A Review of Federated Learning Attacks: Threat Models and Defence Strategies

Fizlin Zakaria, Shamsul KamalAhmad Khalid
Faculty of Computer Science & Information Technology, Universiti Tun Hussein Onn,
86400 Parit Raja, Batu Pahat, Johor, Malaysia

Abstract—**Federated Learning (FL) has emerged as a critical paradigm in privacy-preserving machine learning, enabling collaborative model training across decentralised devices without sharing raw data. While FL enhances privacy by maintaining data locality, it remains susceptible to sophisticated adversarial attacks. This review systematically analyses the FL threat landscape and introduces a novel taxonomy that classifies attack models based on their objectives, capabilities, and exploited vulnerabilities. Major categories include data poisoning, inference attacks, and Byzantine behaviours, each examined in terms of mechanisms, assumptions, and system impact. In addition, the paper evaluates prominent defence strategies—such as differential privacy, secure aggregation, and anomaly detection—by assessing their strengths, limitations, and real-world applicability. Key gaps include the lack of standardised evaluation metrics and limited exploration of adaptive defence mechanisms. Emerging trends such as homomorphic encryption, secure multi-party computation, and blockchain-based verifiability are also discussed. This review is a comprehensive resource for researchers and practitioners aiming to design resilient, privacy-aware FL systems that withstand evolving threats.**

Keywords—Federated learning; threat models; defence strategies; privacy-preserving AI; adversarial attacks

I. Introduction

The healthcare sector has experienced significant technological advancements, underscoring the need to protect sensitive patient data [1]. Federated Learning (FL), a decentralised machine learning approach, offers promising privacy benefits by enabling local data processing. However, this decentralisation also introduces novel security vulnerabilities that require thorough investigation [2]. Despite its advantage in preserving data locality, FL remains susceptible to attacks from malicious participants [2]. Its ability to enable collaborative model training without compromising raw data makes FL particularly valuable in privacy-critical domains. This includes fields such as healthcare and finance, where leveraging distributed datasets without violating privacy regulations is essential [1].

Although interest in FL security has grown, existing literature often focuses on isolated aspects of the threat landscape, resulting in a fragmented understanding of its vulnerabilities and corresponding defence mechanisms [3]. A more holistic and structured review is needed to address this gap, integrating technical, regulatory, and ethical perspectives on safeguarding federated systems.

This paper addresses this need by providing a systematic taxonomy of FL attack vectors and a comparative evaluation of defence strategies. The review classifies attacks based on threat models—including data and model poisoning—and analyses their operational mechanisms. It further assesses defence methods such as differential privacy and secure aggregation in terms of their effectiveness, trade-offs, and applicability. Finally, the paper identifies open research challenges and proposes directions for future work to enhance the resilience, trustworthiness, and scalability of FL systems [4]. The main contributions of this paper are as follows:

- A novel taxonomy of FL attack models, classifying them based on attack goals, vectors, and system vulnerabilities.
- A comparative evaluation of defence strategies regarding performance, scalability, and security robustness.
- Identifying research gaps, including a lack of standardised evaluation frameworks and limited exploration of adaptive defences.
- Discuss emerging trends such as hybrid defence approaches and blockchain-based accountability mechanisms.
- The remainder of the paper is organised as follows: Section II outlines the literature selection and analysis methodology. Section III presents the taxonomy of FL attacks. Section IV evaluates existing defence strategies. Section V discusses current research gaps and future directions. Section VI and Section VII concludes the paper with a summary of findings and recommendations.

To guide this review, the following research question is posed: What are the key adversarial attack vectors that threaten Federated Learning (FL) systems, and how effective are current defence strategies—such as differential privacy, secure aggregation, and blockchain-based mechanisms—in mitigating these threats across different attack surfaces and system configurations? This question frames the scope of the analysis presented in this paper, which aims to systematically categorise existing FL attacks and critically evaluate the corresponding defence mechanisms based on their applicability, strengths, and limitations. To contextualise this study, Section II reviews foundational and recent research on FL threats and defences.

II. Related Work

Federated Learning (FL) has emerged as a prominent privacy-preserving machine learning framework, attracting increasing attention from academia and industry. Numerous studies have examined its security vulnerabilities and mitigation strategies, but many existing reviews are either narrowly scoped or lack updates to reflect the fast-evolving threat landscape. Pioneering works by Shokri et al. [23] and Melis et al. [24] introduced foundational concepts such as membership inference and property inference attacks, exposing the privacy risks inherent in collaborative learning environments. Bagdasaryan et al. [6] further demonstrated the feasibility of model poisoning and backdoor attacks, showing that even minor manipulations can compromise global model integrity. These studies laid the groundwork for analysing adversarial risks in decentralised learning systems.

Subsequent reviews by Hallaji et al. [2] and Nguyen et al. [11] broadened the threat taxonomy to include Byzantine behaviours and gradient leakage. However, their discussions often stop short of connecting these threats to comprehensive, layered defence architectures. Other researchers have explored isolated countermeasures—such as differential privacy [39] and homomorphic encryption [1]—but rarely assess their real-world performance in heterogeneous, large-scale FL deployments.

To address concerns over trust and transparency, blockchain-based FL frameworks (e.g., ShareChain [1], PPBFL [33]) have been proposed. These aim to decentralise trust management, offering tamper-evident logging and client accountability. Despite their conceptual appeal, such frameworks remain in the early stages of development and face considerable challenges related to scalability, latency, and integration with existing FL protocols [35].

This review builds upon previous work by offering a more comprehensive taxonomy of FL attacks, explicitly categorised by attack surface, timing, and intent. Furthermore, it delivers a comparative analysis of defence strategies, focusing on their security effectiveness, computational cost, and deployment trade-offs. This paper also identifies ongoing research gaps, particularly the lack of adaptive, hybrid, and context-aware mechanisms that can evolve alongside sophisticated attack patterns.

III. Methodology

This study follows a structured literature review approach to comprehensively examine Federated Learning (FL) attack models and corresponding defence strategies. The methodology adopted is outlined as follows.

A. Literature Search Strategy

This research employs a systematic literature review methodology, complemented by elements of a narrative review, to investigate the landscape of security threats and defence mechanisms within federated learning systems. The review focuses on identifying, classifying, and synthesising existing research concerning attacks targeting federated learning and corresponding defence strategies designed to mitigate these vulnerabilities [1], [5]. The study's primary objective is to provide a comprehensive overview of the current state-of-the-art, highlighting prominent attack vectors, defence approaches, and open challenges in the field.

The following keywords and Boolean combinations were used during the search:

- "Federated Learning" AND "Attack"
- "Federated Learning" AND "Threat Models"
- "Federated Learning" AND "Security"
- "Federated Learning" AND "Privacy Attacks"
- "Federated Learning" AND "Defence Strategies"
- "Secure Federated Learning" AND "Robust Aggregation"

Additionally, backwards and forward snowballing techniques were applied by reviewing the references of highly cited papers to identify additional relevant studies.

