2024 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress



# Privacy Threats and Countermeasures in Federated Learning for Internet of Things: A Systematic Review

Adel ElZemity and Budi Arief
School of Computing, University of Kent, Canterbury, United Kingdom
Email: ae455@kent.ac.uk, b.arief@kent.ac.uk

***Abstract*—Federated Learning (FL) in the Internet of Things (IoT) environments can enhance machine learning by utilising decentralised data, but at the same time, it might introduce significant privacy and security concerns due to the constrained nature of IoT devices. This represents a research challenge that we aim to address in this paper. We systematically analysed recent literature to identify privacy threats in FL within IoT environments, and evaluate the defensive measures that can be employed to mitigate these threats. Using a Systematic Literature Review (SLR) approach, we searched five publication databases (Scopus, IEEE Xplore, Wiley, ACM, and Science Direct), collating relevant papers published between 2017 and April 2024, a period which spans from the introduction of FL until now. Guided by the PRISMA protocol, we selected 49 papers to focus our systematic review on. We analysed these papers, paying special attention to the privacy threats and defensive measures – specifically within the context of IoT – using inclusion and exclusion criteria tailored to highlight recent advances and critical insights. We identified various privacy threats, including inference attacks, poisoning attacks, and eavesdropping, along with defensive measures such as Differential Privacy and Secure Multi-Party Computation. These defences were evaluated for their effectiveness in protecting privacy without compromising the functional integrity of FL in IoT settings. Our review underscores the necessity for robust and efficient privacy-preserving strategies tailored for IoT environments. Notably, there is a need for strategies against replay, evasion, and model stealing attacks. Exploring lightweight defensive measures and emerging technologies such as blockchain may help improve the privacy of FL in IoT, leading to the creation of FL models that can operate under variable network conditions.**

***Index Terms*—Federated Learning, Internet of Things, Privacy Threats, Defensive Measures, Systematic Literature Review.**

## I. Introduction

The Internet of Things (IoT) consists of interconnected devices that communicate and exchange data, enhancing real-time data collection and analysis across sectors [1]. This connectivity introduces privacy and security challenges, necessitating solutions such as Federated Learning (FL) that train models on decentralised data. FL improves traditional machine learning by addressing issues of accuracy, efficiency, and privacy [2]. However, FL in IoT faces challenges such as resource limitations, data heterogeneity, communication overheads, and privacy issues [3], [4]. These challenges are amplified by the limited computational power and energy resources of IoT devices, increasing potential risks to privacy [5]. Protecting FL data privacy on IoT devices is critical, and it requires robust defensive measures against threats such as inference attacks and data leakage. This review addresses the research gap by systematically analysing privacy threats and evaluating defensive measures within the IoT domain.

In order to address the identified research gap, this review systematically examines recent and pertinent literature. This review advances the knowledge regarding FL's applicability and privacy implications in IoT contexts by developing research questions centred on identifying privacy risks and defensive measures in such settings.

Enhancing the taxonomy for FL privacy in IoT, the systematic classification of existing publications offers insights into privacy properties, potential threats, and defence mechanisms. The importance of this review stems from its comprehensive evaluation of FL's privacy implications in the IoT domain. It is a valuable resource for practitioners and researchers who aim to understand and manage the complex interactions between privacy and AI technologies in the IoT environments, which typically have very limited resources. Other published literature reviews tend to focus only on specific facets of FL or IoT privacy. In comparison, this review is notable for its thorough analysis of FL privacy in the IoT environments.

**Contributions.** The key contributions of our paper are:

- Comprehensive systematic review and analysis of privacy threats in Federated Learning (FL) within the Internet of Things (IoT) environments, including inference attacks, poisoning attacks, and eavesdropping.
- Evaluation of various defensive measures such as Differential Privacy and Secure Multi-Party Computation, assessing their effectiveness in safeguarding privacy without undermining the operational integrity of FL in IoT.
- Identification of critical research gaps, particularly highlighting the need for robust strategies against replay, evasion, and model stealing attacks, to enhance the privacy posture of FL in IoT.

The rest of the paper is organised as follows. Section II introduces the important background of FL, especially in relation to privacy. Section III outlines our methodology, including a detailed explanation of the review's scope. Section IV presents our findings, while Section V discusses the implications of these findings. Finally, Section VI concludes our systematic review and suggests several areas for future research.

2836-3701/24/$31.00 ©2024 IEEE

DOI 10.1109/iThings-GreenCom-CPSCom-SmartData-Cybermatics62450.2024.00072

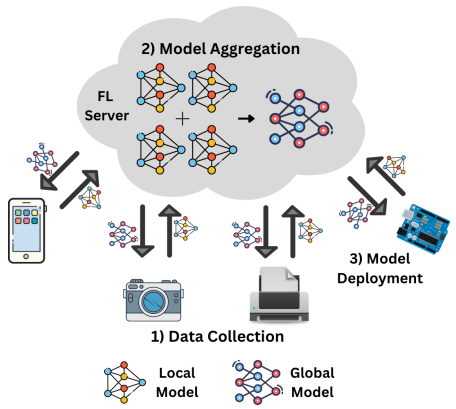


Fig. 1. A High-level Architecture of FL Process

## II. Background

Privacy in the IoT domain faces complex challenges due to the ubiquitous nature of the devices involved, and the vast amount of data they collect. Privacy threats are exacerbated by the diversity and scale of IoT environments, making effective privacy protections crucial, yet difficult to achieve. Various studies highlight the need for robust privacy-preserving measures tailored to IoT's unique constraints, such as device heterogeneity and extensive data generation [6]. Moreover, emerging solutions need to focus on enhancing privacy without compromising the functionality and scalability of IoT systems [7].

Users of wearable and smart IoT devices are more worried than ever about how the personal data they collect is used and shared across services. Because of its volume and diversity, pervasive user data are beneficial for state-of-the-art machine learning and deep learning algorithms, which are being used in these applications more and more. To facilitate learning over a distributed network without transferring the data from each device, Federated Averaging (FedAvg) [8] was presented as a foundational schema. FedAvg literature – and the broader FL literature – examine communication constraints and suggest enhanced learning frameworks, but do not investigate FL in severely constrained IoT environments with limited computing and storage capacity on the device [9].

