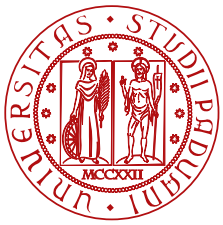

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INFORMATION ENGINEERING

Privacy and security in distributed learning: common threats and countermeasures

Relatore
Prof. Giovanni Perin

Laureando
Alaa Habib

ANNO ACCADEMICO 2024 - 2025

Data di laurea 22/07/2025

To my family thank you for your unconditional love,endless patience, and the strength you have given me every step of the way. Your sacrifices, encouragement, and belief in my potential have been the foundation on which this entire journey was built.I wouldn't ask for a better support system I consider myself very blessed to have you by myside even though the distance is faraway but you never made me feel that I am alone . Your presence was always felt every second . Thank you to my lovely Parents , to my brother and my sisters and my Grandmother who was always supporting me and blessing me with motivational words all the time. And to Prof.Giovanni Perin, thank you for your generous guidance, your steady belief in my work, and your invaluable insight throughout this thesis. Your mentorship has shaped not only the direction of this research, but also my growth as a thinker. To my friends which are all around the world, your support, humor, and presence have been a light during the most challenging moments. Whether through late night talks, thoughtful messages, or quiet understanding, you have made this experience not just bearable, but meaningful. And special thanks to my bestfriends who truly made this journy easier for me : Zayna,Noura,Sarah,Judy,Sara,Hajar you guys are the best. This work is the result of many voices, hearts, and hands. I express deep gratitude to each of you who stood beside me. One last thing I would like to mention , While I had the privilege of receving my Bachelors degree, there are girls and boys just like me in Palestine who cannot. Not because they didn't dream, study, or strive, but because their schools were bombed, their homes were destroyed , their future stolen. Because genocide and injustice have made this milestone unreachable. Today I graduate not just for myself, but in honor of those who never got the chance. Education should never be a casualty. Childhood should never be a target. To the students in Palestine: I mourn for you. I carry your hopes with me.

Abstract

The rapid expansion of federated learning (FL) has introduced new challenges in the domain of machine learning security and privacy. While FL is designed to preserve user data privacy by keeping data decentralized, it remains vulnerable to several types of adversarial attacks. This thesis focuses on two prominent threats: data poisoning attacks, which aim to corrupt the training process by injecting malicious data, and membership inference attacks, which seek to determine whether a specific data point was used during training. A comprehensive analysis of these attack methods is provided, followed by a critical evaluation of the most widely adopted defensive strategies. Based on the limitations observed in existing approaches, a hybrid defense framework is proposed, aiming to balance security, privacy, and model performance. The work concludes by highlighting key open challenges and ethical considerations for the development of secure and trustworthy federated learning systems.

Sommario

La rapida diffusione del Federated Learning (FL) ha introdotto nuove sfide nell'ambito della sicurezza e della privacy nel machine learning. Sebbene il FL sia concepito per tutelare la riservatezza dei dati mantenendoli decentralizzati, esso rimane vulnerabile a diversi tipi di attacchi avversari. Questa tesi si concentra su due minacce principali: gli attacchi di avvelenamento dei dati, che mirano a compromettere il processo di addestramento tramite l'iniezione di dati malevoli, e gli attacchi di inferenza di appartenenza, che cercano di determinare se un dato specifico sia stato utilizzato durante l'addestramento. Viene fornita un'analisi approfondita di queste tecniche di attacco, seguita da una valutazione critica delle strategie difensive più diffuse. Sulla base delle limitazioni rilevate, viene proposto un framework di difesa ibrido che mira a bilanciare sicurezza, privacy e prestazioni del modello. Il lavoro si conclude evidenziando le principali sfide aperte e le considerazioni etiche nello sviluppo di sistemi FL sicuri e affidabili.

Indice

Capitolo 1

Introduction and Research Goals

1.1 Background and Context

In recent years, the increasing need for information privacy has caused substantial transformations within machine learning model training. In centralized learning, information from many users or sources is merged on one server, where the model is implemented. Effective as it is, the method is highly privacy and law risky, particularly in the healthcare and finance domains, where information exposure can have dire consequences (Shokri et al., 2017; Hitaj et al., 2017). Distributed learning is a response wherein machine learning models are trained on decentralized nodes, be it edge devices or institutions, without raw data leaving their location. Distributed learning includes a popular and widely used method known as Federated Learning (FL). Local devices train through their individual data while sharing model updates with a central server for aggregation purposes only (McMahan et al., 2017; Kairouz et al., 2021). Using this method, model development is made possible while retaining the locality of and ownership of the data, thus providing a potentially viable solution for privacy-sensitive applications. But distributed learning is not problem-free. The same qualities that make privacy conducive, that is, decentralization and restricted control, also introduce new risks. Attackers have opportunities for exploitation based on the openness of the system, heterogeneity of the nodes involved, and difficulty of authenticating the credibility of inputs (Bagdasaryan et al., 2020; Kairouz et al., 2019). Following this, providing security and privacy for distributed learning is

an urgent research concern.

1.2 Motivation

The motivation for the thesis stems from the increasing real-world adoption of federated and distributed learning systems and the lack of mature security mechanisms to defend them. FL is being increasingly applied across industries such as mobile health, intelligent transportation systems, and financial analytics because of its ability to comply with regulatory regimes such as GDPR. However, recent evidence indicates that such systems can be attacked (Bagdasaryan et al., 2020; Nasr et al., 2019). In fact, they are even more vulnerable as a consequence of the inability to monitor or control individual participants (Hitaj et al., 2017). Several different parts of distributed systems remain susceptible to attacks because of their extensive attack surface. Attackers may damage model performance while simultaneously breaking user privacy, which undermines the whole reason distributed learning was adopted. Data Poisoning Attacks and Membership Inference Attacks stand out as the most prevalent attacks examined within the research literature due to their practicality and threat severity (Nasr et al., 2019; Bagdasaryan et al., 2020). Building appropriate protective measures requires more than a superficial understanding of internal learning processes because these threats present specific substantial risks.