B. Inclusion and Exclusion Criteria

The literature search encompassed several prominent databases, including Scopus, Web of Science, and IEEE Xplore, selected for their extensive coverage of computer science, engineering, and related disciplines. These databases were queried using keywords and Boolean operators to identify relevant publications on federated learning security [3]. Search terms included variations and combinations of "Federated Learning Attack," "Threat Models," and "Federated Learning Defence" to capture a broad spectrum of research related to the topic. The initial search yielded a substantial number of articles, which were subsequently filtered based on predefined inclusion and exclusion criteria. To ensure the relevance and currency of the review, only peer-reviewed, English-language papers published between 2020 and 2025 were considered for inclusion. This timeframe was chosen to capture the most recent advancements and emerging trends in the federated learning security field—the restriction to peer-reviewed publications aimed to maintain the quality and rigour of the included studies. The English language restriction was applied due to resource constraints and the need for consistent linguistic analysis [1].

C. Study Selection Process

The selection process followed a structured approach, drawing inspiration from the PRISMA guidelines to enhance transparency and reproducibility. The initial screening involved reviewing the titles and abstracts of the identified articles to assess their relevance to the research question. Articles not explicitly addressing security threats or defence mechanisms in federated learning were excluded. Subsequently, the full texts of the remaining articles were examined in detail to determine their eligibility for inclusion based on the predefined criteria. This thorough evaluation ensured that only studies directly relevant to the scope of the review were considered for further analysis. The selected articles were then subjected to a rigorous analysis to extract key information regarding attack types, defence strategies, evaluation metrics, and experimental settings.

D. Data Extraction and Synthesis

The extracted data were organised and synthesised using a comprehensive analysis framework to categorise and classify the identified attacks and defences. The categorisation was based on the nature of the attack, such as data poisoning, model poisoning, inference attacks, and Byzantine attacks [1]. Data poisoning attacks involve manipulating the training data to compromise the global model, while model poisoning attacks target the model aggregation process [4]. Inference attacks, on the other hand, aim to extract sensitive information from the model or the training data. Byzantine attacks encompass various malicious behaviours, including arbitrary data or model manipulation by compromised participants. Defence mechanisms were categorised based on underlying principles, such as differential privacy, secure aggregation, anomaly detection, and robust aggregation techniques. Differential privacy adds noise to the training data or model updates to protect individual privacy, while secure aggregation protocols ensure the confidentiality of model updates during the aggregation process. Anomaly detection techniques aim to identify and filter malicious or abnormal contributions from compromised participants. Robust aggregation methods, such as median or trimmed mean aggregation, are designed to mitigate the impact of outliers or malicious updates on the global model. Decentralised federated learning introduces new privacy threats [2]. For instance, an inference attack can occur in blockchain-based federated learning [1]. In this paradigm, security analysis is necessary [2]. This categorisation facilitated a structured comparison of different approaches and enabled the identification of their strengths and weaknesses. Existing defences against data poisoning cannot be directly applied to federated learning due to the requirement of accessing the training data [6]. Adversaries are improving at hiding malicious behaviours from benign ones [7]. Sophisticated attackers are motivated by very high incentives to manipulate the results of the machine learning models [8]. Data owners update their models while keeping the data localised, guaranteeing zero leakage of individual data since only trained models are shared in the proposed model [1]. The absence of a centralised server in blockchain-enabled federated learning eliminates the single point of failure [1]. By integrating federated learning with blockchain technology, auditing machine learning models is supported without centralising the training data [7].

IV. Results

This section presents a structured literature analysis on Federated Learning (FL) attacks and defence strategies. A taxonomy is developed to categorise attack types based on their objectives, Timing, and attack surfaces. Defence mechanisms are evaluated based on their effectiveness, complexity, and applicability.

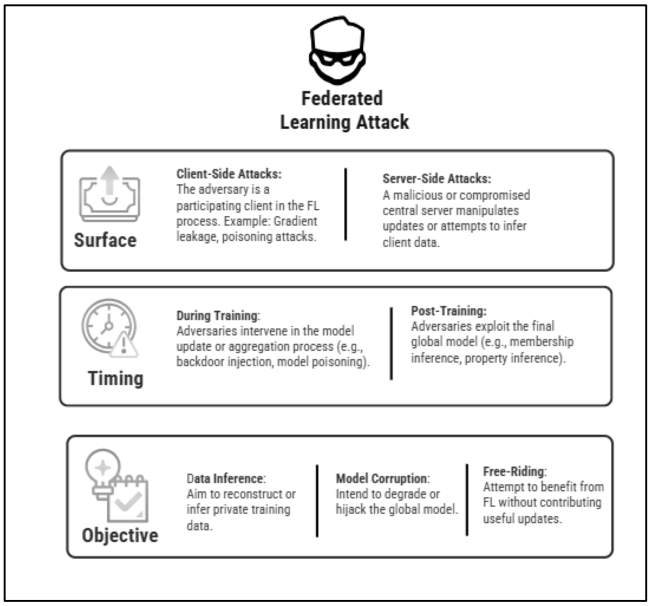
A. Taxonomy of Federated Learning Attacks

Federated Learning (FL) is a distributed machine learning paradigm designed to train models collaboratively across decentralised data sources without transferring raw data. This decentralisation preserves privacy by enabling local computation on each client and aggregating model updates on a central server [1]. However, despite these inherent privacy advantages, FL systems remain vulnerable to a broad spectrum of adversarial threats that can undermine model integrity, compromise data confidentiality, and disrupt system performance [9]. Understanding and classifying these threats is essential for developing effective, context-aware defence mechanisms [5], [9], [10].

Fig. 1 presents a visual taxonomy that categorises FL attacks based on three key dimensions:

- Attack surface (client-side vs. server-side),
- Timing (during training vs. after training),
- Objective (e.g., data inference, model disruption, free-riding).

This classification provides a structured lens to analyse how FL systems are targeted across various operational contexts.



Client-Side Attacks: The adversary is a participating client in the FL process. Example: Gradient leakage, poisoning attacks.

Server-Side Attacks: A malicious or compromised central server manipulates updates or attempts to infer client data.

During Training: Adversaries intervene in the model update or aggregation process (e.g., backdoor injection, model poisoning).

Post-Training: Adversaries exploit the final global model (e.g., membership inference, property inference).

Data Inference: Aim to reconstruct or infer private training data.

Model Corruption: Intend to degrade or hijack the global model.

Free-Riding: Attempt to benefit from FL without contributing useful updates.

Fig. 1. Overview of attack categories in federated learning.

This diagram visualises how FL systems are exposed to threats across client and server surfaces, during various phases (training/post-training), and based on differing attack objectives.