FL is a distributed machine learning technique where clients train locally without sharing personal data with the server [2]. Devices iteratively update a shared global model by aggregating information from each client model. Figure 1 depicts the high-level architecture of the FL process, which usually consists of three phases [4]:

1) **Data Collection and Local Model Update:** The target application and task requirements are determined by the central server during the first phase. The server initialises a global model ($W_G^0$) and transmits it to the chosen local clients, called participants. Every participant uses their local data to create a model. Each client $k$ updates its model parameters ($W_t^i$) to find the optimal parameters that minimise the local loss function ($F_k(W_t^k)$) after receiving the global model ($W_G^t$) (where $t$ denotes the $t^{\text{th}}$ iteration). The local optimal models are then shared with the FL server.
2) **Global Aggregation:** The FL server aggregates the local models provided by the participants to create an updated global model ($W_G^{t+1}$).
3) **Model Deployment:** All of the new participants are given access to the most recent global model. Phases 2 and 3 are repeated until the central server reaches a convergence by minimising the global loss function ($F(W_t^G)$), which can be expressed as follows [10]: ($\min_w f(w) = \sum_{k=1}^{N} P_k F_k(w)$) where $N$ is the total number of devices available, $F_k(w)$ is the expected prediction loss on a sample input of the $k^{\text{th}}$ device on parameter $w$, $P_k(\geq 0)$ indicates the relative impact of each device $k$ while satisfying $\sum_k P_k = 1$, and each device $k$ has $n_k$ samples (where $n = \sum_k n_k$). $P_k = (n_k/n)$ is the expression that can be used to show the relative impact of each local device.

As this section has shown, current implementations of FL still face significant challenges, even though they offer promising paths for privacy-preserving ML, particularly within the IoT. These include protecting against sophisticated cyber threats that take advantage of the particular weaknesses of distributed architectures, managing resource constraints on IoT devices, and guaranteeing data privacy during model training. Significant gaps in privacy have come up from the inadequacies of existing strategies in effectively addressing these concerns, which our research attempts to address. The sections that follow will go into more detail about these issues and provide a new angle on privacy risks and the efficiency of modern defences in IoT environments.

## III. Methodology and Scope of Review

Our systematic literature review follows the PRISMA protocol [11] for a rigorous and transparent approach. We defined research questions to identify privacy threats and defensive measures in FL within IoT contexts. Using a comprehensive search strategy and strict inclusion and exclusion criteria, we systematically analysed recent advances in the field.

To analyse the literature and compare the proposed techniques systematically, we established the following research questions to guide our assessment:

- RQ1: What are the privacy threats present in federated learning within IoT environments?
- RQ2: What are the defensive measures to mitigate these risks without compromising data integrity, user privacy, and confidentiality?

### A. Paper Selection and Data Collection

We selected keywords aligned with our research questions, such as "Federated Learning", "FL", "Decentralised Machine Learning", "Privacy-Preserving Machine Learning", "Resource", "Energy", "Power", "Limited", "Constrain", "Privacy", and "Threat". The search query was: *("Federated*

*Learning" OR "FL" OR "Decentralised Machine Learning") AND ("IoT" OR "Internet of Things") AND ("Resource" OR "Energy" OR "Power") AND ("Limited" OR "Constrain") AND ("Privacy" OR "Threat").*

We used Scopus, IEEE Xplore, Wiley, ACM, and Science Direct to filter articles based on inclusion and exclusion criteria: articles from 2017 to April 2024, written in English, incorporating "federated learning" in the title, mentioning privacy aspects of FL or IoT, originating from reputable journals and conferences, and focusing on threats or defensive measures in IoT environments. Survey and review articles, and book chapters were excluded. Titles, abstracts, and full texts were evaluated against these criteria. Reference lists and citation tracking were used to ensure comprehensive coverage.

### B. Summary of Selected Papers

Following the PRISMA protocol, **980** papers were identified through database searching. Additionally, we employed citation chaining, identifying **30** additional papers through backward and forward snowballing from our core articles to ensure thorough coverage of the literature. After removing duplicates, **970** papers remained, but **715** of which were then excluded after initial title and abstract filtering. Subsequently, the full-text of the remaining articles were subjected to further screening based on the inclusion and exclusion criteria by the researchers involved. In the event of disagreement between the researchers, a third researcher served as a mediator to resolve the selection conflict. Finally, **49** articles were selected for subsequent analysis in this systematic literature review.

## IV. Results

Existing reviews offer insights into the challenges and limitations of FL in IoT. Hosseinzadeh et al. [12] discuss communication efficiency, resource allocation, and client selection in FL, focusing on its advantages without balancing potential drawbacks. Mothukuri et al. [13] highlight security threats such as communication bottlenecks and backdoor attacks in FL, but their review lacks comprehensive coverage of all privacy-related threats and limitations, particularly in resource-constrained IoT environments. Nguyen et al. [14] emphasise security and privacy in FL for IoT networks but do not provide a comprehensive risk analysis. Similarly, Khan et al. [15] discuss privacy challenges such as edge-cloud server inference and malicious user threats but lack an in-depth examination of IoT-specific vulnerabilities. Ferrag et al. [16] focus on attack vectors such as model poisoning and inference attacks but do not extensively evaluate privacy challenges in the IoT context.

These reviews highlight the need for a comprehensive understanding of privacy concerns in FL within resource-constrained IoT environments. FL in IoT faces various threats across its phases. Table I maps these threats to data collection, model aggregation, and model deployment phases, helping identify when specific threats are most likely to occur, which is crucial for developing targeted defences. For instance, inference attacks impact all phases, while model aggregation is most susceptible to various threats.

TABLE I
Privacy Threats to Federated Learning Process

| Threats | Data Collection | Model Aggregation | Model Deployment |
|---|---|---|---|
| Inference attacks | ✓ | ✓ | ✓ |
| Poisoning attacks | ✓ | ✓ | |
| Eavesdropping | | ✓ | |
| Sybil attacks | | ✓ | |
| Backdoor attacks | | ✓ | ✓ |
| Gradient Leakage | | ✓ | |
| Reconstruction | | | ✓ |

### A. Threats

A significant number of papers identify specific attacks that pose significant privacy risks and aim to validate their claims through proof of concept demonstrations. Subsequently, they propose various methods to defend against these identified threats. Table II groups the papers based on the seven privacy threats identified in the literature, and elaborated below.