1.3 Focus of the Thesis

This thesis focuses on two key attacks:

- **Data Poisoning Attacks:** These consist of injecting malicious or deceptive information into the training process by one or several malicious clients. The objective can range from model degradation (availability attacks) to intentional misclassification (backdoor attacks). Due to training on edge devices where there is minimal verification, it becomes possible for at-

tackers to poison the learning process undetected (Bhagoji et al., 2019; Bagdasaryan et al., 2020; Fang et al., 2020).

- **Membership Inference Attacks:** These involve identifying whether a specific record of data was part of the model's training set. This can cause serious privacy violations, most notably where being included within a dataset can, by itself, constitute sensitive information (such as medical notes from a hospital). In such cases, membership inference attacks are based on the observation that models have different behaviour on training as opposed to non-training instances (Shokri et al., 2017).

  Both exploits have had considerable research attention; however, actual defensive strategies are still unavailable. They are studied so that one can know how to make resilient distributed learning systems and how to secure users from direct as well as indirect attacks.

1.4 Research Goals and Objectives

The primary objective of this thesis is the exploration of the nature, feasibility, and effect of Data Poisoning Attacks and Membership Inference Attacks on distributed learning systems. The thesis aims to:

- Describe the mechanisms and techniques involved in such attacks within the context of federated learning.
- Survey and group together the best defence strategies outlined in recent works.
- Evaluate the strengths as well as the drawback of these defense mechanisms.
- Identify the open issues and directions of research on securing distributed learning systems.

Through its attainment of its objectives, this thesis will help further enhance our comprehension of distributed learning vulnerabilities as well as provide insight for future privacy-preserving machine learning developments.

Capitolo 2

Research Questions and Objectives

2.1 Research Problem

Federated Learning (FL) enables distributed learning, which combines local data storage with model updates shared from a central server to preserve privacy throughout operations. New security challenges arise from distributed learning frameworks, which involve decentralized data storage, as these models specifically threaten decentralized computation (Bhagoji et al., 2019; Bagdasaryan et al., 2020; Fang et al., 2020). Data poisoning attacks, alongside membership inference attacks, are among the most critical FL security threats. Malicious clients implement altered data inputs to disrupt the training process, while adversaries aim to identify which training set members provided specific data points (Shokri et al., 2017; Salem et al., 2019). The core values of FL encounter significant challenges because attackers compromise model fidelity and breach privacy boundaries. The available set of defense mechanisms from recent literature shows no sign of achieving a universally efficient and scalable solution.

- **Data Poisoning Attacks** – where malicious clients inject corrupted data to mislead the model.
- **Membership Inference Attacks (MIAs)** – where adversaries aim to determine whether a particular data point was used in training.

Despite recent advances, no defense provides a universally efficient or scalable solution.

2.2 Main Research Question

What are the most effective countermeasures to mitigate common threats, specifically data poisoning and membership inference attacks in distributed learning systems, while preserving model performance and user privacy?

2.3 Research Objectives

This thesis seeks to:

1. Explore the attack vectors and techniques involved in data poisoning within distributed systems (Fang et al., 2020).
2. Analyze the methodology and feasibility of membership inference attacks in federated learning (Shokri et al., 2017; Salem et al., 2019).
3. Evaluate the effectiveness of current defenses, such as differential privacy, robust aggregation, and cryptographic techniques (Bhagoji et al., 2019; Bonawitz et al., 2017).
4. Assess trade-offs between privacy, accuracy, and computational cost introduced by these defenses (Truex et al., 2019).
5. Identify research gaps and open challenges to suggest future directions for more resilient FL systems.

2.4 Sub-Questions

To reach these objectives, the thesis will address the following sub-questions:

- What are the goals, methods, and effectiveness of poisoning attacks in real federated learning environments?
- How do model characteristics(e.g., overfitting, confidence scores) influence vulnerability to membership inference attacks?

- Which defense techniques(e.g., Trimmed Mean, Krum, DP-SGD) are most promising under practical constraints?
- What are the performance implications of integrating these defenses into real-world FL pipelines?
- What are the current limitations of the literature and what future directions could improve the security of distributed learning?

Capitolo 3

Data Poisoning Attacks in Distributed Learning

3.1 Introduction

Distributed machine learning faces data poisoning attacks as one of its most substantial and practical threats to its operations. Attackers disturb the global model outcome by injecting malicious or incorrect information into their individual data pools that feed global model training (Fang et al., 2020). Federated learning delivers model updates instead of client data, so its decentralized setup makes detecting poisoned information very difficult (Bhagoji et al., 2019). The attack method of data poisoning hurts model integrity while also breaking user trust because it distorts decision boundaries in critical applications, including self-driving systems, healthcare systems, and financial operations (Bagdasaryan et al., 2020).

3.2 Types of Data Poisoning Attacks

Poisoning attacks can be broadly categorized into the following types (Bhagoji et al., 2019; Fang et al., 2020):

- Untargeted Attacks:The attacker's goal is to degrade the overall model accuracy or convergence. These attacks aim to make the global model less useful for all users.