1) By Attack surface

a) Client-side attacks: These attacks originate from compromised or malicious clients within the FL network. A typical example is data poisoning, where adversaries inject manipulated training data to introduce specific biases into the global model [4]. Another form, model poisoning, involves clients altering their model updates before aggregation to influence global model behaviour. These threats exploit the decentralised nature of FL, making it difficult for the server to distinguish between benign and malicious updates [11], [12].

b) Server-side attacks: Server-side attacks target the central aggregation server, which collects client updates and distributes the global model. Adversaries may compromise the server to tamper with the aggregation process or launch inference attacks to extract sensitive client data from model updates. Moreover, the server itself can act maliciously by injecting adversarial gradients or modifying aggregation rules [13], [14], [15].

2) By Timing

a) During training: Attacks launched during the training process are commonly referred to as poisoning attacks. These may involve the injection of malicious data, gradients, or even fake client updates (e.g., Sybil attacks) to degrade the model's accuracy or manipulate its outputs [16], [17]. These threats can be particularly stealthy and difficult to detect in asynchronous FL settings.

b) After training (after deployment): Once an FL model is deployed, it becomes susceptible to inference attacks, such as membership inference (determining if a record was part of training) or property inference (inferring sensitive aggregate attributes). These attacks exploit model outputs or gradients to reveal private training data [18], [20]. Attackers can also launch Reconstruction attacks to extract private textual data, exploiting gradient updates from the clients during the training phase [20].

3) By Objective: Federated learning attacks can be classified based on their primary objectives, reflecting the intended outcome that the attacker seeks. Attacks can also be categorised based on their goals, i.e., what the attacker hopes to achieve.

a) Data inference: Data inference attacks aim to extract sensitive information about the clients' local datasets from the shared model updates or the final global model [21]. Membership inference attacks determine whether a specific data point was used to train the model, revealing sensitive information about individual clients. Property inference attacks aim to infer statistical properties of the clients' local datasets, such as the distribution of sensitive attributes.

b) Model disruption: Model disruption attacks, also known as model poisoning attacks, aim to degrade the performance or functionality of the global model. By injecting malicious data or manipulating model updates, attackers can cause the model to make incorrect predictions or exhibit undesirable behaviour [8]. Such attacks can lead to model performance degradation or complete model failure. An attacker may compromise the global model by manipulating local parameters [22].

c) Free-Riding: Free-riding attacks, also known as lazy client attacks, occur when malicious clients contribute minimally to the training process while benefiting from the global model. These clients may submit outdated or irrelevant model updates, slowing convergence and reducing the model's accuracy [1]. In federated learning, free-riding attacks involve lazy clients that do not contribute meaningfully to the training process [1].

The results are further supported by visual representations, including Fig. 1 through Fig. 3 and Table II, which collectively illustrate the taxonomy of FL attacks, threat-defence mappings, and comparative defence evaluations. These visual aids were designed to enhance the clarity and accessibility of the findings, enabling a layered understanding of the attack vectors and mitigation strategies. Including these diagrams strengthens this review's communicative impact and supports readers in grasping complex adversarial relationships and defence limitations within FL systems.

Table I summarises the FL attack classification with attack surface, Timing, and impact.

TABLE I. SUMMARY OF FEDERATED LEARNING ATTACK TYPES

Attack Type	Description	Attack Surface	Timing	Representative Studies
Membership Inference	Determines if a specific data point was in the training set	Client	Post-training	[23]
Property Inference	Infers statistical properties of client data	Client	Post-training	[24]
Model Poisoning	Injects malicious updates to corrupt the global model	Client	During training	[25]
Backdoor Attacks	Embeds hidden misbehaviour triggered by specific inputs	Client	During training	[6]
Gradient Leakage	Reconstructs raw data from shared gradient updates	Server	During training	[26]
Sybil Attacks	Uses fake identities to influence model updates disproportionately	Client	During training	[27]
Free-Rider Attacks	Participates in FL without contributing useful updates	Client	During training	[28]
Collusion Attacks	Multiple adversaries collaborate to subvert learning	Multiple	During training	[2]
Adaptive Multi-Vector	Dynamic use of multiple attack vectors to evade defences	Multiple	Dynamic	[29]

B. Threat Models

The evolving landscape of Federated Learning (FL) security has seen a marked increase in hybrid and adaptive adversarial behaviours. Contemporary studies reveal that attacks are no longer limited to isolated strategies; adversaries now deploy multi-vector approaches that exploit privacy, model integrity, and communication weaknesses. Examples include coordinated multi-client poisoning, collusion-based inference, and gradient inversion attacks augmented with generative models. The growing sophistication of these threats necessitates the development of layered, context-aware defence mechanisms. Fig. 2 illustrates the key threat models encountered in FL environments, offering a taxonomy of representative attack types based on mechanisms and impact.

1) Membership Inference Attacks (MIA): Membership inference attacks (MIA) pose a fundamental threat to privacy in machine learning. In this attack, an adversary attempts to determine whether a specific data instance was part of the training dataset used to develop the model. The implications are severe, especially when dealing with sensitive datasets such as electronic medical records or financial transactions [30].

These attacks exploit subtle behavioural discrepancies between the model's predictions on seen versus unseen data. For example, models typically return higher confidence scores or lower prediction entropy on training data than on unfamiliar inputs [31]. By leveraging such output characteristics, attackers can infer membership with non-trivial accuracy. Additionally, some variants assess model sensitivity to the presence or removal of specific samples, thereby detecting overfitting or memorisation tendencies.

Research has demonstrated the feasibility of MIAs across a range of model architectures and datasets. Key enabling factors include overfitting, model complexity, and the availability of public query APIs that allow repeated probing [23]. Attackers often rely on shadow models—locally trained replicas mimicking the target model's behaviour—to optimise their inference strategies. While simple mitigation techniques like limiting the number of queries have been proposed [13], these alone are insufficient in high-risk applications.

2) Property inference attacks: Property inference attacks, a closely related threat, extend the scope of privacy breaches by targeting aggregate properties of the training dataset rather than individual data points. In property inference attacks, the adversary aims to infer statistical characteristics or sensitive attributes of the training data, such as the proportion of individuals with a specific medical condition or demographic information. These attacks exploit the model's learned representations to extract information about the overall distribution and characteristics of the training data, potentially revealing sensitive information about the population from which the data was drawn. For instance, an adversary might try to determine a specific disease's prevalence among the individuals trained to train a diagnostic model. The effectiveness of property inference attacks depends on the model's tendency to encode statistical patterns and correlations in the training data, which can then be extracted and analysed.

3) Model poisoning attacks: Model poisoning attacks represent a significant threat to machine learning systems, where adversaries inject malicious data into the training dataset to impair the model's performance or manipulate its behaviour. This attack can have devastating consequences, especially in critical applications such as fraud detection, spam filtering, and medical diagnosis, where the accuracy and reliability of machine learning models are paramount [18]. By carefully crafting and injecting poisoned data points, attackers can subtly alter the model's decision boundaries, leading to incorrect predictions or biased outcomes [18]. The success of model poisoning attacks depends on the attacker's ability to inject enough poisoned data without being detected and the model's sensitivity to changes in the training distribution [8]. Security companies often rely on crowd-sourced threat feeds to train their classifiers, making them a natural injection point for such attacks. Furthermore, in federated learning scenarios, where multiple parties collaboratively train a model, a malicious participant can inject poisoned data, affecting the global model's performance.

