Volume 10, Issue 5, May – 2025 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may617

Security Threats and Defence Mechanisms in Federated Learning: A Comprehensive Review

Andrews Ocran[1]; Japhet Effah[2]

[1]Teesside University London
[2]Teesside University London

Publication Date: 2025/06/04

Abstract: Federated Learning (FL) is a promising decentralised machine learning model that enables multiple devices to collaboratively train a shared model without sharing their private data. While this approach enhances data privacy and regulatory compliance, it is significantly vulnerable to a range of security threats and adversarial attacks. This research seeks to investigates various attack vectors in FL, such as poisoning attacks, Byzantine attacks, Sybil attacks, and gradient inversion and also evaluates their impact on model performance and data confidentiality. Through a comprehensive analysis and empirical reviews of existing literature, the study explores mitigation strategies, attack model and threat taxonomy to classify adversarial behaviours. Key findings from the reviews suggests that while existing defence mechanisms show promise, they often suffer from trade-offs between model accuracy, system scalability, and computational overhead. The study was concluded by identifying gaps in current literature, such as the need for adaptive mitigation strategies and more realistic threat models, and offers recommendations for future work. By addressing these challenges, the research strengthens the robustness and trustworthiness of federated learning systems in real-world applications.

Keywords: *Federated Learning (FL); Attacks, Privacy Preservation, Aggregation, Trusted Execution Environments (TEEs), Federated Averaging (FedAvg), Inference Attacks, Resilience Strategies, Machine Learning, Byzantine Attack.*

How to Cite: Andrews Ocran; Japhet Effah (2025). Security Threats and Defence Mechanisms in Federated Learning: A Comprehensive Review. *International Journal of Innovative Science and Research Technology,* 10(5), 3275-3293. https://doi.org/10.38124/ijisrt/25may617

I. INTRODUCTION

Federated learning (FL) is a decentralised machine learning algorithm where multiple devices or servers collaboratively train a shared global model without exchanging raw data but only the model updates necessary to improve data privacy and security (McMahan et al., 2017). This approach addresses an important issue in data privacy, bandwidth efficiency, and compliance regulations such as the GDPR (Truong et al., 2021). Federated learning achieve these tasks by forwarding local model updates to a central server for aggregation.

Despite the privacy-preserving principles of Federated learning, it introduces a security vulnerability as a result of its decentralised and trust-assuming nature (Truong et al., 2021). Federated learning is vulnerable to a range of adversarial attacks that can compromise a model's confidentiality, integrity and availability of its data, hence defeating the very purpose it sought to achieve (Xie et al., 2024).

Federated Learning is vulnerable to attacks such as poisoning attacks, where threat actors posing as benign clients to manipulate local training data or gradient updates to distort the global model (Lenaerts-Bergmans, 2024). These can be either data poisoning, where malicious samples are injected into the training database, or model poisoning, where attackers craft gradient updates to reduce the accuracy of the model or implant backdoors to be exploited later (Xie, Koyejo and Gupta, 2021). Poisoning attacks are difficult to detect and mitigate due to the lack of centralised data oversight.

Again, Federated Learning is vulnerable to privacy inference attacks, such as gradient inversion and membership inference. This is where threat actors seek to exploit shared model parameters (updates) to reconstruct private input data or determine data membership (Guo et al., 2024). Yang et al. (2023) demonstrated in their studies how intermediate feature maps can be reverse-engineered to retrieve sensitive data, thereby raising significant concerns about Federated Learning privacy assurance even when no raw data is exposed during the training. These findings emphasis the need for robust defences that extend beyond naive aggregation or differential privacy techniques.

To mitigate these threats, emerging solutions incorporate trusted execution environments (TEEs) and secure aggregation protocols. Chen et al. (2024) proposed a Federated Learning framework augmented with TEEs to protect the aggregation phase from adversarial attacks. While theses mitigation strategies are effective, they also introduce

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may617

trade-offs between computational efficiency, scalability, and hardware dependency (Zeng et al., 2024).

This paper investigates the taxonomy, methodology, adversarial impact of threats on federated learning systems, and defence strategies to combat federated learning attacks. The main motivation for the studies is to provide a rigorous yet comprehensive overview that informs both theoretical advancements and practical deployments in secure distributed learning systems.

➢ *Motivation for the Study*

The increasing adoption of distributed systems across many sectors such as healthcare, finance, and Internet of things has created a paradigm shift in how machine learning models are trained, to emphasis user privacy and data decentralisation. Due to the decentralised nature of these systems they also introduced significant vulnerabilities that could be exploited by adversaries (Benmalek, Benrekia and Challal, 2022). Some of these threats undermine the integrity and confidentiality of the global model, making the robustness and trustworthiness of FL systems an issue of concern (Almutairi and Barnawi, 2023). Motivated by the complexity of evolving adversarial attacks and inadequate oversight of client activities by central servers, its imperative to explore the varied spectrum of existing attack vectors and their subsequent adversarial impact on the Federated Learning systems. This study is motivated by the need to understand the intricacies of these attacks vectors and evaluate existing mitigation defence strategies to propose a more robust frameworks to safeguard federated learning systems. Moreso, with most evolving applications and systems depending on Federated Learning architecture for secure and privacy-preserving learning, there the need for empirical insights and theoretical models that can guide the future development of robust learning algorithms. By analysing current scholarly works and documented existing Federated Learning adversarial threats, this study seeks to make a immerse contribution to the future development and deployment of a more robust Federated Learning architectures, that fosters trust among users and stakeholders who relies on collaborative learning systems. The study also seeks to identify gaps in existing mitigation strategies and highlight novel research opportunities that can improve the robustness of a secure federated learning system.