*1) Inference Attacks:* **Membership inference attacks** pose a significant risk to users' privacy in resource-constrained IoT environments, determining if a specific data record was used in training a model. This can reveal sensitive information about the data subjects. Zhang et al. [17] discuss a membership inference attack using Generative Adversarial Networks (GANs) in FL, highlighting significant privacy leakages. This attack particularly affects FL models in IoT environments. Chen et al. [18] propose a novel user-level inference attack mechanism in FL, which is a critical concern for privacy in IoT implementations. Nguyen et al. [19] explore an active membership inference attack in FL under local differential privacy settings, demonstrating vulnerabilities in IoT data privacy. Zhao et al. [20] analyse membership inference attacks at a user level within a FL framework deployed in a wireless IoT network. **Model inversion attacks** use model outputs to infer sensitive features of the input data. Salim et al. [21] discuss FL's vulnerability to model inversion attacks in IoT-based social networks and propose a differential privacy-based framework to counter these threats. Xie et al. [22] explore the challenges of resisting model inversion and extraction attacks in IoT using FL, proposing a lightweight privacy protection protocol for edge computing. Zhang et al. [23] address privacy threats, including model inversion, using cryptographic methods within IoT-based healthcare systems employing FL. Zhou et al. [24] discuss protecting against model inversion attacks within a fog computing scenario using FL, focusing on the IoT context. **Property inference attacks** infer properties that hold over the entire training dataset or its subsets, which were not intended to be shared. A study by Shen et al. [25] explores property inference attacks in blockchain-assisted FL within intelligent edge computing, specifically targeting unintended property leakages from model updates. Wang et al. [26] present novel methodologies for carrying out a poisoning-assisted property inference attack that specifically targets FL systems, aiming to infer properties of training data that are unrelated to the learning objective.

TABLE II
REVIEWED PAPERS GROUPED BY PRIVACY THREATS

| Threat | Papers |
|---|---|
| Inference Attacks | [17], [18], [19], [20], [21], [22], [23], [24], [25], [26] |
| Poisoning Attacks | [27], [28], [29], [30] |
| Eavesdropping | [31], [32], [33] |
| Sybil Attacks | [34], [35], [36] |
| Backdoor Attacks | [37], [38], [39], [40] |
| Gradient Leakage | [41] |
| Reconstruction | [42], [43] |

*2) Poisoning Attacks:* Adversaries intentionally manipulate the training data or the model updates to corrupt the learning process, leading to incorrect model outputs or leaking specific data characteristics. Sun et al. [27] discuss data poisoning attacks in FL within IoT systems, highlighting the vulnerability of federated models to such attacks and proposing a novel systems-aware optimisation method to derive optimal attack strategies. Li et al. [28] explore adaptive poisoning attacks in the context of software-defined Industrial IoT (IIoT). They propose a framework that uses a tentacle distribution-based detection algorithm and a stochastic tentacle data exchanging protocol to minimise the impact of poisoned data. Zhang et al. [29] introduce PoisonGAN, a generative poisoning attack model for FL in edge computing. They demonstrate how this model can efficiently reduce attack assumptions and make attacks feasible in practice. Zhang et al. [30] propose RobustFL, a robust FL method for defending against poisoning attacks in IIoT, using an adversarial training framework. This method improves the resistance of the FL model to such attacks.

*3) Eavesdropping:* Unauthorised interception of data during transmission between IoT devices and the central server or amongst the devices themselves, potentially exposing sensitive data. Zheng et al. [31] explore FL as a method to preserve data training privacy from eavesdropping attacks in mobile-edge computing-based IoT. They propose a framework for optimising resource allocation to balance learning accuracy and energy consumption while protecting privacy. Ruzafa-Alcazar et al. [32] discuss the use of FL with differential privacy techniques to safeguard against intrusion and eavesdropping in IIoT environments. Matheu et al. [33] propose an FL approach to detect cyberattacks in IoT-enabled smart cities, integrating it with manufacturer usage descriptions to address eavesdropping and other attacks.

*4) Sybil Attacks:* Attackers create multiple fake identities to influence the training process maliciously or to gain a disproportionate influence over the model. Xiao et al. [34] propose a novel approach for Sybil-based collusion attacks in IIoT FL systems, demonstrating how malicious participants can manipulate model aggregation through Sybil identities. Jiang et al. [35] address Sybil attacks in the context of differential privacy-enhanced FL, proposing defence mechanisms that monitor training loss for anomalies to detect and mitigate such attacks. Fung et al. [36] introduce "FoolsGold", a defence against Sybil-based poisoning attacks in FL, which identifies malicious Sybils by examining the diversity of client updates.

*5) Backdoor Attacks:* Embedding hidden malicious functionality in the FL model, which can be activated to cause intended misbehaviour or to extract data. Hou et al. [37] discuss a defence mechanism against backdoor attacks in IIoT applications using FL, incorporating federated backdoor filters with explainable AI models. Ranjan et al. [38] propose graph-theoretic algorithms to identify and isolate backdoor attackers in FL systems, improving the robustness of the system. Yang et al. [39] explore clean-label poisoning attacks on FL in IoT environments, focusing on stealth and robustness of the attacks. Liu et al. [40] enhance the effectiveness of early-stage backdoor attacks in FL by leveraging information leakage about the whole population's data distribution.

*6) Gradient Leakage:* Even though raw data does not leave local devices, sharing model gradients can still leak information about the original data. Zhu et al. [41] focus on defending against inference attacks in FL within IoT, using parameter compression to mitigate the risk of gradient leakage.

*7) Reconstruction:* Attackers use the gradients or model parameters shared during FL updates to reconstruct the inputs used in training. Techniques might involve solving optimisation problems that aim to find data points that would produce similar gradients. Li et al. [42] discuss the vulnerabilities of FL models to gradient-based reconstruction attacks, particularly in complex IoT environments. They propose a defence strategy suitable for resource-constrained IoT devices, emphasising adaptive communication to ensure model security and decrease communication overhead. Na et al. [43] reevaluate the effectiveness of current privacy-preserving techniques against reconstruction attacks in FL, proposing a new lightweight solution called Fragmented Federated Learning (FFL).