4) Backdoor attacks: Backdoor attacks, a particularly insidious form of model poisoning, involve injecting specific triggers into the training data that cause the model to misbehave only when those triggers are present in the input [1]. These triggers can be subtle patterns or features imperceptible to human observers but easily recognised by the compromised model. When a backdoored model encounters an input containing the trigger, it will produce a predetermined, incorrect output, bypassing the model's intended functionality. Backdoor attacks are particularly concerning because they can be challenging to detect, as the model performs normally on most inputs and only exhibits malicious behaviour when the trigger is present. One study notes that backdoor attacks are highly effective when applied to computer vision models without many poisoned examples [12]. Explanation-guided backdoor poisoning attacks have also been investigated for malware classifiers.

5) Gradient leakage attacks: Gradient leakage attacks exploit the information contained in the gradients of a model during training to infer sensitive information about the training data. In distributed learning settings, where model updates are shared among multiple participants, attackers can intercept these gradients and reconstruct the training data or extract sensitive attributes. Gradient leakage attacks can reveal individual data points, statistical properties of the data, or even the labels associated with specific data points. These attacks pose a significant threat to privacy in federated learning and other distributed learning scenarios, where data is decentralised and sensitive.

6) Free-rider attacks: Free-rider attacks occur in collaborative learning settings, such as federated learning, where participants contribute to training a shared model. In a free-rider attack, a malicious participant exploits the contributions of others without contributing meaningfully to the training process. The free-rider may copy the model updates from other participants without performing any local training or contribute low-quality updates that do not improve the model's performance. By exploiting the contributions of others, free-riders can gain access to a high-quality model without incurring the computational costs or expending the effort required for training. The proposed framework prevents attribute disclosure, homogeneity, and background knowledge threats by using local differential privacy, which adds noise to the data [1].

7) Sybil attacks: Sybil attacks, also relevant in distributed learning scenarios, involve an attacker creating multiple fake identities to gain disproportionate influence over the training process. By controlling various "Sybil" clients, the attacker can manipulate the model's updates, inject bias, or even cause the model to learn incorrect patterns. Sybil attacks are particularly challenging to defend against because it can be difficult to distinguish legitimate clients from fake ones. To solve the single point of failure in federated learning, researchers have explored a blockchain-assisted decentralised federated learning that will completely prohibit malicious clients from poisoning the whole learning process or protect the entire process [1]. A blockchain-based federated learning model with secure multi-party computing model verification has been suggested to counter poisoning attacks [1], [19]. Furthermore, secure aggregation protocols can be employed to ensure that the global model update is computed correctly, even in the presence of malicious participants [2], [6]. Federated Learning (FL) systems face increasingly complex and adaptive threats, including multi-vector and coordinated attacks. These attack types vary by mechanism, intent, privacy, or model performance impact. Fig. 2 summarises key threat models in FL, including membership inference, property inference, model poisoning, backdoor attacks, and Sybil attacks, highlighting their mechanisms and potential consequences. Understanding these models provides the foundation for evaluating appropriate countermeasures.

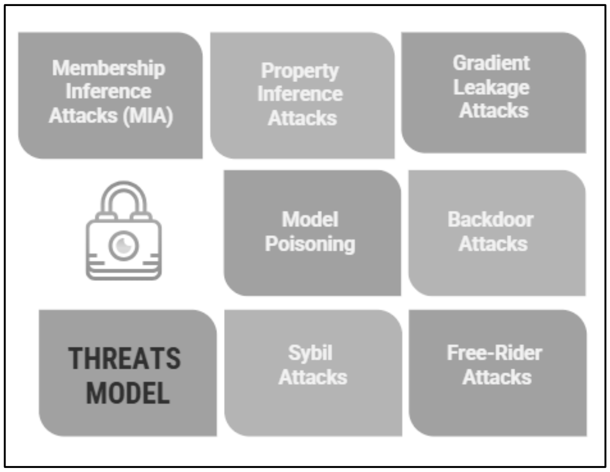


Fig. 2. Threat model of federated learning attacks.

C. Defence Strategies Against Federated Learning Attacks

Recent studies have revealed vulnerabilities in federated learning, demonstrating that model updates can inadvertently leak private information, necessitating robust defence mechanisms [9], [32]. A multifaceted approach encompassing data-level, model-level, and system-level defences is crucial to bolster the security of federated learning systems. Hybrid strategies integrate multiple defence mechanisms and offer a promising avenue for enhanced security and resilience against diverse attacks [2]. Multiple defence mechanisms have been proposed to combat various threats in FL, each addressing specific vulnerabilities. These include privacy-preserving techniques, robust aggregation methods, and decentralised trust frameworks. Fig. 3 maps common FL attack vectors to corresponding defence strategies and highlights the limitations associated with each countermeasure. This visual summary helps clarify the relationships between threats and mitigation approaches, allowing a deeper evaluation of specific techniques.

1) Data-level defences: Data-level defences constitute the first line of protection, focusing on safeguarding individual data points before they contribute to the global model [9]. Local Differential Privacy is a prominent technique, injecting noise into the client's local updates to obscure the contribution of individual data points [1]. Homomorphic Encryption offers another layer of protection by allowing computations on encrypted data, ensuring that the server only processes encrypted model updates, thus preventing access to raw data [19]. Applying these approaches can effectively mitigate privacy risks associated with sensitive training data [19]. Employing these methods can substantially diminish the risk of privacy breaches, reinforcing the protection of sensitive training data.

2) Model-level defences: Model-level defences operate at the aggregation stage, aiming to mitigate the impact of malicious or noisy updates on the global model [5]. Robust aggregation techniques, such as Krum, Median, Bulyan, and Trimmed Mean, identify and discard or down-weight potentially malicious updates, ensuring the global model remains robust despite adversarial clients. On the other hand, certified defences provide provable guarantees on the model's robustness against specific attacks, offering higher security assurance. These strategies play a key role in mitigating malicious influence and preserving the integrity of the global model.

3) System-level defences: System-level defences fortify the overall federated learning infrastructure, addressing potential vulnerabilities in the communication and coordination processes. Blockchain technology can be leveraged to create a decentralised and immutable record of model updates, enhancing transparency and preventing tampering [1]. Learning model on the distributed ledger [7].

a) Blockchain-based validation: Hybrid approaches represent the most promising avenue for defending against federated learning attacks, combining multiple defence mechanisms to provide comprehensive protection. Integrating privacy-preserving techniques, such as Local Differential Privacy or Homomorphic Encryption, with robust aggregation methods can protect data privacy and mitigate the impact of malicious updates. This multi-layered approach creates a synergistic effect, enhancing the overall security and resilience of the federated learning system. For instance, incorporating differential privacy with blockchain can improve privacy management, allowing data owners to manage their privacy preferences [1]. This combination balances data privacy and model accuracy, making federated learning a viable option for sensitive applications [33].

b) Client verification protocols: Client verification protocols authenticate clients before they participate in training, mitigating the risk of unauthorised or malicious clients joining the federation. By verifying participants' identities on the blockchain, frameworks can prevent single points of failure and poisoning attacks [1]. Blockchain helps trace malicious attacks by storing model participation, user fingerprints, and other key information. A permissioned blockchain-based federated learning method can chain incremental updates to an anomaly detection machine.