➢ *Overview of Federated Learning*

As discussed earlier in the introduction, federated learning is a distributed machine learning algorithm that enables collaborative model training across multiple devices or clients without these devices having to share their individual raw data during the model training (Criado et al., 2022). Federated learning facilities collaborative training and development of a model while preserving data privacy and reducing communication costs (Cao et al., 2022). In Federated Learning systems, the clients perform learning using their local datasets under the control of a central server (Liu, Xu and Wang, 2022). Federated Learning can be implemented in both centralised and decentralised architectures. In centralised federated learning, the client devices devices rely on a single central server for updates, while decentralised approaches like Chain Federated Learning utilise blockchain technology to distribute model data across multiple network nodes, mitigating the risk of single-point failure (Cao et al., 2022).

➢ *Federated Averaging.*

Federated Averaging (FedAvg) is the foundational algorithm used within federated learning platform to aggregate the model updates from the participating devices. When the central server receives the model updates from the clients' devices, it determines the weighted average of these updates by computing the weighted average before sending back the aggregated update to the client for further training and learning (Sun, Li and Wang, 2023). Federated averaging is designed in such a way that the global model represents the real contributions of all devices proportionally, taking into account the data amount each device sent (McMahan et al. 2017). The process of aggregation is usually performed using one of the two methods: Simple Averaging or Weighted Averaging, based on the size of the local datasets.

- *Simple Averaging:*

Simple Averaging combines the model updates from all clients by giving equal weight to each client's contribution, regardless of how much data they have (Betul Yurdem et al., 2024). After the local model training done by the client on their own data, the updated parameters are sent back to the central server to perform the averaging of individual parameters to produce a new global model (McMahan et al. 2017). In situations where computational simplicity and efficiency matter greatly, this approach is usually applied, e.g., mobile device personalisation, Internet of Things (IoT) networks, and healthcare diagnostics, particularly for the case when the clients have limited computing capabilities (Qi et al., 2024).

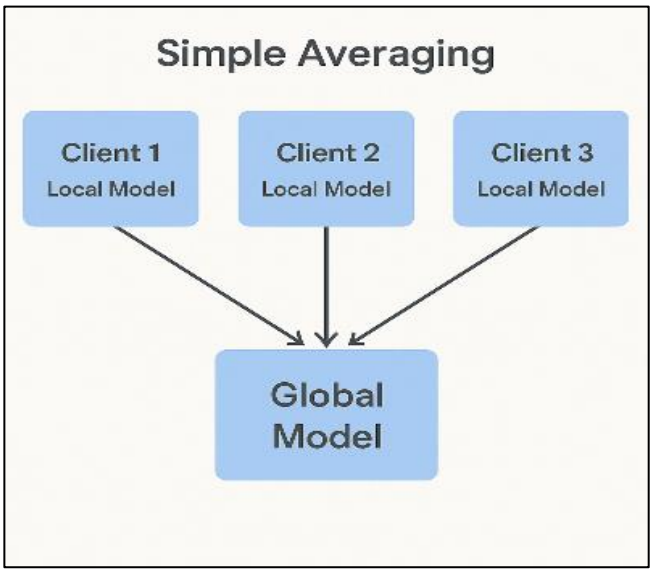

Fig 1 Simple Averaging

Simple Averaging is computed using the formula:

$$w = \frac{1}{n}\sum_{i=0}^{n} \text{wi}$$

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may617

Where:

n: Number of clients (devices/participants) that trained a local model.

w_i : Model weight (or set of weights) produced by the i^{th} client after local training.

w : The new global model weight, obtained by simply averaging all the client updates.

For instance, if 3 three clients are participating in learning, and all clients have amount of training data, each contributes equally to the final model.

w_1=1.5, w_2=2.0, w_3=2.5

Table 1 Simple Averaging Aggregation

Client	Local weight w_i
1	1.5
2	2.0
3	2.5

$$w = \frac{1}{3}(1.5 + 2.0 + 2.5) = \frac{6.0}{3}. = 2.0$$

The computed value of 2.0 represents the new global model parameter (weight) after aggregating the local model updates from the clients. The new weight of 2.0 is significant because it becomes the shared global weight for the next round of training.

- *Weighted Averaging:*

In weighted Averaging, the central server computes an aggregates of each model updates received from participating devices, giving more influence to clients that contributed more data towards the learning. After each client performs local training on its private dataset, it forwards its updated model parameters to the central server. The server then computes a new global model using a weighted average of all the received updates (Betul Yurdem et al., 2024).