- Targeted Attacks: The attacker manipulates data to ensure specific misclassifications, such as making the model misclassify images of one class as another.
- Backdoor Attacks: :A subcategory of targeted attacks, where the attacker implants a specific trigger (e.g., a pattern in an image) so that any input with this trigger will be misclassified in a desired way (Tolpegin et al., 2020).
- Sybil Attacks: The attacker manages multiple fake or compromised client devices that work together in the learning procedure. The perpetrator achieves more potent poisoning effects through coordinated malicious updates while circumventing defenses that depend on single-client behavior (Fang et al., 2020).
- Optimization-Based Attacks:Attackers produce fraudulent examples specifically made to guide a global model toward predefined objectives through optimal design of their falsified data. Attackers use stealthy approaches because these methods successfully bypass defensive measures, according to the findings of (Bagdasaryan et al., 2020; Fang et al., 2020).

  Each of these attacks exploits the fact that individual client contributions in FL are aggregated without a full inspection of their origin or intent.

3.3 Assumptions and Attack Scenarios in Federated Learning

In federated learning, attackers typically assume the role of malicious clients. These clients:

- Possess full control over their local data.
- May collaborate with other attackers (sybil attacks).
- Can selectively manipulate their local model before sending updates to the central server.

- Cross-device FL vs. cross-silo FL: In cross-device FL, malicious users are often individuals with weaker devices; in cross-silo FL, institutions (e.g., hospitals) may collude or be compromised (Kairouz et al., 2021).
- Partial Participation: At any given time, only certain clients are selected to participate in training rounds in federated learning systems. This random selection mechanism allows attackers to opportunistically join and control multiple training sequences (Fang et al., 2020).

  Federated learning protects user privacy by preventing the release of local data to the server, thus rendering detection attempts very difficult (Kairouz et al., 2021). The deployment of non-IID data distributions during attacks helps adversaries conceal unauthorized activities from the aggregation process (Bhagoji et al., 2019; Fang et al., 2020).

3.4 Impact on Model Performance and Privacy

The impact of data poisoning can be substantial:

- Untargeted poisoning can lead to a significant decrease in overall model accuracy, especially when the number of poisoned clients exceeds a critical threshold (Fang et al., 2020).
- Bias and fairness issues: Targeted poisoning can skew predictions for specific classes or demographics.
- Security breaches: Backdoor attacks may be exploited to bypass systems (e.g., unlocking a phone using a specific sticker pattern).
- Attacks can remain undetected for long periods, especially in systems with asynchronous or sparse client participation (Tolpegin et al., 2020).
- Generalization failure : Models demonstrating good testing outcomes with poisoned information exhibit poor results when deployed in actual environments since they show unpredictable reactions to novel conditions (Fang et al., 2020).

- Influence on aggregation algorithms: Some poisoning attacks specifically target aggregation schemes including FedAvg and Krum to manipulate their behavior until the server emits biased model updates resulting in irreversible model convergence changes (Bagdasaryan et al., 2020).

3.5 Defense Mechanisms

Several defenses have been proposed:

Tabella 3.1: Defense mechanisms in federated learning

Defense	**Approach**	**Effectiveness**	**Reference**
Krum	Chooses the update closest to others	Good for small # attackers	Blanchard et al., 2017
Trimmed Mean/Median	Filters outliers by coordinate-wise aggregation	Works well with symmetric attacks	Yin et al., 2018
Norm Clipping	Limits the size of updates	Helps against extreme updates	McMahan et al., 2017
Differential Privacy (DP)	Adds noise to updates to mask individual data	Mitigates poisoning + inference risk	Abadi et al., 2016 or McMahan et al., 2017.
Anomaly Detection	Flags clients with strange update patterns	Hard in non-IID settings	Sun et al., 2019 (Can You Really Backdoor Federated Learning); Chen et al., 2020 (Activation Clustering)
FLTrust	Uses a small trusted dataset on the server to scale and validate updates	Highly effective, but requires centralized trust	Cao et al., 2021
Adaptive Federated Optimization (AFO)	Dynamically adjusts aggregation based on client behavior	Adaptive to changing threats, experimental	Xie et al., 2021

The security defenses each focus on blocking specific attacks but generally need users to choose between stronger security measures and better model performance. The use of differential privacy in these systems results in lower model accuracy and robust aggregators such as Krum have difficulties combating Sybil attacks. The FLTrust solution requires access to trusted data but such conditions might not exist in all possible applications. The research communi-

ty now supports integrated protective strategies which unite powerful statistics alongside adaptable aggregation and privacy protection methods to challenge security threats spanning across diverse fields found in federated learning (Cao et al., 2021; Xie et al., 2021).

3.6 Case Studies

Case Study 1: Tolpegin et al. (2020) – Data Poisoning in FL. Tolpegin et al. Tolpegin et al. 2020 established through experimentation that malice from only 10% of clients resulted in over 40% reduction of test accuracy in their federated image classification task. The researchers discovered that repetitive label-flipping attacks delivered significant results.

Case Study 2: Lyu et al. (2020) – Survey and Evaluation of FL Attacks. The paper by Lyu et al. Lyu, Yu e Yang 2020 established an extensive taxonomy of poisoning methods followed by an evaluation of protection strategies in different suspected attack situations. Among different defenses studied, there exists no solution that can defend against all attacks while stealthy and adaptive attackers render defenses ineffective.

Case Study 3: Bagdasaryan et al. (2020) – Stealthy Backdoor Injection with Minimal Poisoning. This study shows that even a single malicious client in FL can inject a backdoor by modifying only a small portion of their local data with a visible trigger (e.g., a pixel pattern). The backdoor succeeds without affecting overall model accuracy, making it very hard to detect Bagdasaryan et al. 2020.