Developing effective defence strategies against federated learning attacks is an ongoing area of research, with new techniques and approaches emerging regularly. These techniques help ensure federated learning systems’ privacy, security, and reliability, enabling their deployment in real-world applications where data privacy is paramount [34]. It is essential to address these challenges to unlock the full potential of federated learning and ensure its responsible and secure deployment across various domains [35].

As summarised in Table II, which supports a structured comparison of defences across attack vectors, validation metrics, and deployment constraints. Vectors, validation metrics, and deployment constraints.

TABLE II. COMPARATIVE SUMMARY OF FL DEFENCE STRATEGIES

Defence Type	Technique	Targeted Attacks	Strengths	Limitations
Data-Level	Local Differential Privacy	Gradient leakage, inference	Decentralised, lightweight	Privacy-utility trade-off
Data-Level	Homomorphic Encryption	Server-side attacks	Strong privacy, data never exposed	Heavy computation
Model-Level	Krum / Trimmed Mean	Model poisoning, backdoor	Effective against outliers	Not Sybil-proof
Model-Level	DP-RFA	Poisoning + privacy attacks	Formal robustness guarantees	Requires a trusted setup
System-Level	Secure Aggregation	Gradient leakage	Aggregated updates only	Poisoning still possible
System-Level	Blockchain Validation	Free-riders, Sybil, poisoning	Immutable audit trail, client trust	Computational & consensus latency

4) Hybrid and emerging strategies: Hybrid defence strategies are gaining prominence in federated learning as they comprehensively address the multifaceted security challenges in decentralised machine learning. These strategies combine multiple defence mechanisms to create a layered security architecture that is more resilient to attacks than any single defence alone. For instance, combining differential privacy with secure aggregation can provide robust protection against inference and model poisoning attacks [10]. Differential privacy adds noise to the model updates to protect individual client data, while secure aggregation ensures that the server only receives the aggregated model updates without revealing individual contributions [1]. Furthermore, hybrid defence strategies can adapt to different attack scenarios and data distributions, providing a more flexible and effective defence mechanism [1]. This method is more suitable for sophisticated real-world situations because it combines the advantages of various defence strategies. The resilience of federated learning systems can be significantly increased by combining the benefits of blockchain technology with differential privacy techniques.

Emerging trends in federated learning defence strategies focus on developing more adaptive and intelligent defence mechanisms that automatically detect and respond to evolving attack patterns. These strategies leverage anomaly detection, reinforcement learning, and meta-learning to identify and mitigate attacks in real-time. Anomaly detection algorithms can identify suspicious model updates or client behaviour that deviates from the norm. Reinforcement learning can be used to train a defence agent that learns to adjust defence parameters based on the observed attack patterns adaptively. Meta-learning can be used to train a model that can quickly adapt to new attack scenarios by learning from previous attacks. Through a case study on federated intrusion detection systems, the capabilities in detecting anomalies and securing critical infrastructure without exposing sensitive network data can be demonstrated [4]. Such cutting-edge strategies hold tremendous promise for improving federated learning's security and resilience in the face of constantly changing threats. To stay ahead of the attackers, companies should use the most advanced technology and work together in a way that allows them to contribute their insights securely.

Advanced encryption techniques, such as homomorphic encryption, safeguard data privacy during federated learning. Homomorphic encryption allows computations on encrypted data without decrypting, ensuring that sensitive information remains protected throughout the learning process [1]. This is particularly useful in federated learning scenarios where data is distributed across multiple devices and cannot be shared in plaintext. By leveraging homomorphic encryption, the server can aggregate client model updates without seeing the underlying data, preventing potential privacy breaches [15]. The advent of sophisticated encryption strategies guarantees that sensitive data is protected during federated learning activities, thus increasing confidence in the technology's use across various industries.

To enhance clarity and summarise the defensive landscape, Fig. 3 visually maps key attack vectors in Federated Learning (FL), the vulnerabilities they exploit, corresponding defence mechanisms, and their associated limitations. This diagram supports textual taxonomy and comparative tables by providing a holistic view of how threats are addressed and where residual risks remain.

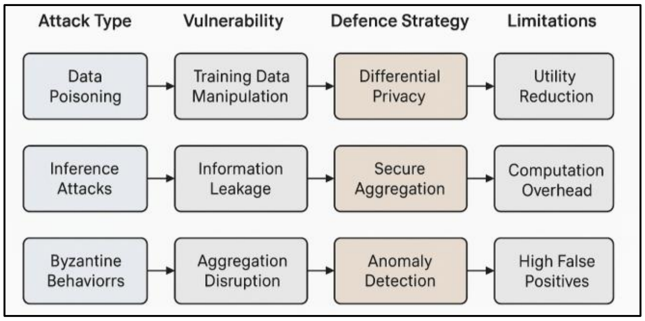


Fig. 3. Threat–defence–limitation mapping in federated learning security.

This flowchart visualises the relationship between common attack types in FL (e.g., model poisoning, inference attacks), the system vulnerabilities they exploit, corresponding mitigation strategies (e.g., differential privacy, secure aggregation), and each defence mechanism's inherent trade-offs or limitations.

V. Discussion

A. Key Findings

Current security frameworks often lack the sophistication to discern subtle anomalies indicative of malicious intent, highlighting the need for intelligent, context-aware systems [2]. Such systems should integrate real-time monitoring of client activities, employing statistical analysis and machine learning techniques to identify deviations from expected behaviour [3]. Furthermore, the dynamic adjustment of privacy budgets, aggregation rules, or trust scores based on the assessed risk level is crucial for maintaining data privacy while preserving model accuracy. This adaptive approach requires sophisticated algorithms that balance the trade-off between security and utility, ensuring that defensive measures do not unduly impede learning [36].

The structured use of figures and summary tables was intentional to enhance reader comprehension, particularly given the complex multi-dimensional nature of threats and defences in FL systems. These visual tools complement the narrative synthesis and provide a concise reference for practitioners evaluating real-world applicability.