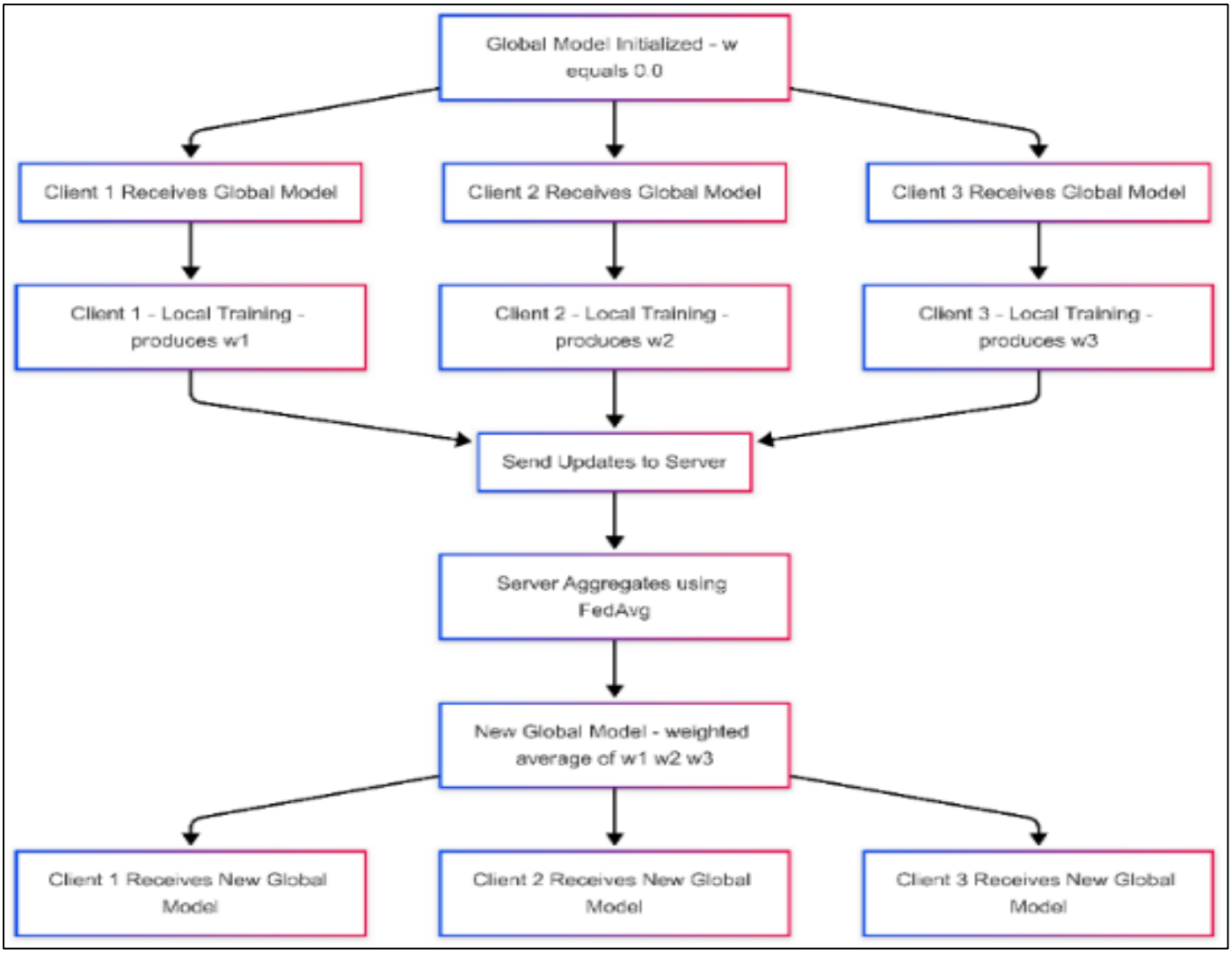

Fig 2 Federated Learning Weighted Averaging

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may617

Weighted average is represented by the formula:

$$w_{\text{global}} = \frac{\sum_{i=1}^{n} n_i \cdot w_i}{\sum_{i=1}^{n} n_i}$$

Where:

- n_i: number of training samples for client i
- w_i : model weights from client i
- w_{global} : the new global model after aggregation

Suppose we have 3 clients participating in federated learning. Each client trained their local model on different amounts of data:

- Client 1: $w1$ = 1.5, trained on 3 data samples
- Client 2: $w2$=2.0, trained on 4 data samples
- Client 3: w3=2.5 trained on 3 data samples

Table 2 Weighted Averaging

Client	Local Weight w_i	Number Of Samples n_i
1	1.5	3
2	2.0	4
3	2.5	3

- Client1: $\frac{3}{10} = 30\%$
- Client2: $\frac{4}{10} = 40\%$
- Client3: $\frac{3}{10} = 30\%$

$$w_{\text{global}} = \frac{3 \cdot 1.5 + 4 \cdot 2.0 + 3 \cdot 2.5}{3 + 4 + 3} = \frac{4.5 + 8.0 + 7.5}{10} = \frac{20.0}{10} = 2.0$$

The new final Global Model Weight after the aggregation is w_{global}=2.0.

The more data the client has, the more their contribution to the final model is, while the clients with less data also contribute, but the amount of their contribution is relative to their data size. For example, the update of Client 2 has a higher effect on the global model than that of Clients 1 and 3 combined, but the latter pair still has an effect, however small. Although the clients have smaller data, they are still part of the training process and they aid in creating the global model, but only until a certain extent.

II. RELATED WORK

Federated learning (FL) has gained considerable attention as a decentralised learning system that preserves data privacy by keeping training data localised on client devices. However, recent studies show that this architecture is susceptible to a variety of security vulnerabilities and adversarial threats. A growing body of research has emerged to examine these vulnerabilities and propose mitigating strategies.

Xie et al. (2021) provided a foundational survey of poisoning attacks in Federated Learning, highlighting how adversarial can inject poisoned data or manipulate model updates to mislead the global model. Qayyum, Janjua and Qadir (2022) propose a hybrid learning-based detection strategy to identify poisoned parameter updates, achieving high in detection rates against label-flipping attacks. Their work highlights the need for intelligent, context-aware defences that go beyond simple aggregation filters. From a privacy preserving perspective, Zhang et al. (2023) demonstrated that features shared during training can leak sensitive information. In their study, attackers are able to reconstruct private data from these shared parameters, thereby diffusing the very objectives that Federated Learning guarantees privacy. Complementing Zhang et al's perspective, Zhu, Liu and Han (2019) explored clean-label data poisoning attacks, where adversarial data samples appear legitimate to clients in the model training but are engineered to disrupt performance in federated learning systems.

Further, Xie et al. (2024) presents a taxonomy of Federated Learning vulnerabilities, grouped attacks into data-to-model, model-to-data, and model-to-model. Their survey emphasises the evolving sophistication of attacks, from overt gradient inversion to subtle perturbations in model parameters. Kasyap and Tripathy (2024) extend this by introducing hyperdimensional computing techniques to generate adversarial samples, showing that FL models are vulnerable to attacks that do not require access to model internals or labels.

These works collectively highlight the fragility of Federated Learning systems in adversarial environments. Insights from review of related literature also demonstrate that effective and resilient mitigation strategies must account for both the decentralised architecture and the heterogeneity of participating devices. Despite promising developments in secure aggregation and trusted execution, the literature continues to emphasise the need for adaptive, scalable, and lightweight solutions to protect federated learning systems in real-world deployments.