### *B. Defensive Measures*

We also systematically evaluated the measures used within the literature to protect FL processes in IoT environments. Based on the specific privacy threats they address, we group seven defensive measures into three key categories: (i) Encryption and Obfuscation, (ii) Differential Privacy and Noise Injection, and (iii) Secure Multi-Party Computation and Anonymisation. These are detailed below. Table III provides a mapping of all defensive mechanisms against the types of threats they address. Related to this, Table IV provides a summary of quantitative metrics on various privacy threats and the effectiveness of corresponding defensive measures, showing the metrics before and after applying these defensive measures.

*1) Encryption and Obfuscation:* These measures encrypt or alter data to prevent direct access or interpretation by unauthorised parties. **Gradient obfuscation** conceals sensitive data by altering gradient samples within FL processes to prevent direct inference attacks without sacrificing model performance. It protects data by making it difficult to reverse-engineer or identify sensitive information from gradients [44]. Yue et al. [63] present an analysis of how gradient obfuscation, including quantisation and perturbation, provides a false sense of security in FL by demonstrating the feasibility of data reconstruction attacks despite these privacy measures. Fu et

TABLE III
MAPPING OF PRIVACY DEFENSIVE MEASURES AGAINST PRIVACY THREATS

| | Inference Attacks | Poisoning Attacks | Eavesdropping | Sybil Attacks | Backdoor Attacks | Gradient Leakage | Reconstruction |
|---|---|---|---|---|---|---|---|
| **Gradient Obfuscation** | [44] | | | | | | |
| **Parameter Compression** | [41], [45] | | | | | | |
| **Compressed Sensing** | [46] | | | | [47] | | |
| **Differential Privacy** | [48], [49], [32], [50], [51] | [48], [49], [32], [50], [51] | | | | | |
| **Decentralised Perturbation** | [52], [53], [54], [55], [56] | | | [53] | | | |
| **Secure Multi-Party Computation** | [57], [58], [59], [24] | [57], [58], [59], [24] | [60] | | [58] | [61] | |
| **Anonymisation and Siamese Networks** | [62] | | | [62] | | | [63] |

TABLE IV
SUMMARY OF QUANTITATIVE METRICS ON PRIVACY THREATS AND DEFENSIVE MEASURES IN FEDERATED LEARNING

| Paper | Attack Type | Defence Measure | Metrics Before Defence | Metrics After Defence |
|---|---|---|---|---|
| Zhu et al. (2023) [41] | GAN-Based Privacy Inference | FLPC[a] | **Accuracy:** 0.9591 → 0.9507 (Baseline) Decrease: 0.84% | **Accuracy:** 0.9565 → 0.9557 (FLPC, Comlevel = 0.001) Decrease: 0.08% |
| Song et al. (2020) [62] | User-Level Privacy Attack | mGAN-AI[b] | **Accuracy (MNIST Training):** 0.9438<br>**Accuracy (MNIST Testing):** 0.9247<br>**Accuracy (AT&T Training):** 0.9435<br>**Accuracy (AT&T Testing):** 0.9267 | **Passive mGAN-AI Inception Score:** 1.42±0.02<br>**Active mGAN-AI Inception Score:** 1.61±0.05<br>**Passive mGAN-AI Accuracy:** Similar to baseline<br>**Active mGAN-AI Accuracy:** Slightly lower |
| Liu et al. (2021) [64] | Label-Flipping | PEFL[c] | **Attack Success Rate:** 0.001 → 1<br>**True Positive Rate:** 0.98 → 0<br>**Non-Source Class Accuracy:** 0.95 → 0.51 | **Attack Success Rate:** 0 → 0.03<br>**True Positive Rate:** 0.95 → 0.88<br>**Non-Source Class Accuracy:** 0.97 → 0.76 |
| | Backdoor | PEFL[c] | **Attack Success Rate:** 0.001 → 1<br>**True Positive Rate:** 0.98 → 0<br>**Non-Source Class Accuracy:** 0.95 → 0.64 | **Attack Success Rate:** 0 → 0.04<br>**True Positive Rate:** 0.95 → 0.88<br>**Non-Source Class Accuracy:** 0.97 → 0.76 |
| Liu et al. (2023) [65] | Gradient-Based Data Reconstruction | Privacy-Encoded FL | **PSNR:** 28.99 → 0.6838 (Baseline)<br>**Test Accuracy:** 89.76% | **PSNR:** 29.45 → 3.54<br>**Test Accuracy:** 87.78% → 89.86% |
| Jiang et al. (2020) [66] | Sybil Attacks | DP[d] and Anomaly Detection | **CNN Error Rate:** 0.03 → 0.14<br>**MLP Error Rate:** 0.59 → 0.63 | **CNN Error Rate:** 0.03 → 0.03<br>**MLP Error Rate:** 0.59 → 0.59 |
| Miao et al. (2022) [67] | Backdoor Attacks | CND[e] with DP[d] | **CIFAR-10 Accuracy:** 88% → 84%<br>**EMNIST Accuracy:** 99% → 90%<br>**CIFAR-10 Attack Success:** 0% → 80%<br>**EMNIST Attack Success:** 0% → 100% | **CIFAR-10 Accuracy:** 82% → 81%<br>**EMNIST Accuracy:** 95% → 75%<br>**CIFAR-10 Attack Success:** 0% → 3%<br>**EMNIST Attack Success:** 0% → 5% |
| Li et al. (2023) [68] | Gradient-Based Inference, Byzantine | PBA[f] | **Global Accuracy:** 88% | **GA[g] (f=2):** ∼87%; **GA[g] (f=6):** ∼83%<br>**LFA[h] (f=2):** ∼85%; **LFA[h] (f=6):** ∼60%<br>**Running Time:** 9.391 s |
| Asad et al. (2020) [69] | DP[d], Homomorphic Encryption, Backdoor | DP[d], HE[i], Secure Aggregation | **DP Accuracy (PB[j]=0.1):** 70%<br>**DP Accuracy (PB[j]=0.5):** 55%<br>**DP Accuracy (PB[j]=1.0):** 40%<br>**DP Accuracy (PB[j]=2.0):** 30%<br>**HE Accuracy (SP[k]=32):** 85%<br>**HE Accuracy (SP[k]=64):** 75%<br>**HE Accuracy (SP[k]=96):** 65%<br>**HE Accuracy (SP[k]=128):** 60% | **Secure Aggregation:** ∼80%<br>**Partial Secure Aggregation:** ∼75%<br>**Backdoor (5 rounds):** ∼0%<br>**Backdoor (10 rounds):** ∼10%<br>**Backdoor (60 rounds):** ∼50%<br>**Backdoor (80 rounds):** ∼60% |

