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2208.14933

Lyu, L., Yu, H., & Yang, Q. (2020). Threats to Federated Learning: A Survey (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2003.02133

Ma, J., Zhou, Y., Li, Q., Sheng, Q. Z., Cui, L., & Liu, J. (2024). The Power of Bias: Optimizing Client Selection in Federated Learning with Heterogeneous Differential Privacy (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2408.08642

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. y. (2016). Communication-Efficient Learning of Deep Networks from Decentralized Data. https://doi.org/10.48550/ARXIV.1602.05629

McMahan, H. B., Ramage, D., Talwar, K., & Zhang, L. (2017). Learning Differentially Private Recurrent Language Models (Version 3). arXiv. https://doi.org/10.48550/ARXIV.1710.06963

Melis, L., Song, C., De Cristofaro, E., & Shmatikov, V. (2018). Exploiting Unintended Feature Leakage in Collaborative Learning (Version 3). arXiv. https://doi.org/10.48550/ARXIV.1805.04049

Mills, J., Hu, J., & Min, G. (2023). Faster Federated Learning With Decaying Number of Local SGD Steps. IEEE Transactions on Parallel and Distributed Systems, 34(7), 2198–2207. https://doi.org/10.1109/TPDS.2023.3277367

Mohassel, P., & Rindal, P. (2018). ABY3: A Mixed Protocol Framework for Machine Learning. Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, 35–52. https://doi.org/10.1145/3243734.3243760

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. 2019 IEEE Symposium on Security and Privacy (SP), 739–753. https://doi.org/10.1109/SP.2019.00065

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. https://doi.org/10.1109/TIFS.2017.2787987

Pillutla, K., Kakade, S. M., & Harchaoui, Z. (2019). Robust Aggregation for Federated Learning. https://doi.org/10.48550/ARXIV.1912.13445

Rao, B., Zhang, J., Wu, D., Zhu, C., Sun, X., & Chen, B. (2025). Privacy Inference Attack and Defense in Centralized and Federated Learning: A Comprehensive Survey. IEEE Transactions on Artificial Intelligence, 6(2), 333–353. https://doi.org/10.1109/TAI.2024.3363670

Ren, M., Zhang, Q., & Zhang, J. (2019). An introductory survey of probability density function control. Systems Science & Control Engineering, 7(1), 158–170. https://doi.org/10.1080/21642583.2019.1588804

Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., & Backes, M. (2019). ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. Proceedings 2019 Network and Distributed System Security Symposium. Network and Distributed System Security Symposium, San Diego, CA. https://doi.org/10.14722/ndss.2019.23119

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., & Jitsev, J. (2022). LAION-5B: An open large-scale dataset for training next generation image-text models (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2210.08402

Schulman, P. (1984). Bayes' Theorem—A Review. Cardiology Clinics, 2(3), 319–328. https://doi.org/10.1016/S0733-8651(18)30726-4

Shejwalkar, V., & Houmansadr, A. (2021). Manipulating the Byzantine: Optimizing Model Poisoning Attacks and Defenses for Federated Learning. Proceedings 2021 Network and Distributed System Security Symposium. Network and Distributed System Security Symposium, Virtual. https://doi.org/10.14722/ndss.2021.24498

Shen, J., Tian, J., Wang, Z., & Cai, H. (2022). Friendship links-based privacy-preserving algorithm against inference attacks. Journal of King Saud University - Computer and Information Sciences, 34(10), 9363–9375. https://doi.org/10.1016/j.jksuci.2022.09.014

Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., & Zettlemoyer, L. (2023). Detecting Pretraining Data from Large Language Models (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2310.16789

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017a). Membership Inference Attacks Against Machine Learning Models. 2017 IEEE Symposium on Security and Privacy (SP), 3–18. https://doi.org/10.1109/SP.2017.41

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017b). Membership Inference Attacks Against Machine Learning Models. 2017 IEEE Symposium on Security and Privacy (SP), 3–18. https://doi.org/10.1109/SP.2017.41

So, J., Guler, B., & Avestimehr, A. S. (2020). Turbo-Aggregate: Breaking the Quadratic Aggregation Barrier in Secure Federated Learning (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2002.04156