Decentralised trust infrastructures, potentially leveraging blockchain and distributed ledger technologies, offer a promising avenue for enhancing accountability and security in large-scale FL systems [1]. The inherent immutability and transparency of blockchain provide a robust platform for validating client updates, auditing model evolution, and establishing tamper-proof records of all transactions [1]. By decentralising the trust mechanism, the reliance on a central aggregator is reduced, mitigating the risk of single points of failure and bolstering the overall resilience of the FL system [1]. Smart contracts can further automate and enforce security policies, ensuring participant compliance with predefined rules [1]. However, implementing blockchain-based solutions in FL environments presents challenges related to scalability and computational overhead [1]. Innovative approaches are needed to optimise the performance of blockchain networks in the context of FL, such as using consortium blockchains or sharding techniques [35].

Addressing the resource constraints of edge devices and IoT deployments is paramount for the widespread adoption of FL in diverse application domains. Lightweight cryptographic and differential privacy mechanisms must be developed to minimise the computational burden on these devices, enabling them to participate effectively in the learning process without compromising their performance or battery life. Techniques such as homomorphic encryption and secure multi-party computation can provide strong security guarantees, but often incur significant computational costs. Therefore, research is needed to develop more efficient variants of these techniques tailored to the specific requirements of FL. Moreover, novel approaches to differential privacy, such as local and distributed differential privacy, can offer enhanced privacy protection while minimising the impact on model accuracy. Integrating Blockchain, LDP, and FL enables stronger protection of trained models' integrity by preventing several attacks, such as poisoning attacks, background knowledge, and inference attacks [1].

The robustness of federated models can be significantly enhanced through adversarial training techniques that incorporate adversarial examples and simulated attacks into the training phase. Exposing the model to diverse potential threats can improve its ability to generalise to unseen data and resist malicious inputs. This approach generates adversarial examples to fool the model into making incorrect predictions. These examples can be crafted using various techniques, such as gradient-based methods and generative adversarial networks. Furthermore, simulated attacks can evaluate the model's resilience to adversarial behaviours, such as data poisoning and model poisoning. By incorporating these adversarial examples and simulated attacks into the training process, the model can learn to defend itself against a broader range of threats.

Establishing standardised evaluation frameworks is essential for facilitating reproducibility and consistent comparisons of different FL security techniques. Open-source benchmarks and evaluation metrics should be developed to provide a common platform for assessing the performance of FL systems under various attack scenarios. These benchmarks should include a diverse range of datasets, model architectures, and attack strategies, reflecting the complexity and heterogeneity of real-world FL deployments. Furthermore, the evaluation metrics should capture not only the accuracy and efficiency of the model but also its privacy and robustness. Differential privacy is often implemented to protect data privacy by obscuring model parameter data that each client transmits. This helps improve system security and preserve user data privacy [37] while enhancing model robustness and confidentiality [1]. However, current FL setups make it difficult for participants to verify the machine learning model's authenticity [19]. Standardised evaluation frameworks would enable researchers and practitioners to objectively compare security techniques and identify the most effective solutions for specific FL applications [11]. Recent studies demonstrate that an adversarial attacker may raise privacy concerns by launching inference attacks against other participants, even when the iterations are performed only a few times [30]. These attacks can be defended against by methods such as differential privacy. Compared to the taxonomy proposed by Nguyen et al. [11], which broadly classifies attacks into data and model categories, our model introduces a finer-grained surface-based classification. This facilitates more precise identification of attack entry points and informs layered defence design. Furthermore, our comparative defence summary (Table II) validates the practicality of proposed strategies by mapping them to real-world limitations such as computational cost and detection accuracy—current Research Gaps.

A significant gap in current research is the lack of effective defences against complex multi-client collusion attacks, where multiple malicious participants coordinate their actions to compromise the global model [2]. These attacks pose a substantial threat, as they can be challenging to detect and mitigate due to their distributed nature and the potential for subtle manipulation of the learning process. Further complicating the landscape is the trade-off between robustness and model performance, where defences designed to enhance robustness against adversarial attacks can inadvertently degrade the accuracy and generalisation capabilities of the model [19].

This trade-off requires careful consideration and the development of adaptive defence strategies that can dynamically adjust their strength based on the detected threat level and the desired level of model performance. The lack of comprehensive scalability testing in real-world federated learning scenarios also presents a challenge, as many existing defences have not been rigorously evaluated under large-scale deployments with heterogeneous data and diverse client capabilities. Addressing these gaps is essential for building trustworthy and reliable federated learning systems that can be confidently deployed in real-world applications.

Within the intricate domain of contemporary technological progression, the convergence of artificial intelligence (AI) and cloud computing has emerged as a transformative paradigm, revolutionising data processing, analysis, and dissemination across diverse sectors. According to study [38], federated learning decentralises AI model training by enabling computations directly on edge devices, enhancing model personalisation while preserving user privacy. As discussed by [39], cloud infrastructure provides scalable and elastic computational resources that support real-time AI workloads and large-scale model deployment. Additionally, the study [40] emphasises the role of this convergence in enabling privacy-aware and latency-efficient learning systems, particularly in applications involving sensitive or distributed datasets.

This convergence is not merely a confluence of technologies but rather a synergistic amalgamation that amplifies the capabilities of both domains, fostering innovation and driving efficiency gains across a spectrum of applications. The essence of this convergence lies in the ability of cloud computing to provide a scalable and cost-effective infrastructure for AI algorithms, enabling them to process vast datasets and execute complex computations with unprecedented speed and efficiency. As we navigate an era defined by the exponential growth of data, fueled by the proliferation of interconnected devices, the need for robust and scalable data processing solutions has become paramount. With its inherent elasticity and ability to dynamically allocate resources, cloud computing offers an ideal platform for AI algorithms to thrive, providing the necessary computational power and storage capacity to handle the ever-increasing volume of data.

The traditional paradigm of AI development, characterised by centralised data processing and algorithmic training, often encounters limitations regarding scalability, generalisation, and real-time responsiveness. To address these challenges, federated learning has emerged as a groundbreaking approach, enabling collaborative model training across decentralised devices while preserving data privacy [38]. Federated learning empowers AI models to learn from vast amounts of data distributed across numerous edge devices without centralising sensitive information. Federated learning, in essence, brings the learning process to the edge, directly onto devices, harnessing the computational capabilities of heterogeneous devices to enhance model quality [38], [39]. This distributed approach to machine learning addresses critical challenges associated with data privacy, security, and access in traditional centralised systems.

Federated learning operates on the principle of distributed model training, where each participating device locally updates a shared model using its data, subsequently transmitting the model updates to a central server for aggregation. This iterative process of local model training and global model aggregation ensures that the learned model generalises well to the diverse data distributions across different devices, without compromising the privacy of individual datasets [38]. Federated learning distinguishes itself through its ability to train models across decentralised devices, offering a solution to privacy concerns and scalability challenges inherent in centralised systems [38]. By keeping data localised on devices, federated learning minimises the risk of data breaches and misuse, while utilising vast datasets distributed across numerous sources [41]. Moreover, federated learning enhances model generalisation by leveraging the diversity of data across different devices, resulting in more robust and accurate AI models that can effectively address real-world problems.