[a]FLPC: Federated Learning Parameter Compression, [b]mGAN-AI: Generative Adversarial Network for Adversarial Inference,
[c]PEFL: Privacy-Enhanced Federated Learning, [d]DP: Differential Privacy, [e]CND: Clip Norm Decay, [f]PBA: Privacy Robust Aggregation,
[g]GA: Gaussian Attack, [h]LFA: Label Flipping Attack, [j]PB: Privacy Budget, [k]SP: Security Parameter, [i]HE: Homomorphic Encryption

al. [61] propose VFL, a verifiable FL framework for big data in the IIoT, enhancing privacy through Lagrange interpolation and blinding technology to safeguard gradient privacy. Gade et al. [60] introduce a privacy-preserving distributed learning method using obfuscated stochastic gradients to enhance privacy against honest-but-curious adversaries in an FL setup.

**Parameter compression** reduces detailed information sharing in FL, preventing attackers from reconstructing private data from model parameters. Zhu et al. [41] address privacy inference attacks in FL for IoT via parameter compression, preserving privacy and model accuracy. Chen et al. [45] discuss an adaptive federated optimisation algorithm that balances computation, communication, and precision in IoT environments using parameter compression.

**Compressed sensing as encryption** uses compressed sensing as a dual method for data compression and encryption,

safeguarding gradients and labels against inference attacks. Miao et al. [46] design an efficient privacy-preserving FL scheme based on compressed sensing, which serves both as a compression and encryption method. This approach ensures that gradients do not disclose private information, making it suitable for IoT scenarios. Li et al. [47] propose FL algorithms based on compressed sensing, enhancing communication efficiency in IoT environments. These algorithms allow for model updates between IoT clients and a central server, improving performance over traditional methods.

*2) Differential Privacy and Noise Injection:* These measures use noise to mask data, adhering to differential privacy standards to ensure individual data points remain indiscernible. **Differential privacy-injected noise** incorporates artificial noise based on differential privacy to protect local parameters, balancing privacy with model accuracy. Shen et al. [48] have developed a performance-enhanced DP-based FL algorithm for IoT, introducing a classifier-perturbation regularisation method to improve the robustness of the trained model against DP-injected noise. Cui et al. [49] have designed an improved differentially private FL system for anomaly detection in IoT infrastructures, optimising data utility throughout the training process. Ruzafa-Alcazar et al. [32] provide a comprehensive evaluation of differential privacy techniques in the training of an FL-enabled intrusion detection system for IIoT. Yin et al. [50] propose a new hybrid privacy-preserving method for federal learning that employs sparse differential gradient to improve transmission efficiency in social IoT scenarios. He et al. [51] introduce adaptive local differential privacy mechanisms in FL for heterogeneous IoT data, focusing on balancing the trade-off between privacy and utility.While differential privacy techniques have shown promising results in controlled environments, their practical application in real-world IoT scenarios often faces challenges such as maintaining utility while ensuring privacy. Recent studies [70] and [49] have highlighted the need for adaptive mechanisms that balance this trade-off effectively.

**Decentralised perturbation techniques** distribute the task of injecting noise across federated nodes to protect privacy, enhancing the scalability and robustness of privacy measures [52]. Mothukuri et al. [53] propose an FL-based anomaly detection for IoT security, utilising decentralised data processing to enhance privacy and model accuracy. Mantey et al. [54] introduce a Secure Recommendation and Training Technique (SERTT) that leverages both FL and blockchain for privacy-preserved data management in the Internet of Medical Things (IoMT). Alotaibi [55] proposes a biserial correlative Miyaguchi–Preneel blockchain-based Ruzicka-indexed deep multi-layer perceptive learning (BCMPB-RIDMPL) method for improving malware detection in IoMT. Alamleh et al. [56] have developed a standardisation and bench-marking framework for machine-learning based intrusion detection systems using FL in IoMT environments.

*3) Secure Multi-party Computation and Anonymisation:* These measures focus on collaborative techniques that enable secure and private computations among multiple parties without revealing individual data inputs. **Secure multiparty computing** employs secure multiparty computing to enable private information exchange between FL participants, enhancing data privacy through complex protocols [57]. Liu et al. [58] propose a privacy-preserving FL scheme for Internet of Medical Things, which includes secure authentication and aggregation to protect data during model training. Lu et al. [59] have designed a blockchain and FL-based architecture for secure data sharing in IIoT, maintaining data privacy by sharing the data model instead of the actual data. Zhou et al. [24] present an FL scheme in fog computing that enhances privacy and efficiency by integrating secure multi-party computing techniques. **Anonymisation and Siamese networks** use anonymisation strategies along with advanced network architectures to protect client identity and data during the training process, making re-identification challenging. Song et al. [62] propose a framework incorporating GANs with a multi-task discriminator to analyse user-level privacy leakage in FL, developing a siamese network to re-identify anonymised updates and measuring the similarity of representatives effectively in IoT scenarios.

## V. Discussion

This comprehensive review systematically explores the landscape of privacy threats in FL within IoT environments and evaluates the effectiveness of various defensive measures. We identify common threats such as inference and poisoning attacks and discuss lesser-covered threats such as Sybil and backdoor attacks in the context of IoT devices. The defensive measure of Differential Privacy is prominently featured, highlighting its critical role across various phases of the FL process. Additionally, we extend the current understanding by contrasting our findings with previous reviews which often focus on a narrower range of threats or do not address the unique challenges posed by the IoT environment. Notable papers by Mothukuri et al. [13] and Ferrag et al. [16] primarily highlight security aspects without delving into the nuanced impacts of these environments on privacy and security strategies.

### A. Current Landscape of Threats and Defences

In the reviewed papers, inference attacks are extensively studied, while gradient leakage and reconstruction are notably less addressed, indicating significant research gaps within federated learning in IoT (see Table II). Replay, evasion, and model stealing attacks also emerge as critical yet under-researched threats. The lack of focus on these vulnerabilities is concerning due to their potential to disrupt federated models' integrity and effectiveness. We emphasise the need for IoT system designers to incorporate robust defences early in the design phase. Defensive strategies such as differential privacy and secure multi-party computation, though promising, must be tailored to IoT constraints such as limited computational power and energy resources. Addressing these gaps is crucial for safeguarding systems against sophisticated cyber threats, ensuring reliability and trustworthiness in applications such as autonomous driving and medical diagnostics. In practical

applications, secure multiparty computation in IoT has shown varying success. For instance, Liu et al. [58] demonstrate its feasibility in medical IoT systems but noted significant computational overhead. Similarly, Lu et al. [59] highlight integrating blockchain with FL to enhance data integrity and privacy in IIoT, though it requires substantial computational resources that may not be available in all IoT settings.