A. Gaps in Existing Literature.

Despite growing research interest in the security of federated learning (FL), several critical gaps remain unaddressed in the current literature. One of such limitation is the fragmentation of defence strategies. Currently most proposed solutions are highly specialised and target a single attack type, such as data poisoning or gradient inversion, without accounting for more complex, multi-vector adversarial attacks. Robust aggregation methods, such as Krum and Trimmed Mean, are a good approach to outliers, but very bad for backdoor attacks that are stealthy and blend with benign updates (Xie et al., 2021).

One significant gap identified during literature review is the lack of consistent and scalable frameworks for the detection and mitigation of adversarial attacks. Zhang et al.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may617

(2023) indicated that differential privacy can be quite effective in the face of the inversion of gradients, but most of these methods will still go against the use of the model, and very few studies have offered a rational framework for the trade-off between privacy and performance.

Although there is a good classification of FL threats by Xie et al. (2024), we observe a scarcity of empirical papers that discover identity-based and coordination-based attacks. Most of the defenses strategies are based on the ideal scenario of using a partial trust model or the static client behavior. However, these do not correspond to reality in cases of open or cross-device systems, where the attackers can easily increase client participation. Again, there is a noticeable gap in socio-technical considerations, such as client trust modelling, authentication, and incentive mechanisms. Current studies largely ignore the economic and behavioural dimensions of Federated Learning systems, assuming that all clients are either benign or malicious, with no spectrum in between. This binary framing overlooks real-world nuances, such as semi-honest participants or accidental failures that may resemble adversarial behaviour. The evaluation metrics and benchmarks used in many studies lack standardisation, which hinders the reproducibility and comparability of their proposed defences. The absence of comprehensive framework suites that simulate realistic attack scenarios across diverse data distributions, model types, and network conditions limits the ability to rigorously assess the resilience of Federated Learning systems.

B. *Components of Federated Learning (FL)*

➢ *Central Server:*

The central server is responsible for coordinating the entire process of Federated Learning (FL). The central server initializes the global machine learning model, by selecting a subset of participating devices, and aggregating the updates received from these clients to refine the global model. The central server does not have direct access to raw data on the client's device. This is necessary to ensure data privacy and regulatory compliance (McMahan et al., 2017). An example of a central server in Federated Learning can be found in Google's Gboard application. The central server deploys a text prediction model to millions of Android devices. Each device updates the model locally using the user's typing data and sends only the updated model parameters such as gradients back to the server. The server then performs federated averaging (FedAvg) to aggregate these updates to improve the global model without ever seeing individual user inputs (McMahan et al., 2017).

➢ *Clients:*

Clients are the participating devices that hold local datasets and perform computations on them. The clients ranges personal mobile phones and IoT devices to large institutions like hospitals or banks. Each client receives the current version of the global model from the central server, trains it using local data, and then sends only the updated model parameters back to the server (Qayyum, Janjua and Qadir, 2022). In healthcare for instance, multiple hospitals may participate as clients to collaboratively train a diagnostic model for detecting pneumonia from chest X-rays. Each hospital keeps its patient data secure and private but contributes to a more robust and generalised model through localized training (Qayyum, Janjua and Qadir, 2022). A case study involving such a setup demonstrated how FL could be used to train a COVID-19 detection model from CT scans across hospitals in different regions while complying with patient data privacy laws like GDPR (Sheller et al. 2020)

➢ *Communication Protocols:*

The communication protocol defines how model updates are exchanged between the central server and the clients. The protocols must be both secure and bandwidth-efficient given that more clients' device will be participating in the learning process. Techniques such as quantisation and sparsification are used to compress data to reduce the size of updates. Technique such as secure aggregation is deployed to ensure that updates cannot be reverse-engineered to reveal sensitive information (Bonawitz et al., 2019). Apple has deployed secure communication protocol to improve Siri and dictation services. Apple uses encrypted protocols combined with differential privacy to ensure that updates sent from iOS devices to Apple servers cannot be traced back to individual users. This has improved the accuracy of Apple's voice recognition and text prediction models while maintaining a high standard of user privacy (Truong et al., 2021).