Case Study 4: Xie et al. (2021). Here, several clients each embed part of a trigger into the model updates. The full backdoor is activated only when all pieces come together during inference, allowing the attack to evade most existing defense strategies Xie et al. 2021.

3.7 Summary

The chapter studied one of the main threats that affects distributed learning through data poisoning attacks. These attacks take advantage of the decentralized nature of Federated Learning, which conducts local training operations on each device, but distributes only model update content. The security approach allows detrimental clients to influence training while maintaining secrecy about their motives. The first step involved dividing poisoning attacks into untargeted attacks, targeted attacks, and backdoor attacks. The goal of untargeted attacks is to lower model accuracy throughout the system, while targeted attacks make the model make mistakes on particular input patterns. A hidden, clandestine behavior enters models through backdoor attacks because they normally execute correctly but provide wrong outputs during specific pattern activation. These newer forms of attack included the Sybil attack, which enables one adversary to operate multiple fake clients, and optimization-based attacks that use specifically crafted updates to stealthily damage the model. The danger level of attacks increases in federated settings due to non-IID data alongside partial client system participation. Attackers find it simpler to hide their harmful behaviour because of these system characteristics. Small numbers of poisoned clients will significantly alter the global model results because incorrect update assessment and analysis processes are not performed. We examined the effects that result from poisoning attacks for successful execution. Model accuracy degradation, along with problematic behaviour, unfair treatment of classes and users, and potential malicious backdoor usage, stand as major consequences of poisoning attacks. Traditional validation testing often produces incorrect results because poisoned models continue to function properly even though they fail when actually deployed. Detecting such models turns out to be more challenging for these reasons. Various defense methods have been developed by researchers to fight the identified risks. Two types of aggregation methods (Krum and Trimmed Mean) alongside norm clipping constraints and privacy measures through differential privacy defend FL but eliminate potential outlier updates. FLTru-

st represents one of the modern solutions in consent-based federated learning, which brings a trusted server-side dataset to perform server-side validation of client updates. Most defense strategies for federated learning face limitations because they depend on ineffective assumptions and performance reduction, or can be circumvented by sophisticated attackers. We concluded with an overview of four detailed case studies from recent literature. These showed how attacks can be carried out using simple label flips, carefully crafted updates, or even complex coordinated behavior across multiple malicious clients. The takeaway from these examples is clear: poisoning attacks are practical, evolving, and still difficult to defend against reliably. This chapter emphasizes that protecting federated learning from poisoning attacks isn't something a single solution can handle. It takes a flexible, layered strategy that can adapt as threats evolve. As FL continues to be used in sensitive areas such as healthcare and finance, ensuring the trustworthiness of its models is not only a technical challenge but a necessary step for safe deployment.

Capitolo 4

Membership Inference Attacks in Distributed Learning

4.1 Introduction

The rising popularity of distributed learning methods makes indirect privacy violation an urgent problem. Attackers facing the challenge of Membership Inference try to establish whether specific records exist within the training pool of machine learning models. The attacks represent critical privacy breaches which particularly threaten sensitive information including medical records and personal user data according to Shokri et al. (2017). FL within distributed systems presents an enhanced privacy threat because its architecture creates additional concerns. The privacy-boosting mechanism of FL through local data storage remains at risk because MIAs prove effective when studying update exchanges between clients and the server (Nasr et al., 2019). The discovered vulnerabilities challenge FL's ability to deliver expected privacy benefits which has prompted researchers to pursue solutions for their remediation.

4.2 Attack Principles

The tendency of models to remember training patterns forms the basis of MIA attacks. During the training process models exhibit different behavior on expo-

sed data when compared to new unseen input based on measures of confidence scores and output distributions or gradients according to Shokri et al. (2017).

MIA setting occurs in two principal formats:

- **Black-box attacks:** The adversary only observes the model's output (e.g., confidence scores). This is the most common scenario in cloud-deployed ML APIs.
- **White-box attacks:** The attacker has internal access to the model, such as the gradients, parameters, or intermediate layers. This is more applicable in FL, where clients may access local models.

  The adversary adopts two different approaches to distinguish training set members from non-members by using shadow modeling or statistical techniques.

4.3 MIAs in Federated Learning

Raw client data stays within FL systems while the exchanged gradients and updates might reveal details about what the clients train with. Nasr et al. (2019) proved that monitoring model updates through time would enable white-box MIAs to launch successful attacks against federated systems. Key vulnerabilities in FL include:

- Repeated participation of clients, which allows pattern tracking.
- Overfitting on local data, making member samples stand out.
- Heterogeneous data distributions (non-IID) make it easier to distinguish rare or unique data records (Nasr et al., 2019; Melis et al., 2019).

  These properties make FL models more prone to membership leakage than originally assumed.

4.4 Evaluation Metrics and Assumptions

The success of MIAs is usually evaluated using:

- Attack Accuracy: Fraction of correct membership predictions.
- Precision and Recall: Used when the attacker is selective.
- Shadow Model Effectiveness: How well a mimic model performs on data drawn from similar distributions.

Typical assumptions include:

In real-world settings, these assumptions may not fully hold, but partial access is often enough to mount meaningful attacks.