### B. Advances and Innovations

There is a critical need for developing lightweight privacy-preserving algorithms optimised for the IoT contexts. Our findings suggest that while differential privacy offers a balanced approach to privacy and efficiency, secure multi-party computation and other high-cost measures may require significant optimisation to be feasible in IoT contexts. Furthermore, emerging technologies like blockchain could offer scalable solutions but need thorough evaluation in real-world IoT settings to determine their operational viability. Additionally, empirical studies assessing the real-world applicability and resilience of proposed defensive mechanisms under varied IoT conditions and attack scenarios would greatly benefit the field. The expansion of IoT devices in sensitive areas (such as healthcare and smart cities) underscores the urgency of addressing privacy in FL. As IoT devices become more pervasive, ensuring the privacy and security of FL systems will be crucial in maintaining user trust and regulatory compliance.

### C. Limitations

There are several limitations to our research, starting with the exclusion of gray literature and non-English publications, which might contain relevant data and insights. Additionally, the rapid evolution of both threats and technologies in this domain means that our findings might require continuous updates to remain relevant.

## VI. Conclusion

In conclusion, this systematic review critically assesses privacy threats and defensive measures in Federated Learning (FL) within IoT environments. Analysing literature from 2017 to April 2024, we identified persistent challenges such as inference and poisoning attacks that compromise FL model robustness. The review highlights the need for innovative defensive strategies tailored to IoT constraints, balancing computational efficiency with privacy safeguards. Our findings emphasise integrating advanced measures such as Differential Privacy and Secure Multi-Party Computation to mitigate privacy risks. However, under-explored threats – such as replay, evasion, and model stealing attacks – pose significant risks, necessitating further research. Practical implementation of defensive measures in IoT settings reveals some potential but also exposes gaps requiring further research. Effective deployment demands addressing computational constraints and ensuring robust performance under variable network conditions, as shown in recent studies [58], [70]. Future research should prioritise developing lightweight, optimised privacy-preserving algorithms and explore emerging technologies such as blockchain to enhance FL privacy. Additionally, developing FL models that operate under variable network conditions while maintaining edge device privacy is crucial. Further exploration into FL adaptations for edge computing to reduce latency and improve response times in privacy-critical applications is also essential.


## Acknowledgements

This work was supported by the funding received from the UK EPSRC project EP/X036707/1 on Countering HArms caused by Ransomware in the Internet Of Things (CHARIOT). The authors would also like to thank the anonymous reviewers for their constructive feedback.


## References


[1] S. Madakam, R. Ramaswamy, and S. Tripathi, "Internet of Things (IoT): A Literature Review," *J. of Comp. Chemistry*, vol. 3, pp. 164–173, 2015.

[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*, pp. 1273–1282, 2017.

[3] M. Aledhari, R. Razzak, R. M. Parizi, and F. Saeed, "Federated learning: A survey on enabling technologies, protocols, and applications," *IEEE Access*, vol. 8, pp. 140699–140725, 2020.

[4] A. Imteaj, U. Thakker, S. Wang, J. Li, and M. H. Amini, "A survey on federated learning for resource-constrained iot devices," *IEEE IoT J.*, vol. 9, no. 1, pp. 1–24, 2022.

[5] K. Kopparapu, E. Lin, J. G. Breslin, and B. Sudharsan, "Tinyfedtl: Federated transfer learning on ubiquitous tiny iot devices," in *2022 IEEE Int'l Conf. on Pervasive Computing and Commun. Workshops and other Affiliated Events (PerCom Workshops)*, pp. 79–81, 2022.

[6] L. Tawalbeh, F. Muheidat, M. Tawalbeh, and M. Quwaider, "Iot privacy and security: Challenges and solutions," *Applied Sciences*, 2020.

[7] Y. Qu, S. Yu, W. Zhou, S. Peng, G. Wang, and K. Xiao, "Privacy of things: Emerging challenges and opportunities in wireless internet of things," *IEEE Wireless Commun.*, vol. 25, pp. 91–97, 2018.

[8] T. Sun, D. Li, and B. Wang, "Decentralized federated averaging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, pp. 4289–4301, 2021.

[9] S. K. Lo, Q. Lu, C. Wang, H.-Y. Paik, and L. Zhu, "A systematic literature review on federated machine learning: From a software engineering perspective," *ACM Computing Surveys*, vol. 54, no. 5, pp. 1–39, 2021.

[10] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proc. of Machine learning and systems*, vol. 2, pp. 429–450, 2020.

[11] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and t. PRISMA Group*, "Preferred reporting items for systematic reviews and meta-analyses: the prisma statement," *Ann. Intern. Med.*, vol. 151, no. 4, pp. 264–269, 2009.

[12] M. Hosseinzadeh, A. Hemmati, *et al.*, "Federated learning-based iot: A systematic literature review," *Int. J. of Comm. Systems*, vol. 35, 2022.

[13] V. Mothukuri, R. Parizi, S. Pouriyeh, Y. ping Huang, A. Dehghantanha, and G. Srivastava, "A survey on security and privacy of federated learning," *Future Gener. Comput. Syst.*, vol. 115, pp. 619–640, 2021.

[14] D. C. Nguyen, M. Ding, *et al.*, "Federated learning for internet of things: A comprehensive survey," *IEEE Commun. Surveys & Tutorials*, vol. 23, pp. 1622–1658, 2021.

[15] L. U. Khan, W. Saad, *et al.*, "Federated learning for internet of things: Recent advances, taxonomy, and open challenges," *IEEE Commun. Surveys & Tutorials*, vol. 23, pp. 1759–1799, 2020.

[16] M. A. Ferrag, O. Friha, *et al.*, "Federated Deep Learning for Cyber Security in the Internet of Things: Concepts, Applications, and Experimental Analysis," *IEEE Access*, vol. 9, pp. 138509–138542, 2021.