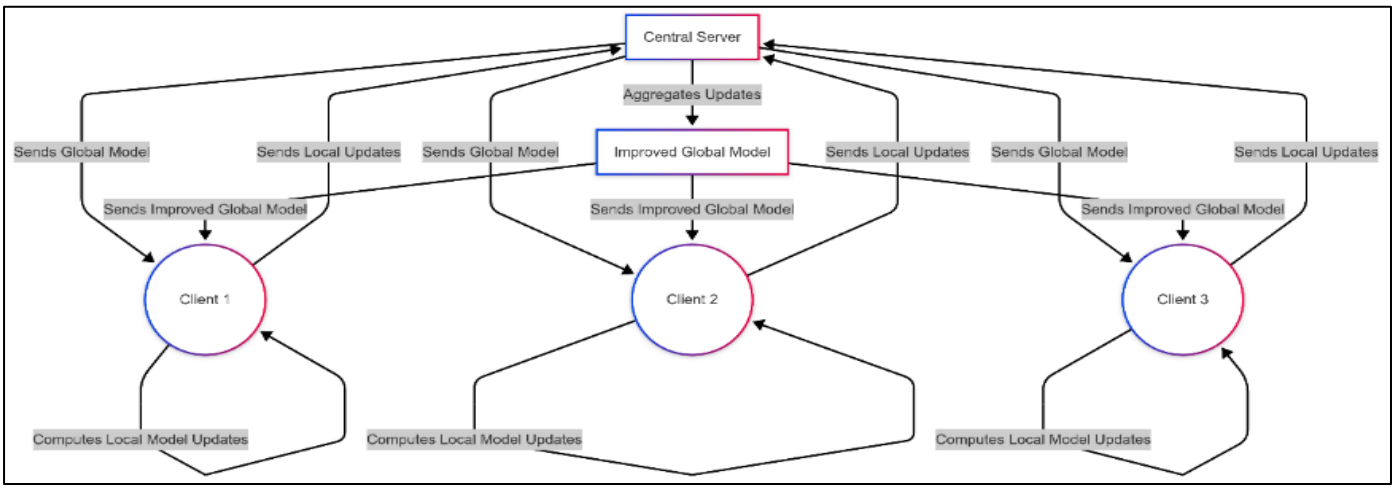

Fig 3 Federated Learning Architecture

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may617

C. Workflow of Federated Learning

The workflow begins with the central server initializing and distributing the model to client devices that are targeted. Each of the clients then trains the model utilizing their own local data privately and then transfers the update to the server. The server collects these updates via algorithms like Federated Averaging, thus generating an enhanced global model (Betul Yurdem et al., 2024). The process of iterations, which is depicted in Figure 4, goes on for the entire communication rounds until the stage of completeness and stability is reached.

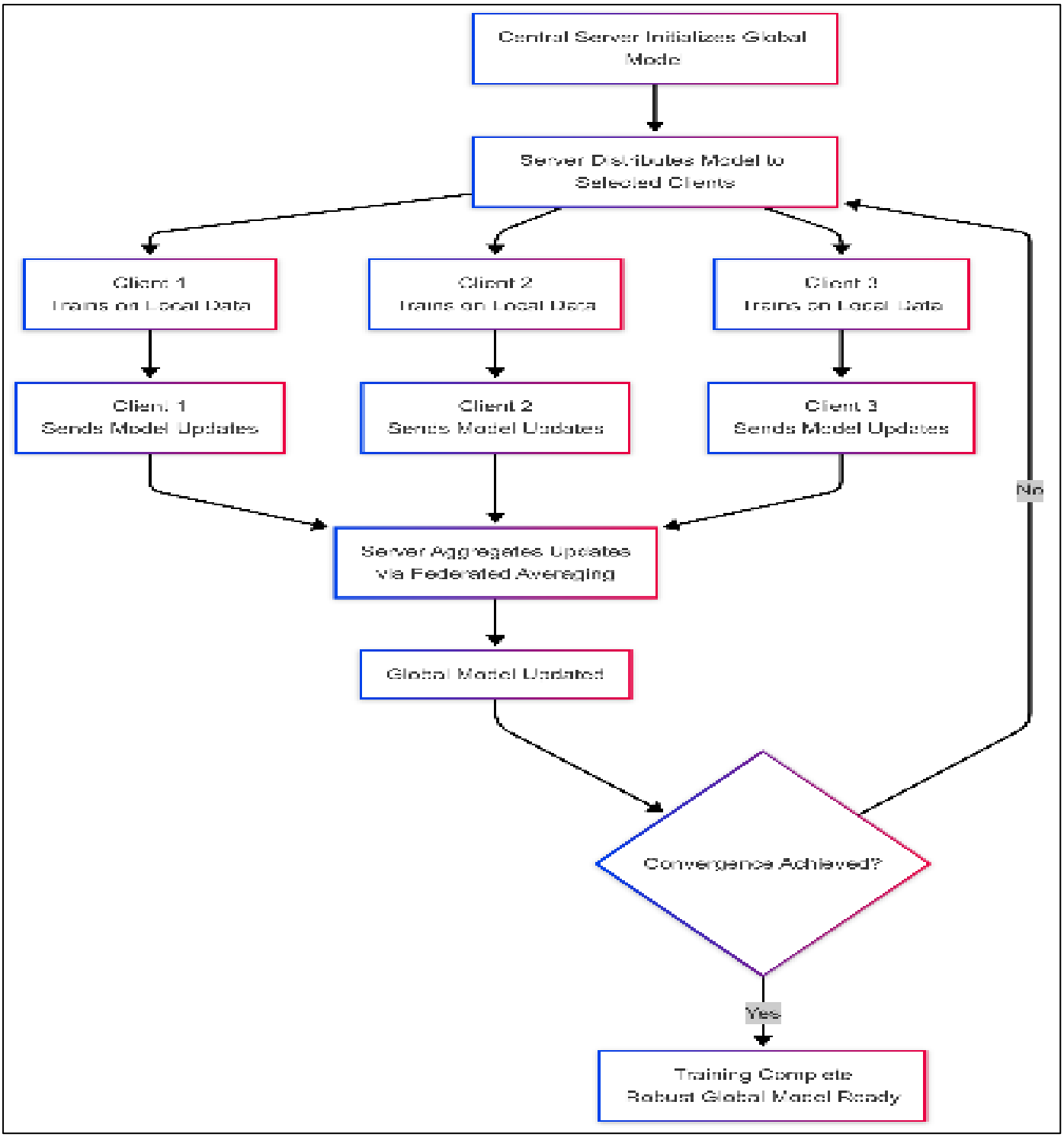

Fig 4 Federated Learning Workflow

- *Initialisation:*

At the beginning of federated learning process, the central server initialises a global machine learning model. The model's weights are either set randomly or initialised using pre-trained parameters. McMahan et al. (2017) introduced the FedAvg algorithm, where a simple neural network is initialised before being distributed to client devices.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may617

- *Client Selection:*

In this stage, the server selects a portion of clients from a larger pool of available clients. It can be a random selection, or a specific selection based on conditions, like the device is availability, the network connectivity, battery level, and the usefulness local data is useful. An example is selection of the devices in Google Gboard where a devices are chosen to participate in the learning process following a device availability protocol, thus healthy only devices meeting the conditions can participate in the learning (Xu et al., 2023).

- *Model Distribution:*

After successful selecting clients, the server sends the current version of the global model to the selected devices. The model is transmitted over the network, using compression techniques to reduce communication overhead. Model quantisation and pruning compression techniques can be applied to transmit smaller data, especially in environments with limited bandwidth like mobile networks (Liang et al., 2021).