Federated learning, a burgeoning technique in distributed machine learning, facilitates model training across many decentralised devices, obviating the need for centralised data collection. This approach is particularly salient in scenarios where data privacy is paramount, such as healthcare, finance, and IoT applications [38]. By enabling collaborative model training without direct access to sensitive data, federated learning upholds user privacy and complies with stringent data protection regulations [38]. The core tenet of federated learning lies in its ability to train models across diverse data distributions, enhancing generalisation and robustness. The decentralised nature of federated learning obviates the need to consolidate data in a central repository, which can be daunting due to logistical hurdles, regulatory constraints, and security considerations [38]. Synchronous federated learning requires all participants to complete local computations before updates are sent to a central server. In contrast, asynchronous federated learning allows the central server to immediately integrate updates from any ready participant [39].

To further enhance the privacy guarantees of federated learning, differential privacy techniques can be integrated to obfuscate individual contributions during model training [39]. Differential privacy introduces carefully calibrated noise to the model updates, ensuring that the presence or absence of any single data point has a negligible impact on the final model, thereby protecting the privacy of individual users. Federated learning not only addresses privacy concerns but also improves the efficiency and scalability of machine learning models. This approach is beneficial when data are located in multiple clinical systems or when learning from sensitive personal data.

The convergence of federated learning and differential privacy holds immense promise for many applications, particularly in domains where data privacy and security are paramount. Differential privacy enhances data privacy in federated learning by adding noise during data queries and model updates [39]. These areas include healthcare, finance, and the Internet of Things.

VI. FUTURE DIRECTIONS

Future research directions should prioritise the development of adaptive, context-aware defences that can dynamically adjust their parameters and strategies based on the specific characteristics of the data, the threat environment, and the available resources. By tailoring the defences to the particular context, it becomes possible to optimise their effectiveness and minimise their impact on model performance [3]. Combining secure aggregation with differential privacy dynamically represents another promising avenue for future research. Secure aggregation ensures that individual client updates are aggregated without revealing the underlying data, while differential privacy adds noise to the aggregated updates to protect against inference attacks [19]. Dynamically adjusting the parameters of these techniques based on the detected threat level and the desired privacy-utility trade-off can further enhance the security and performance of federated learning systems. Decentralised, blockchain-supported federated learning security offers a novel approach to improving trust and transparency in federated learning systems [7]. Attack detection and early warning systems in federated learning are crucial for proactively identifying and mitigating potential threats before they can cause significant damage. Integrating federated learning with blockchain technology allows for creating immutable audit trails and decentralised governance mechanisms, improving the learning process's accountability and trustworthiness [33]. A consortium blockchain-based federated learning framework enables decentralised, reliable, and secure federated learning without a centralised model coordinator [42]. Blockchain can address key challenges, propel the field forward, and potentially enhance data privacy and improve trust and security [35].

VII. Conclusion

Federated Learning (FL) represents a transformative approach in privacy-preserving machine learning, enabling decentralised model training across distributed clients without exposing raw data. However, the shift from centralised to federated paradigms introduces novel vulnerabilities that adversaries can exploit to compromise model integrity, data privacy, and system reliability.

Despite the comprehensive scope of this review, several limitations should be acknowledged. First, the analysis is constrained by the availability and maturity of published studies between 2020 and 2025, which may not fully capture the most recent developments or unpublished techniques in FL security. Second, while the taxonomy and comparative evaluation provide structured insights, they rely on secondary data rather than experimental benchmarking. Furthermore, the real-world effectiveness of many proposed defences—especially in large-scale, heterogeneous FL deployments—remains underexplored due to a lack of unified validation frameworks. These limitations highlight the need for ongoing empirical evaluation and longitudinal studies to assess defence robustness in dynamic adversarial settings.

To advance the field, future research must prioritise the development of adaptive, intelligent defence mechanisms capable of responding dynamically to evolving threats. Emphasis should also be placed on achieving a balance between security, privacy, and performance, particularly in large-scale, heterogeneous environments. Hybrid defence architectures, federated adversarial training, and lightweight cryptographic techniques hold promise in addressing these challenges. Ultimately, the continued evolution of FL security requires collaborative efforts across disciplines, integrating insights from cryptography, distributed systems, artificial intelligence, and regulatory policy. Ensuring robust, trustworthy FL systems will support privacy-aware innovations across critical sectors such as healthcare, finance, and autonomous systems.

Limitations: While this review provides a comprehensive classification of FL attack vectors and defence mechanisms, it is limited by the scope of current literature and the lack of unified benchmarking datasets. Furthermore, the practical deployment challenges—such as computation overhead, privacy-utility trade-offs, and client heterogeneity—require empirical validation in future work.

Acknowledgment

The author(s) sincerely thank Pejabat Penerbit, University Tun Hussein Onn Malaysia and the Faculty of Computer Science and Technology for their support throughout this research. Thanks to Dr. Shamsul Kamal bin Ahmad Khalid for the valuable guidance, feedback, and encouragement. The authors also acknowledge the use of publicly available resources and literature that have significantly contributed to the completion of this review. This work was conducted by the academic and ethical standards set forth by the institution. Communication of this research is made possible through monetary assistance by the University Tun Hussein Onn Malaysia and the UTHM Publisher’s Office via Publication Fund E15216.

References

[1] L. Javed, A. Anjum, B. M. Yakubu, M. Iqbal, S. A. Moqurrab, and G. Srivastava, “ShareChain: Blockchain-enabled model for sharing patient data using federated learning and differential privacy,” *Expert Systems*, vol. 40, no. 5, p. e13131, Aug. 2022.

[2] E. Hallaji, R. Razavi-Far, M. Saif, B. Wang, and Q. Yang, “Decentralised Federated Learning: A survey on security and privacy,” *IEEE Transactions on Big Data*, vol. 10, no. 2, pp. 194–213, April 2024.

[3] S. M. Mathews and S. A. Assefa, “Federated Learning: Balancing the thin line between data intelligence and privacy,” *arXiv*, arXiv:2204.13697, Jan. 2022.

[4] H. J. Huang, B. Iskandarov, M. Rahman, H. T. Otal, and M. A. Canbaz, “Federated learning in adversarial environments: Testbed design and poisoning resilience in cybersecurity,” *arXiv*, arXiv: 2409.09794, Sep. 2024.

[5] J. Li, X. Li, and C. Zhang, “Analysis on security and privacy-preserving in federated learning,” *Highlights in Science Engineering and Technology*, vol. 4, pp. 349–358, Jul. 2022.

[6] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, “How to backdoor federated learning,” in *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 108, pp. 2938–2948, 2020.

[7] D. Preuveneers, V. Rimmer, I. Tsingenopoulos, J. Spooren, W. Joosen, and E. Ilie-Zudor, “Chained anomaly detection models for federated learning: An intrusion detection case study,” *Applied Sciences*, vol. 8, no. 12, p. 2663, Dec. 2018.