[17] J. Zhang, J. Zhang, J. Chen, and S. Yu, "Gan enhanced membership inference: A passive local attack in federated learning," *ICC 2020 - 2020 IEEE Int'l Conf. on Commun. (ICC)*, pp. 1–6, 2020.

[18] J. Chen, J. Zhang, *et al.*, "Beyond model-level membership privacy leakage: an adversarial approach in federated learning," *2020 29th Int. Conf. on Computer Commun. and Networks (ICCCN)*, pp. 1–9, 2020.

[19] T. D. T. Nguyen, P. Lai, *et al.*, "Active membership inference attack under local differential privacy in federated learning," *ArXiv*, vol. abs/2302.12685, 2023.

[20] Y. Zhao, J. Chen, J. Zhang, Z. Yang, H. Tu, H. Han, K. Zhu, and B. Chen, "User-Level Membership Inference for Federated Learning in Wireless Network Environment," *Wirel. Commun. Mob. Comput.*, 2021.

[21] S. Salim, N. Moustafa, B. Turnbull, and I. Razzak, "Perturbation-enabled deep federated learning for preserving internet of things-based social networks," *ACM Trans. on Multimedia Computing, Commun., and Applications (TOMM)*, vol. 18, pp. 1 – 19, 2022.

[22] H. Xie, Y. Guo, K. He, Y. Song, and Y. Liu, "Privacy-preserving edge intelligent computing based on federated learning," *2021 7th Int. Conf. on Computer and Commun. (ICCC)*, pp. 1371–1377, 2021.

[23] L. Zhang, J. Xu, *et al.*, "Homomorphic Encryption-Based Privacy-Preserving Federated Learning in IoT-Enabled Healthcare System," *IEEE Trans. Netw. Sci. Eng.*, vol. 10, pp. 2864–2880, 2023.

[24] C. Zhou, A. Fu, *et al.*, "Privacy-Preserving Federated Learning in Fog Computing," *IEEE IoT J.*, vol. 7, pp. 10782–10793, 2020.

[25] M. Shen, H. Wang, *et al.*, "Exploiting unintended property leakage in blockchain-assisted federated learning for intelligent edge computing," *IEEE IoT J.*, vol. 8, pp. 2265–2275, 2021.

[26] Z. Wang, Y. Huang, M. Song, L. Wu, F. Xue, and K. Ren, "Poisoning-assisted property inference attack against federated learning," *IEEE Trans. on Dep. and Sec. Computing*, vol. 20, pp. 3328–3340, 2023.

[27] G. Sun, Y. Cong, J. Dong, Q. Wang, and J. Liu, "Data poisoning attacks on federated ml," *IEEE IoT J.*, vol. 9, pp. 11365–11375, 2020.

[28] G. Li, J. Wu, S. Li, W. Yang, and C. Li, "Multitentacle fl over software-defined industrial internet of things against adaptive poisoning attacks," *IEEE Trans. Ind. Informat.*, vol. 19, pp. 1260–1269, 2023.

[29] J. Zhang, B. Chen, X. Cheng, H. T. T. Binh, and S. Yu, "Poisongan: Generative poisoning attacks against federated learning in edge computing systems," *IEEE IoT J.*, vol. 8, pp. 3310–3322, 2021.

[30] J. Zhang, C. Ge, F. Hu, and B. Chen, "Robustfl: Robust federated learning against poisoning attacks in industrial iot systems," *IEEE Trans. Ind. Informat.*, vol. 18, pp. 6388–6397, 2022.

[31] J. Zheng, K. Li, *et al.*, "Exploring Deep-Reinforcement-Learning-Assisted Federated Learning for Online Resource Allocation in Privacy-Preserving EdgeIoT," *IEEE IoT J.*, vol. 9, pp. 21099–21110, 2022.

[32] P. Ruzafa-Alcazar, P. Fernandez-Saura, *et al.*, "Intrusion Detection Based on Privacy-Preserving Federated Learning for the Industrial IoT," *IEEE Trans. Ind. Informat.*, vol. 19, pp. 1145–1154, 2023.

[33] S. N. Matheu, E. Mármol, J. L. Hernández-Ramos, A. Skarmeta, and G. Baldini, "Federated Cyberattack Detection for Internet of Things-Enabled Smart Cities," *Computer*, vol. 55, pp. 65–73, 2022.

[34] X. Xiao, Z. Tang, C. Li, B. Xiao, and K. Li, "Sca: Sybil-based collusion attacks of iiot data poisoning in federated learning," *IEEE Trans. Ind. Informat.*, vol. 19, pp. 2608–2618, 2023.

[35] Y. Jiang, Y. Li, *et al.*, "Sybil Attacks and Defense on Differential Privacy based Federated Learning," *IEEE 20th Int. Conf. on Trust, Security and Privacy in Computing and Commun. (TrustCom)*, pp. 355–362, 2021.

[36] C. Fung, C. J. M. Yoon, and I. Beschastnikh, "Mitigating sybils in federated learning poisoning," *ArXiv*, vol. abs/1808.04866, 2018.

[37] B. Hou, J. Gao, X. Guo, T. Baker, Y. Zhang, Y. Wen, and Z. Liu, "Mitigating the backdoor attack by federated filters for industrial iot applications," *IEEE Trans. Ind. Informat.*, vol. 18, pp. 3562–3571, 2022.

[38] P. Ranjan, A. Gupta, F. Coró, and S. K. Das, "Robust federated learning against backdoor attackers," *IEEE INFOCOM 2023 - IEEE Conf. on Computer Commun. Workshops*, pp. 1–6, 2023.

[39] J. Yang, J. Zheng, *et al.*, "Clean-label poisoning attacks on federated learning for iot," *Expert Systems*, vol. 40, 2022.

[40] T. Liu, X. Hu, and T. Shu, "Facilitating early-stage backdoor attacks in federated learning with whole population distribution inference," *IEEE IoT J.*, vol. 10, pp. 10385–10399, 2023.

[41] Y. Zhu, H. Cao, *et al.*, "Defending Privacy Inference Attacks to Federated Learning for Intelligent IoT with Parameter Compression," *Security and Comm. Networks*, 2023.

[42] Y. Li, Y. Li, H. Xu, and S. Ren, "An adaptive communication-efficient federated learning to resist gradient-based reconstruction attacks," *Secur. Commun. Networks*, vol. 2021, pp. 9919030:1–9919030:16, 2021.