- *Local Training:*

Selected client locally trains it model based on the the received model from the central server. The training usually involves the of uses optimisation algorithms like stochastic gradient descent (SGD) (Vungarala, 2023). Clients perform multiple local training to improve model performance using only their data. In a hospital's Federated Learning system, models can be trained on patient imaging data without necessarily revealing its sensitive data. This approach is useful in other privacy-sensitive domains like IoT and finance (Bonawitz et al., 2019).

- *Update Upload:*

After completing the model training, each client sends back the updated model parameters to the central server. The updates to be forwarded to the central server are encrypted to preserve user privacy. Techniques like differential privacy add carefully calibrated noise to updates, ensuring that individual data points cannot be reverse-engineered from the shared gradients. Secure aggregation techniques used ensure that the server cannot see individual updates but only the combined result from each participating device (Wei and Rao, 2024).

- *Aggregation:*

Updates received from multiple clients are aggregated to form a new version of the global model at the server. Federated Averaging (FedAvg) is the most common aggregation technique used for updates aggregation (Sun, Li and Wang, 2023).

- *Convergence Check:*

Finally, the server evaluates the performance of the updated global model using a validation dataset or client feedback. If the model's performance has reached a satisfactory level or a preset number of rounds have been completed, training stops. Otherwise, the process loops back to client selection until model convergence (Nanayakkara, Pokhrel and Li, 2024).

D. Types of Federated Learning

Federated Learning (FL) can be categorised into distinct types based on how data is distributed across clients. The classifications Horizontal FL, Vertical FL, and Federated Transfer Learning (FTL) offer an insight into which FL approach suits a particular collaborative learning. Each type addresses different issues among participating devices.

➢ *Horizontal Federated Learning (HFL):*

Horizontal FL, also known as sample-partitioned Federated Learning, is a type of Federated learning where participating clients share common feature space but differ in sample space. The clients have similar types of data but different users or instances (Naik and Naik, 2024). Horizontal Federated Learning is applicable in cases like mobile devices, where for instance, each user's smartphone may log similar features such as app usage time, location data, or accelerometer readings, but this data is exclusive to individual users (Lutho Ntantiso et al., 2023). Federated Averaging (FedAvg) algorithm, is used for updates aggregation (Yang et al., 2019). Google's Gboard and Apple's predictive text systems are practical examples of HFL, where model training occurs locally on millions of devices, each contributing distinct user-specific data points.

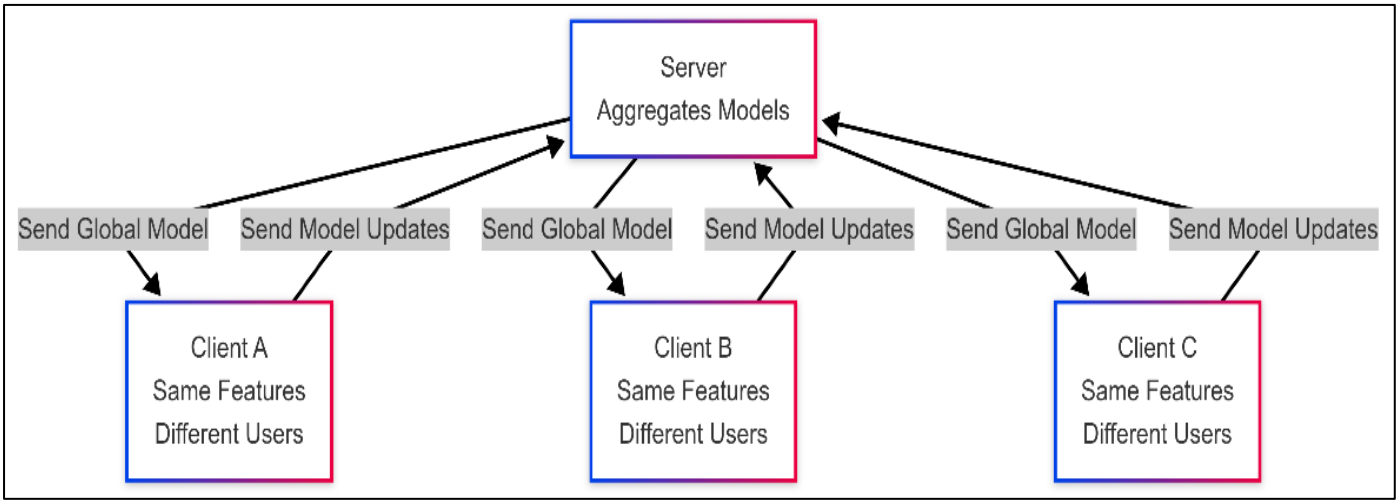

Fig 5 Horizontal Transfer Learning

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may617

➢ *Vertical Federated Learning (VFL):*

Vertical Federated Learning also known as feature-partitioned FL, is used in situations where clients hold different sets of features about the same set of users or samples. This often occurs in business collaborations where clients share customers but gather different types of data (Lutho Ntantiso et al., 2023). An of Vertical Federated Learning is a collaboration between a bank and an e-commerce company. The bank holds users' financial data, while the retailer holds purchase history. VFL aims to jointly train models that benefit from this complementary information, requiring sophisticated privacy-preserving techniques such as secure multi-party computation (SMPC) and homomorphic encryption to align and utilize overlapping samples without revealing raw features (Yang et al., 2019).