4.5 Defense Mechanisms

Several defenses have been proposed to protect against MIAs:

Tabella 4.1: Defense mechanisms against membership inference attacks

Defense	**Idea**	**Effectiveness**	**Sources**
Differential Privacy (DP)	Adds noise to gradients or updates	Effective, but may reduce accuracy	Abadi et al. (2016)
Regularization (L2, dropout)	Reduces overfitting	Inexpensive and effective in many cases	Shokri et al. (2017)
Early Stopping	Prevents overtraining	Simple but can hurt performance	Nasr et al. (2019)
Adversarial Regularization	Penalizes information leakage	Promising but still experimental	Jayaraman & Evans (2019)
Update Sanitization	Filters sensitive gradients before sharing	Trade-off between privacy and utility	Nasr et al. (2019)

Most defenses involve a trade-off between privacy and utility, and no single solution is universally adopted

4.6 Case Studies

Case Study 1: Shokri et al. (2017) – The First MIA Framework.Shokri et al. 2017 The first conceptual framework and evaluation framework for black-box membership inference attacks was established by Shokri et al. Shadow modeling was used by Shokri et al. as they developed a framework to replicate target model functionality while an attack model identified member versus non-members. The results achieved 80% model accuracy in specific situations yet demonstrated how models with overfitting became easy targets.

Case Study 2: Nasr et al. (2019) – White-box MIAs in FL.Nasr, Shokri e Houmansadr 2019 Nasr et al. implemented gradient analysis on MIAs to create a system for federated learning security. FL clients who participate in the process are able to monitor local updates and establish membership through this capability. The researchers found that MIA success rates were better when performed in a FL environment compared to centralized experiments on CIFAR-100 and Purchase datasets.

Case Study 3: Shokri et al. (2017) – Shadow Model Membership Inference.Shokri et al. 2017 Shokri et al. introduced in 2017 what remains a prominent example of demonstrating Membership Inference Attacks (MIAs). The research presented a black-box attack framework which allowed an adversary to determine if particular training data points were utilized in the development of target machine learning models although they lacked knowledge about model parameters. The main conceptual advancement in their approach involved shadow models. Attackers create shadow models through training that replicates target models while operating on data distributions that match those of the target. The attacker studies the differences in shadow model responses to acquired and unacquired inputs to discover patterns that differentiate member from non-member cases.The authors developed an attack model through which they trained a classifier to spot differences between target model output responses for member versus non-member inputs using prediction confidence metrics, entropy measurements,or probability vector patterns. The attack model learns

to detect membership status through processing real target model outputs after training procedure completion.

4.7 Results and Effectiveness

The study tested this method on several datasets, including:

- Purchase-100 (a binary classification dataset with 600 dimensions)
- Texas Hospital Discharge Data
- CIFAR-10, a widely-used image classification dataset.

Across all cases, the membership inference attack was significantly more accurate than random guessing, achieving:

- Over **90%** accuracy on the Purchase dataset,
- Between **60%–80%** accuracy on CIFAR-10, depending on the model configuration.

The attack was particularly effective when:

- The model was highly overfitted to the training data.
- The dataset had imbalanced or rare classes
- The output layer provided rich probability distributions instead of just final predicted labels.

The conducted research proved that models remain exposed to privacy breaches even when attackers lack direct access to the model (black-box scenario). The study demonstrated that preventing model weight access or algorithm knowledge alone is insufficient for protecting privacy.

Relevance to Federated Learning

Despite their study being built upon centralized models, Shokri et al.'s at- tack framework works directly for federated learning (FL). FL systems allow attackers to perform client-based actions as well as intercept model update distributions during periods of operation. The distributed training process in FL tends to overfit data repeatedly because of non-IID data which leads to higher leakage of membership information. FL security suffers from two limitations that allow clients to see only partial model updates because they help build shadow versions of models independently while distributing the vulnerability detections. Researchers conducted examinations which applied Shokri's approach to FL deployments to demonstrate that decentralization of training does not eliminate privacy threats. The study acts as fundamental knowledge for understanding how Member- ship Inference Attacks operate while demonstrating the limitations of standard privacy protection techniques particularly in distributed teamwork.

4.8 Why Overfitting Leads to Membership Leakage

Deep neural networks, along with other machine learning models, face success in membership inference attacks because they overfit training data. A model becomes overfit when it obtains training patterns that do not apply to new data instances. An unintentional side effect of this behaviour generates unnoticeable behavioural shifts between model responses for training samples versus new samples not included in training. The difference between how MIAs process trained data versus new data is what enables adversaries to determine data usage in the training phase. During training, an overfitted model obtains higher prediction confidence levels for data points it encounters. Unseen data requires models to generate less accurate outcomes than normal because their predictions are typically more uncertain. A model-based attacker can use this pattern by measuring both the model outcomes and their certainty ratings. The authors of Shokri et al. (2017) developed a shadow model that replicated the target mo-

del's output to demonstrate this approach. The analysis of probability output data between training set members and non-members enabled the attack model to identify characteristic patterns regarding high-confidence score distributions. The research showed that the attack proved effective by achieving results better than chance probability estimates, even when operating under black-box constraints. Small datasets held by clients represent a specific challenge in federated learning since the problem of overfitting becomes more difficult to resolve. A weakness of local models is their tendency to memorize unique training points, causing them to leak information about shared model parameters or updates. The repeated client participation in FL leads to their overfitted local updates becoming progressively more exposed since their training occurs over multiple rounds. Researchers have formally defined the connection between overfitting and privacy leakage in numerous studies while also clearly demonstrating their understanding. A model's susceptibility to MIAs directly correlates with the generalization gap, which represents the training accuracy vs. test accuracy difference according to Yeom et al. (2018). A wider difference between training and test accuracy makes it more likely for the model to retrieve previously seen training data that attackers could exploit. Dropout, data augmentation, regularization, and early stopping constitute primary defensive tactics to combat MIAs because they control overfitting. These techniques improve the model's generalization while also minimizing its behavioural discrepancies across member and non-member samples to decrease MIA success.