[43] S. Na, H. Hong, J. Kim, and S. Shin, "Closing the loophole: Rethinking reconstruction attacks in federated learning from a privacy standpoint," *Proc. of the 38th Annual Computer Security Applications Conf.*, 2022.

[44] J. Wu, M. Hayat, M. Zhou, and M. Harandi, "Defense against privacy leakage in federated learning," *ArXiv*, vol. abs/2209.05724, 2022.

[45] Z. Chen, H. Cui, E. Wu, and X. Yu, "Computation and communication efficient adaptive federated optimization of federated learning for internet of things," *Electronics*, 2023.

[46] Y. Miao and S. Chen, "Efficient privacy-preserving federated learning against inference attacks for iot," *2023 IEEE Wireless Commun. and Networking Conf. (WCNC)*, pp. 1–6, 2023.

[47] C. Li, G. Li, *et al.*, "Communication-efficient federated learning based on compressed sensing," *IEEE IoT J.*, vol. 8, pp. 15531–15541, 2021.

[48] X. Shen, Y. Liu, and Z. Zhang, "Performance-enhanced federated learning with differential privacy for internet of things," *IEEE IoT J.*, vol. 9, pp. 24079–24094, 2022.

[49] L. Cui, Y. Qu, *et al.*, "Security and privacy-enhanced federated learning for anomaly detection in iot infrastructures," *IEEE Trans. Ind. Informat.*, vol. 18, pp. 3492–3500, 2021.

[50] L. Yin, J. Feng, H. Xun, Z. Sun, and X. Cheng, "A Privacy-Preserving Federated Learning for Multiparty Data Sharing in Social IoTs," *IEEE Trans. Netw. Sci. Eng.*, vol. 8, pp. 2706–2718, 2021.

[51] Z. He, L. Wang, and Z. Cai, "Clustered federated learning with adaptive local differential privacy on heterogeneous iot data," *IEEE IoT J.*, vol. 11, pp. 137–146, 2024.

[52] P. C. M. Arachchige, P. Bertók, *et al.*, "Privacy preserving distributed machine learning with federated learning," *ArXiv*, vol. abs/2004.12108, 2020.

[53] V. Mothukuri, P. Khare, R. Parizi, S. Pouriyeh, A. Dehghantanha, and G. Srivastava, "Federated-learning-based anomaly detection for iot security attacks," *IEEE IoT J.*, vol. 9, pp. 2545–2554, 2021.

[54] E. A. Mantey, C. Zhou, J. H. Anajemba, Y. Hamid, and J. K. Arthur, "Blockchain-enabled technique for privacy-preserved medical recommender system," *IEEE Access*, vol. 11, pp. 40944–40953, 2023.

[55] A. Alotaibi, "Biserial miyaguchi–preneel blockchain-based ruzicka-indexed deep perceptive learning for malware detection in iomt," *Sensors (Basel, Switzerland)*, vol. 21, 2021.

[56] A. Alamleh, O. Albahri, *et al.*, "Federated Learning for IoMT Applications: A Standardization and Benchmarking Framework of Intrusion Detection Systems," *IEEE J. of Biomedical and Health Informatics*, vol. 27, pp. 878–887, 2022.

[57] Y. Li, Y. Zhou, A. Jolfaei, D. Yu, G. Xu, and X. Zheng, "Privacy-preserving federated learning framework based on chained secure multiparty computing," *IEEE IoT J.*, vol. 8, pp. 6178–6186, 2020.

[58] J. Liu, J. Zhang, M. Jan, R. Sun, L. Liu, S. Verma, and P. Chatterjee, "A comprehensive privacy-preserving federated learning scheme with secure authentication and aggregation for internet of medical things," *IEEE journal of biomedical and health informatics*, 2023.

[59] Y. Lu, X. Huang, Y. Dai, S. Maharjan, and Y. Zhang, "Blockchain and federated learning for privacy-preserved data sharing in industrial iot," *IEEE Trans. Ind. Informat.*, vol. 16, pp. 4177–4186, 2020.

[60] S. Gade and N. Vaidya, "Privacy-preserving distributed learning via obfuscated stochastic gradients," in *2018 IEEE Conf. on Decision and Control (CDC)*, pp. 184–191, 2018.

[61] A. Fu, X. Zhang, N. Xiong, Y. Gao, H. Wang, and J. Zhang, "Vfl: A verifiable federated learning with privacy-preserving for big data in industrial iot," *IEEE Trans. Ind. Informat.*, vol. 18, pp. 3316–3326, 2020.

[62] M. Song, Z. Wang, Z. Zhang, Y. Song, Q. Wang, J. Ren, and H. Qi, "Analyzing user-level privacy attack against federated learning," *IEEE J. on Selected Areas in Commun.*, vol. 38, pp. 2430–2444, 2020.

[63] K. Yue, R. Jin, C.-W. Wong, D. Baron, and H. Dai, "Gradient Obfuscation Gives a False Sense of Security in Federated Learning," *ArXiv*, vol. abs/2206.04055, pp. 6381–6398, 2022.

[64] X. Liu, H. Li, G. Xu, Z. Chen, X. Huang, and R. Lu, "Privacy-enhanced federated learning against poisoning adversaries," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 4574–4588, 2021.

[65] H. Liu, B. Li, C. Gao, P. Xie, and C.-l. Zhao, "Privacy-encoded federated learning against gradient-based data reconstruction attacks," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 5860–5875, 2023.

[66] Y. Jiang, Y. Li, *et al.*, "Mitigating sybil attacks on differential privacy based federated learning," *ArXiv*, vol. abs/2010.10572, 2020.

[67] L. Miao, W. Yang, *et al.*, "Against backdoor attacks in federated learning with differential privacy," in *ICASSP 2022 - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 2999–3003, 2022.

[68] Q. Li, X. Wang, and S. Ren, "A privacy robust aggregation method based on federated learning in the iot," *Electronics*, vol. 12, p. 2951, 2023.

[69] M. Asad, A. Moustafa, and C. Yu, "A critical evaluation of privacy and security threats in federated learning," *Sensors*, vol. 20, 2020.

[70] X. He, Y. Liu, *et al.*, "Adaptive local differential privacy mechanisms for heterogeneous iot data in federated learning," *J. Priv. Confid.*, 2024.