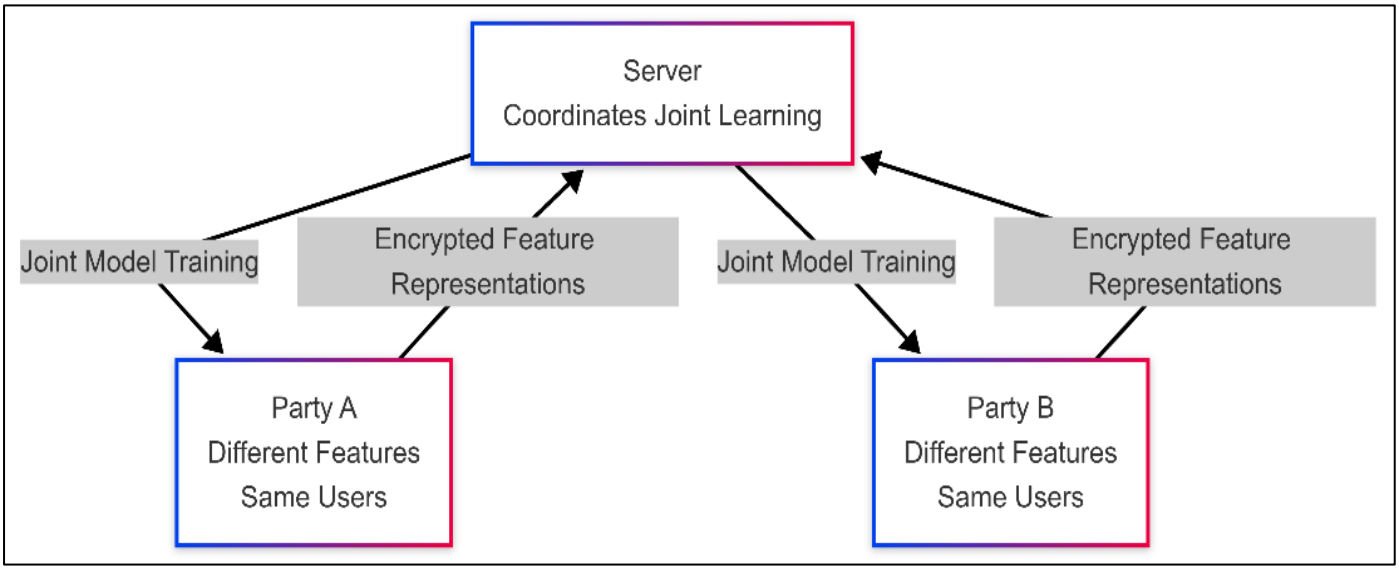

Fig 6 Vertical Federated Learning

➢ *Federated Transfer Learning (FTL):*

Clients have both different sample spaces and feature spaces meaning their data distributions overlap neither in users nor in data types (Hu et al., 2024, Lutho Ntantiso et al., 2023). This situation arises in international or cross-industry collaborations where data collection standards, features, and populations diverge widely. FTL leverages transfer learning techniques to enable knowledge sharing through shared latent representations or pre-trained models (Huang et al., 2021). Federated Transfer Learning implantation could involve a hospital in one country using FTL to enhance its model with insights from an unrelated institution's data in a different region, despite differences in demographics (Liu et al., 2020).

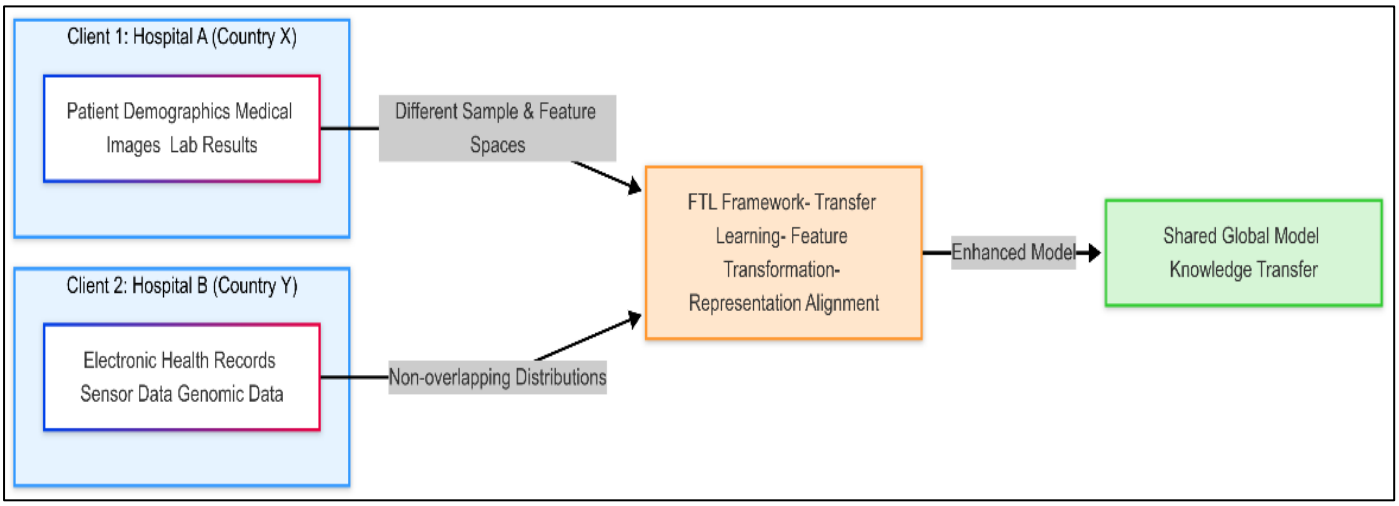

Fig 7 Federated Transfer Learning

E. Threat Model

The threat model encompasses a range of adversarial threats that seeks to exploit the system's decentralised nature, heterogeneity of participating clients, and limited transparency into client-side computations. Adversaries in FL may be either internal, posing as legitimate users within the system or external, with the ability to intercept, infer, or manipulate communications. Given that clients maintain control over their local data and model training, malicious behaviour can easily go undetected unless sophisticated mitigation strategies are in place.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may617

F. Adversary Capabilities:

The adversary is modelled as a participant-level device that is legitimately part of the FL system but behaves maliciously. The adversary has full control over their local training model, allowing them to manipulate their training data, customise model update process, or injection of crafted malicious updates. The adversary may also act as a Sybil attacker, controlling multiple malicious clients to increase its influence during aggregation. The adversary is assumed to know the system architecture, including the aggregation algorithm, but does not have access to the internal training processes or data of other legitimate clients.