4.9 Label and Attribute Inference Attacks

The ultimate objective of Membership Inference Attacks (MIAs) is to ascertain training set membership, but label inference and attribute inference attacks extend this pursuit with additional data collection objectives. The different variants enable adversaries to obtain supplementary sensitive information about data records by working with partial input knowledge.

Label Inference Attacks

An attacker who receives data attribute values but lacks knowledge about the appropriate target value conducts a label inference attack. The attacker attempts to determine the label selection that the model applied to that particular input for training purposes. This attack method proves harmful, specifically when the labels hold sensitive details such as medical diagnoses and user' political affiliations discovered through observation of their recorded behaviors. Discrete information leakage occurs through analyzing classification probabilities and confidence scores produced by a model. An overconfident or over-fitted model produces results that expose its judgment regarding correct labels when these responses stem from the training data samples. Departures can perform a dual attack where they identify whether training data includes a specific record along with its attached label.

Attribute Inference Attacks

The objective of an attribute inference attack involves determining unknown characteristics of data records. Attackers who have information about a target's age, along with their income level and education status, attempt attribute inference attacks that aim to guess the target's undisclosed marital status using the model's behavior and predicted values. This type of attack serves as a threat because it enables the discovery of private characteristics that the parties never disclosed to others. The attackers succeed in obtaining private information from partially observable data outputs without needing access to model parameters. The basic idea behind attribute inference attacks assumes that ML models learn strong relationships between training input characteristics, so developers can reverse-engineer system weaknesses that lead to overfitting or incorrect generalization.

Real-World Implications

Machine learning models demonstrate the ability to learn feature-label relationships that enemies can exploit during attacks and training data memorization procedures. The risks for both attacks multiply in federated learning operations that demand repeated updates on private client data. A malicious client who participates in the network can determine sensitive user attributes about other users through analysis of shared updates or gradients.

Standard MIAs share the same conditions for label and attribute inference attacks to become more effective when:

- The model is overfitted
- The attacker has auxiliary data
- The model's outputs reveal fine-grained probabilities

Research in this area is still developing, but these attack forms are already being used to analyze the limits of privacy in deep learning systems, especially under realistic, black-box conditions.

4.10 Summary

The exposure of training data through membership inference attacks poses an essential privacy risk to distributed learning operations. The overfitting of the model and additional information, along with the access to the gradient or confidence score, determine the successful results of these attacks. The privacy-preserving system of federated learning becomes vulnerable to attacks due to repeated client involvement along with heterogeneous data types. Defenses that use differential privacy together with regularization add substantial performance limitations to their deployment. Distributed security machine learning research focuses heavily on solving the ongoing problem of balancing system functionality with precision alongside privacy protection, while this challenge persists as an active research focus.

Capitolo 5

Proposed Hybrid Defense Framework

5.1 Motivation for a Hybrid Approach

Evaluating defenses against data poisoning and membership inference attacks (MIAs) reveals that there is no single option that provides thorough protection without problems. Methods like differential privacy (Abadi et al., 2016) make the system less vulnerable but may reduce accuracy.On the other hand,defenses using aggregation, such as Krum (Blanchard et al., 2017), are weak in the face of Sybil attacks. Because of these issues, there is a need for an approach that efficiently considers security, privacy, and how the model is used. We suggest a mixed defense system that joins helpful methods designed for federated learning (FL), mainly in cases where the data is not identical, trust is limited, and resources are scarce.

5.2 Framework Overview

The proposed framework includes five integrated components:

1. **Client Trust Scoring via FLTrust**
    - A small, trusted dataset on the server side is used to scale incoming client updates (Cao et al., 2021).
    - Low-trust clients are down-weighted or excluded from aggregation.
2. **Differential Privacy with Gradient Clipping**

- Each client's model updates are clipped to a norm bound and noised using DP (e.g., DP-SGD).

3. **Anomaly Detection Module**
    - Statistical tests (e.g., Mahalanobis distance) are run on incoming updates to flag anomalies (Sun et al., 2019; Chen et al., 2020).
    - Outliers are filtered before aggregation, targeting stealthy poisoning attacks.
4. **Adaptive Aggregation Strategy (e.g., Krum + Median)**
    - The framework selects an aggregation strategy based on update distribution (Yin et al., 2018).
    - Example: Krum for sparse client activity, median for symmetric non-IID data.
5. **Client Update Logging for Post-hoc Auditing**
    - Optional: lightweight blockchain or secure logging mechanism to track model updates (Kim et al., 2020).
    - Helps diagnose failed defenses and monitor suspicious participation patterns.

5.3 Benefits and Limitations

Benefits:

- Combines **proactive and reactive** defenses.
- Adapts to varying **attack surfaces and client behavior**.
- Maintains **model utility** better than isolated defenses.
- Offers **explainability and traceability** with optional logging.

Limitations:

- Requires **computational overhead** on the server.
- Depends on a small trusted dataset (FLTrust), which may be unavailable.
- DP noise may still lower model precision if privacy budget is too strict.

5.4 Integration into Real Systems

This framework is suitable for:

- **Healthcare federated systems**, where trust can be partially established (e.g., among hospitals).
- **Edge-device deployments** with large-scale participation and sparse activity.
- **Financial services**, where sensitive data must be protected under the regulation.