Wainakh, A., Ventola, F., Müßig, T., Keim, J., Cordero, C. G., Zimmer, E., Grube, T., Kersting, K., & Mühlhäuser, M. (2021). User-Level Label Leakage from Gradients in Federated Learning (Version 4). arXiv. https://doi.org/10.48550/ARXIV.2105.09369

Wan, Y., Qu, Y., Ni, W., Xiang, Y., Gao, L., & Hossain, E. (2024). Data and Model Poisoning Backdoor Attacks on Wireless Federated Learning, and the Defense Mechanisms: A Comprehensive Survey. IEEE Communications Surveys & Tutorials, 26(3), 1861–1897. https://doi.org/10.1109/comst.2024.3361451

Wang, Z., Huang, Y., Song, M., Wu, L., Xue, F., & Ren, K. (2023). Poisoning-Assisted Property Inference Attack Against Federated Learning. IEEE Transactions on Dependable and Secure Computing, 20(4), 3328–3340. https://doi.org/10.1109/TDSC.2022.3196646

Wei, K., Li, J., Ding, M., Ma, C., Yang, H. H., Farokhi, F., Jin, S., Quek, T. Q. S., & Vincent Poor, H. (2020). Federated Learning With Differential Privacy: Algorithms and Performance Analysis. IEEE Transactions on Information Forensics and Security, 15, 3454–3469. https://doi.org/10.1109/TIFS.2020.2988575

Wei, W., & Liu, L. (2022). Gradient Leakage Attack Resilient Deep Learning. IEEE Transactions on Information Forensics and Security, 17, 303–316. https://doi.org/10.1109/TIFS.2021.3139777

Wei, W., Liu, L., Loper, M., Chow, K.-H., Gursoy, M. E., Truex, S., & Wu, Y. (2020). A Framework for Evaluating Gradient Leakage Attacks in Federated Learning (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2004.10397

Wu, L., Du, X., Wu, J., Liu, J., & Dragut, E. C. (2018). An Accountable Anonymous Data Aggregation Scheme for Internet of Things (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1803.07760

Xia, G., Chen, J., Yu, C., & Ma, J. (2023). Poisoning Attacks in Federated Learning: A Survey. IEEE Access, 11, 10708–10722. https://doi.org/10.1109/ACCESS.2023.3238823

Yadav, P., Gupta, N., & Sharma, P. K. (2023). A comprehensive study towards high-level approaches for weapon detection using classical machine learning and deep learning methods. Expert Systems with Applications, 212, 118698. https://doi.org/10.1016/j.eswa.2022.118698

Yamato, H. (1971). SEQUENTIAL ESTIMATION OF A CONTINUOUS PROBABILITY DENSITY FUNCTION AND MODE. Bulletin of Mathematical Statistics, 14(3/4), 1–12. https://doi.org/10.5109/13049

Yang, H., Ge, M., Xue, D., Xiang, K., Li, H., & Lu, R. (2024). Gradient Leakage Attacks in Federated Learning: Research Frontiers, Taxonomy, and Future Directions. IEEE Network, 38(2), 247–254. https://doi.org/10.1109/MNET.001.2300140

Yang, X., & Wu, W. (2023). A federated learning differential privacy algorithm for non-Gaussian heterogeneous data. Scientific Reports, 13(1), 5819. https://doi.org/10.1038/s41598-023-33044-y

Ye, J. (2011). Cosine similarity measures for intuitionistic fuzzy sets and their applications. Mathematical and Computer Modelling, 53(1–2), 91–97. https://doi.org/10.1016/j.mcm.2010.07.022

Zhang, Q., Xin, C., & Wu, H. (2021). GALA: Greedy ComputAtion for Linear Algebra in Privacy-Preserved Neural Networks (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2105.01827

Zhang, X., Kang, Y., Chen, K., Fan, L., & Yang, Q. (2023). Trading Off Privacy, Utility, and Efficiency in Federated Learning. ACM Transactions on Intelligent Systems and Technology, 14(6), 1–32. https://doi.org/10.1145/3595185

Zhang, Z., Yan, C., & Malin, B. A. (2022). Membership inference attacks against synthetic health data. Journal of Biomedical Informatics, 125, 103977. https://doi.org/10.1016/j.jbi.2021.103977

Zhu, L., Liu, Z., & Han, S. (2019). Deep Leakage from Gradients (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1906.08935