[8] K. Aryal, M. Gupta, and M. Abdelsalam, “Analysis of Label-Flip poisoning attack on machine learning based malware detector,” in *2021 IEEE International Conference on Big Data (Big Data)*, pp. 4236–4245, Dec. 2022.

[9] S. Truex *et al.*, “A hybrid approach to privacy-preserving federated learning,” *Informatik-Spektrum*, vol. 42, no. 5, pp. 356–357, Aug. 2019.

[10] S. Fuladi *et al.*, “A reliable and privacy-preserved federated learning framework for real-time smoking prediction in healthcare,” *Frontiers in Computer Science*, vol. 6, p. 1494174, Jan. 2025.

[11] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. V. Poor, “Federated Learning for Internet of Things: A comprehensive

survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1622–1658, Jan. 2021.

[12] Y. Zhao *et al.*, "Privacy-preserving blockchain-based federated learning for IoT devices," *IEEE Internet of Things Journal*, vol. 8, no. 3, pp. 1817–1829, Aug. 2020.

[13] R. Xu, S. R. Pokhrel, Q. Lan, and G. Li, "Post quantum secure Blockchain-Based Federated Learning for Mobile Edge Computing," *arXiv,* arXiv:2302.13258, Jan. 2023.

[14] A. Joshi, "Federated Learning: Enhancing data privacy and security in machine learning through decentralised training paradigms," *Journal of Artificial Intelligence & Cloud Computing*, vol. 1, no. 1, pp. 1–5, Jan. 2022.

[15] K. Yadav and B. B. Gupta, "Clustering-based rewarding algorithm to detect adversaries in federated machine learning based IoT environment," in *2021 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 1–6, Jan. 2021.

[16] G. Sun, Y. Cong, J. Dong, Q. Wang, L. Lyu, and J. Liu, "Data poisoning attacks on federated machine learning," *IEEE Internet of Things Journal*, vol. 9, no. 13, pp. 11365–11375, Nov. 2021.

[17] M. Usama, J. Qadir, A. Al-Fuqaha, and M. Hamdi, "The adversarial machine learning conundrum: Can the insecurity of ML become the Achilles' heel of cognitive networks?" *IEEE Network*, vol. 34, no. 1, pp. 196–203, Oct. 2019.

[18] E. R. H. P. Isaac and J. Reno, "AI Product Security: A primer for developers," *arXiv,* arXiv:2304.11087, Jan. 2023.

[19] A. P. Kalapaaking, I. Khalil, and X. Yi, "Blockchain-Based Federated Learning with SMPC model verification against poisoning attack for healthcare systems," *IEEE Transactions on Emerging Topics in Computing*, vol. 12, no. 1, pp. 269–280, Apr. 2023.

[20] M. Balunović, D. Dimitrov I., N. Jovanović, and M. Vechev, "LAMP: Extracting text from gradients with language model priors," in *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*, Article 555, pp. 7641–7654, Jan. 2022.

[21] J. Sengupta, S. Ruj, and S. D. Bit, "SPRITE: a scalable privacy-preserving and verifiable collaborative learning for industrial IoT," in *2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, pp. 249–258, May 2022.

[22] C. Benzaid and T. Taleb, "AI for beyond 5G networks: A cyber-security defence or offence enabler?" *IEEE Network*, vol. 34, no. 6, pp. 140–147, Sep. 2020.

[23] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, May 2017.

[24] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 691–706, May 2019.

[25] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Model poisoning attacks in federated learning," in *Proceedings Workshop on Security in Machine Learning (secML) 32nd Conference on Neural Information Processing Systems (neurIPS)*, pp. 1–23, Dec. 2018.

[26] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems,* Article 1323, pp. 14774–14784, Jan. 2019.

[27] C. Fung, C. J. M. Yoon, and I. Beschastnikh, "Mitigating sybils in federated learning poisoning," *arXiv*, arXiv:1808.04866, Jan. 2018.

[28] J. Lin, M. Du, and J. Liu, "Free-riders in Federated Learning: Attacks and defences," arXiv.1911.12560, Jan. 2019.

[29] Z. Chen, H. Cui, E. Wu, and X. Yu, "Dynamic asynchronous anti-poisoning federated deep learning with blockchain-based reputation-aware solutions," *Sensors*, vol. 22, no. 2, p. 684, Jan. 2022.

[30] X. Zhang, H. Hou, Z. Fang, and Z. Wang, "Industrial internet federated learning driven by IoT equipment ID and blockchain," *Wireless Communications and Mobile Computing*, vol. 2021, no. 1, Jan. 2021.

[31] M. Veale, R. Binns, and L. Edwards, "Algorithms that remember: model inversion attacks and data protection law," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol.. 376, no. 2133, p. 20180083, Oct. 2018.

[32] D. Enthoven and Z. Al-Ars, "An overview of federated deep learning privacy attacks and defensive strategies," in *Federated Learning Systems. Studies in Computational Intelligence*, vol 965, M. H. u. Rehman and M. M. Gaber, Eds. Cham: Springer, 2021, pp. 173–196.

[33] Y. Li, C. Xia, W. Lin, and T. Wang, "PPBFL: a privacy-protected blockchain-based federated learning model," *arXiv,* arXiv:2401.01204, Jan. 2024.

[34] J. Wan, H. Xun, X. Zhang, J. Feng, and Z. Sun. "A privacy-preserving and correctness audit method in multi-party data sharing," in *Proceedings of the 2020 International Conference on Cyberspace Innovation of Advanced Technologies (CIAT 2020)*, pp. 414–419, Jan. 2021.

[35] M. M. Orabi, O. Emam, and H. Fahmy, "Adapting security and decentralised knowledge enhancement in federated learning using blockchain technology: Literature review," *Journal of Big Data*, vol. 12, p. 55, Mar. 2025.

[36] J. Mao, C. Cao, L. Wang, J. Ye, and W. Zhong, "Research on the security technology of Federated Learning privacy preserving," *Journal of Physics Conference Series*, vol. 1757, no. 1, p. 012192, Jan. 2021.

[37] L. Lyu, H. Yu, and Q. Yang, "Threats to Federated Learning: A survey," *arXiv,* arXiv:2003.02133, Jan. 2020.

[38] H. M. Asif, M. A. Karim, and F. Kausar, "Federated Learning and its applications for security and communication," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 8, Jan. 2022.

[39] F. Shan, S. Mao, Y. Lu, and S. Li, "Differential privacy Federated Learning: A comprehensive review," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 7, Jan. 2024.

[40] H. K. Shinwari and R. U. Amin, "Harnessing the power of Federated Learning: A systematic review of light weight deep learning protocols," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 1, Jan. 2025.

[41] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated Learning: challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, May 2020.

[42] Y. Qi, M. S. Hossain, J. Nie, and X. Li, "Privacy-preserving blockchain-based federated learning for traffic flow prediction," *Future Generation Computer Systems*, vol. 117, pp. 328–337, Dec. 2020.