Capitolo 6

Real-World Implications and Ethical Considerations

6.1 The Stakes of Federated Learning in Sensitive Domains

Federated learning (FL) is increasingly deployed in high-stakes domains such as healthcare, finance, and transportation, and now often uses Federated Learning (FL). It is very important in these areas for personal data to remain private, models to be reliable, and those involved to be responsible. However, having a decentralized network is what attracts users, but it also makes FL more vulnerable. For example, by using FL, hospitals in healthcare can teach their diagnostic models together without disclosing data on their patients. However, a successful membership inference attack could uncover if a single patient's record was accessed, which is against the Health Insurance Portability and Accountability Act (HIPAA). In addition, data poisoning in this area can cause incorrect diagnoses, increasing the risk of wrong treatment choices being made. FL in financial services promotes fraud discovery and managing risks in companies that operate in different places. If fraudulent clients alter the model with false information or wrong behavior, there could be errors in credit scoring, rejected loans, or fraud that goes unnoticed. When these instances occur, it is the law and ethics that require algorithmic accountability.

6.2 Ethical Considerations

Distributed learning systems face unique ethical dilemmas:

- **Informed Consent**: Users are generally unaware when their data indirectly influence FL models. Without using central storage, gradients may still share information, making people concerned about implied consent.
- **Power Imbalance**: FL models can be set up by big companies for smaller users without making the change clear, which reduces the user's freedom.
- **Trust and Explainability**: Users and organizations must trust the system's fairness and integrity. Poisoned models or membership leaks undermine trust and challenge **AI explainability** standards.

Ethical AI frameworks, like those proposed by the**EU High-Level Expert Group on AI**, call for:

- Privacy preservation
- Technical robustness
- Transparency
- Accountability

Federated systems must explicitly adopt these principles to ensure responsible AI deployment.

6.3 Legal and Regulatory Implications

Several legal frameworks are directly affected by vulnerabilities in FL *Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation)* 2016; *Health Insurance Portability and Accountability Act of 1996* 1996; *Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)* 2021:

- **GDPR (General Data Protection Regulation)** *Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation)* 2016: Even if the data isn't released directly, if systems can be reversed to expose particular people (for example through MIAs), it may still breach the data minimization and privacy by design requirements.
- **HIPAA (U.S. healthcare law)** *Health Insurance Portability and Accountability Act of 1996* 1996: Membership inference can reveal patient involvement, violating privacy standards.
- **AI Act (EU Proposal)** *Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)* 2021: Emphasizes risk assessment and human oversight, both of which become challenging in opaque federated systems.

Additionally, **auditability** becomes a legal necessity. Organizations deploying FL models may be required to demonstrate:

- Provenance of updates
- Defensive strategies in place
- Response plans in case of breaches

6.4 Hypothetical Attack Scenarios

The following hypothetical scenarios help demonstrate the practical risks of these attacks:

Scenario 1: Poisoned Medical Diagnosis

An attacker gains access to a local hospital's FL client and injects manipulated training data to misclassify a rare disease. The global model performs well on common diagnoses but fails for this condition putting specific patients at risk.

Scenario 2: MIA in Political Opinion Dataset

A social media company uses FL to train a sentiment analysis model. A malicious insider launches a membership inference attack to confirm whether

a user posted political content. This violates user anonymity and may lead to reputational or legal consequences.

6.5 The Path Forward: Building Responsible FL Systems

Federated learning has great potential to transform data privacy, but only if built with **defense-first** and **ethics-aware** principles. Developers, researchers, and regulators should collaborate to:

- **Embed privacy safeguards** by default (e.g., DP, secure aggregation)
- **Ensure transparency** in how user data influences models
- **Mandate ethical audits** and risk assessments for high-stakes applications
- **Create redress mechanisms** for users harmed by model misuse or attack

Capitolo 7

Conclusion and Open Challenges

7.1 Conclusion

The research investigated two major threats targeting distributed learning systems, namely data poisoning attacks and membership inference attacks (MIAs). The research in this work demonstrates that FL achieves its intended privacy goals; however, distributed machine learning systems continue to face serious threats from privacy-based and integrity-based vulnerabilities. Research has demonstrated that data poisoning attacks successfully lead to major model accuracy decline as well as deliberate backdoor insertion for malicious purposes. FL makes it difficult to detect poisoning attacks because client data are non-IID, there is restricted access to raw data, and the update process occurs asynchronously (Bagdasaryan et al., 2020; Fang et al., 2020). Security mechanisms such as robust aggregation, along with anomaly detection and norm clipping, provide limited protection in real FL environments but struggle with assumptions that do not apply in practice and face efficiency limitations (Kairouz et al., 2021). User data protection is endangered by membership inference attacks because adversaries can determine if specific input points participated in model training. Studies by Nasr et al. (2019) and Shokri et al. (2017) have revealed that data decentralization in FL remains vulnerable, since gradient updates can reveal membership information during repeated client participation or model overfitting conditions. Performance trade-offs from differential privacy techniques, along with regularization methods, reduce their usefulness for practical use. The examination

showed that no isolated defence approach can fully protect against current protection threats. Several defence mechanisms specifically designed for a system's unique characteristics and its tolerance for risks need to be properly combined for comprehensive protection.

7.2 Open Research Challenges

Despite substantial progress in understanding and addressing privacy and security threats in distributed learning, several **open challenges** remain:

Challenge 1: Adaptive and Stealthy Adversaries Today's security approaches deal with predefined attacker methods. Attacks carried out by actual perpetrators adjust their strategies during operational periods to slip past security detection systems. Scientific analysis must explore defensive mechanisms that adjust their approach to comply with the modifications of the attacker method (Fang et al., 2020; Kairouz et al., 2021)..

Challenge 2: Realistic Assumptions and Datasets

Most evaluations are conducted under simplified or ideal conditions. Future work should focus on:

- Heterogeneous (non-IID) datasets,
- Partial participation scenarios,
- Resource-limited edge devices, to better reflect real deployment conditions (Kairouz et al., 2021)

Challenge 3: Balancing Utility and Security

The implementation of defense systems usually results in decreased model utility for enhanced privacy or robustness. Researchers need to create energy-efficient model algorithms that protect privacy while sustaining high performance levels according to Kairouz et al., 2021.

Challenge 4: Formal Verification and Guarantees

Privacy and robustness assurances remain almost nonexistent in current defensive models. FL systems gain trustworthiness through the development of models that offer demonstrated security limitations against MIAs and poisoning attacks (Kairouz, P., McMahan, H. B., Avent, B., et al. (2021). Advances and Open Problems in Federated Learning. Foundations and Trends in Machine Learning).

Challenge 5: Integrating Cryptographic Solutions

Although tools like **Secure Multiparty Computation (SMPC)** and **Homomorphic Encryption** are promising, they remain underused due to complexity and computational cost. Research on **scalable cryptographic solutions** for FL is needed.

7.3 Final Remarks

AI privacy definitions will change through distributed learning, yet it creates multiple entry points that demand vigilant protection. The deployment of FL in security-sensitive domains requires essential protection measures for models that include robustness and privacy preservation. This thesis helps to achieve this goal through threat analysis of essential threat classes along with current defense assessments and a discussion of open challenges alongside trade-offs. Additional research should create integrated defensive frameworks that merge security and privacy functionality together with efficient system operation to reach trustworthy decentralized artificial intelligence.

Bibliografia

Abadi, Martin et al. (2016). «Deep Learning with Differential Privacy». In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 308–318.

Bagdasaryan, Eugene et al. (2020). «How To Backdoor Federated Learning». In: *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2938–2948.

Bhagoji, Arjun Nitin et al. (2019). «Analyzing Federated Learning through an Adversarial Lens». In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 634–643.

Blanchard, Peva, Rachid Guerraoui e Julien Stainer (2017). «Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent». In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 119–129.

Bonawitz, Keith et al. (2017). «Practical Secure Aggregation for Privacy-Preserving Machine Learning». In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 1175–1191.

Cao, Xiaoyu, Mikhail Khodak e Neil Zhenqiang Gong (2021). «FLTrust: Byzantine-Robust Federated Learning via Trust Bootstrapping». In: *Network and Distributed System Security Symposium (NDSS)*.

Chen, Bryan et al. (2019). «Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering». In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 1, pp. 2727–2735.

European Commission High-Level Expert Group on Artificial Intelligence (2019). *Ethics Guidelines for Trustworthy AI*. url: `https://ec.europa.eu/futurium/en/ai-alliance-consultation`.

Fang, Minghong et al. (2020). «Local Model Poisoning Attacks to Byzantine-Robust Federated Learning». In: *29th USENIX Security Symposium*, pp. 1605–1622.

Health Insurance Portability and Accountability Act of 1996 (1996). url: `https://www.hhs.gov/hipaa/`.

Jayaraman, Bargav e David Evans (2019). «Evaluating Differentially Private Machine Learning in Practice». In: *USENIX Security Symposium*, pp. 1895–1912.

Kairouz, Peter et al. (2021). «Advances and Open Problems in Federated Learning». In: *Foundations and Trends in Machine Learning* 14.1–2, pp. 1–210.

Kim, Hyoungshick et al. (2020). «BlockFL: Blockchain-enabled Federated Learning for Privacy-Preserving and Secure Decentralized Learning». In: *IEEE Transactions on Parallel and Distributed Systems* 32.9, pp. 2156–2168.

Lyu, Lingjuan, Han Yu e Qiang Yang (2020). «Threats to Federated Learning: A Survey». In: *arXiv preprint arXiv:2003.02133*.

McMahan, H. Brendan et al. (2017). «Communication-Efficient Learning of Deep Networks from Decentralized Data». In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282.

Melis, Laura et al. (2019). «Exploiting Unintended Feature Leakage in Collaborative Learning». In: *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, pp. 691–706.

Nasr, Milad, Reza Shokri e Amir Houmansadr (2019). «Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning». In: *IEEE Symposium on Security and Privacy*, pp. 739–753.

Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) (2021). url: `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206`.

Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation) (2016). url: `https://eur-lex.europa.eu/eli/reg/2016/679/oj`.

Salem, Ahmed et al. (2019). «ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models». In: *Network and Distributed System Security Symposium (NDSS)*.

Shokri, Reza et al. (2017). «Membership Inference Attacks Against Machine Learning Models». In: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, pp. 3–18.

Tolpegin, Vahideh Nasr et al. (2020). «Data Poisoning Attacks Against Federated Learning Systems». In: *European Symposium on Research in Computer Security (ESORICS)*, pp. 480–501.

Truex, Stacy et al. (2019). «Demystifying Membership Inference Attacks in Machine Learning as a Service». In: *Proceedings of the 2019 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 418–432.

Xie, Chulin et al. (2021). «DBA: Distributed Backdoor Attacks against Federated Learning». In: *arXiv preprint arXiv:2012.06785*. url: `https://arxiv.org/abs/2012.06785`.

Yeom, Samuel et al. (2018). «Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting». In: *IEEE Computer Security Foundations Symposium*, pp. 268–282.

Yin, Dong et al. (2018). «Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates». In: *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 5636–5645.

Zhang, Xinyang et al. (2019). «Can You Really Backdoor Federated Learning?» In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 32, pp. 4933–4943.