G. Attack Goals:

The objectives of the attacker in this model can be categorised into four main goals:

➢ *Integrity Violation:*

The attacker aims to degrade the accuracy, reliability, or fairness of the global model. This includes data and model poisoning attacks where the goal is to mislead the model into incorrect predictions, either in a general sense or for specific targeted classes. An attacker might cause systematic misclassification by injecting poisoned data samples with flipped labels or crafted gradients (Xie et al., 2021).

➢ *Privacy Breach:*

A primary motive of most passive attack scenarios is to exfoliate sensitive information from other participating devices detected. This can includes attacks such as gradient inversion, where attackers reconstruct private training data from shared model updates (Du et al., 2023), and membership inference attacks, which determine if a specific data sample was used in a client's training set (Xie et al., 2024).

➢ *Availability Disruption:*

The adversary seeks to disrupts the overall learning process by preventing the model from converging. This is be achieved by submitting noisy, irrelevant, or contradictory updates over several rounds, thus introducing instability into the optimisation process.

➢ *Evasion and Stealth:*

Advanced adversaries prioritise stealth to remain undetected while pursuing their objectives. Clean-label poisoning attacks exemplify this goal by using legitimate labels with subtly crafted features to evade detection while still influencing the global model (Zhou et al., 2024). Backdoor attacks embed specific triggers in the model that activate only under rare input patterns, allowing the adversary to maintain the model's performance on clean data while acting on attacker-defined inputs (Kasyap & Tripathy, 2024).

H. Federated Learning Attacks

The decentralised nature of the Federated Learning exposes the system to various types of attacks that inherently compromise the privacy, security, and performance of the model. Attacks can be carried out by an adversary who seeks to exploit vulnerabilities within the federated learning system. Discussed below are various attack types that an adversary can perpetuate against a federated learning systems, each with unique objectives and methods of execution.

➢ *Poisoning Attack*

Poisoning attack is a deliberate attempt by a malicious adversary to change training data or model updates to compromise the integrity and reliability of the global model. Due to the decentralised nature of the Federated Learning system, poisoning attacks can be executed covertly, making them particularly difficult to detect and mitigate (Sagar et al., 2023). Poisoning attacks can be classified into two types: DPA (Data Poisoning Attack) and MPA (Model Poisoning Attack) (Sagar et al., 2023).

- *Data Poisoning Attacks:*

Data poisoning attack involves an attacker gaining access to the training data of one client and manipulating or corrupting the data that is used to train a Federated learning model (Aljanabi et al., 2023). The primary motive of data poisoning is to infiltrate the system with malicious data to influence the model's performance and behaviour, resulting in incorrect or biased prediction (Verde, Marulli and Marrone, 2021). Data poisoning attacks can further be classified into targeted attacks and untargeted attacks (Huang et al., 2011). Targeted attacks occur when an adversary tries to influence a model's behaviour with a specific objective. Targeted attacks are difficult to control as the attacker has set a specific goal to achieve with the attack, but can have major and far-reaching adversarial impact on the model. Untargeted or random data poisoning attacks aims to alter the model's dataset in order to other to reduce the model's accuracy and overall performance (Lyu, Yu and Yang, 2020). Data poisoning attacks include following attack types.

- *Label-Flipping:*

Label-flipping is a dirty-label attack where adversaries change a portion of the training data after gaining access to the system, but reserve the remaining portion with the intent to manipulate the Federated learning model (Sagar et al., 2023). Instead of altering the actual input features, attackers flip or modify the labels of certain samples, causing the model to learn incorrect associations (Moharram et al., 2022). For example, in an image classification learning system, the adversary might change the label of a "dog" image to "cat." If the attacker is able to introduce more mislabelled samples, the model will begin misclassifying legitimate inputs, thereby reducing accuracy and reliability of the model.

- *Backdoor Poisoning:*

The motive of a backdoor attacker is to modifies a small portion of the original training model to embeds a trigger of a specific pattern to create a backdoor in other to influence the model to behave to the whims of the attacker. When the model is deployed, the trigger in the model will make a predetermined incorrect decision hence compromising the Federated learning architecture (Lyu, Yu and Yang, 2020).

➢ *Model Poisoning:*

An adversary performing model poisoning attack aims is to poison the local model before forwarding the model update to the central server for aggregation. The attacker

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may617

injects enough corrupted data to the model that will cause the model to misclassify set of predetermined input with certainty (Bagdasaryan et al., 2019).

- *Inference Attacks:*

Inference attacks are threat vector used by adversaries to exfiltrate sensitive information from a learning model, by exploiting the model's behaviour and outputs. Inference attacks does not require access to the underlying training data but instead rely on querying the model and analysing its responses or updates to make inference about the data used for training (Chen et al., 2020).

- *Membership Inference Attack:*

Bad actors use membership inference attacks vectors determine whether a particular data sample was part of the training dataset. Even if the attacker cannot directly access the training data, they can query the model with various inputs and analyse the model's responses to make inferences about whether a specific data sample was used during training (Chen et al., 2020).

- *Attribute Inference Attack:*

An attribute inference attack involves exfiltrating sensitive attributes such as medical conditions, age, gender, etc. from the model's output, which might occur when models inadvertently reveal too much information about the underlying data (Struppek et al., 2023). The adversary queries the model to infer specific attributes that might be of interest about a client, based on the model's predictions (Zi et al., 2021). For example, if a model is trained to predict medical conditions, an attacker might use the model's output to infer whether a person has a particular condition.

- *Sybil Attacks:*

Proposed by Douceur (2002), is an attack where the attacker intends to controls multiple fake or duplicate clients with intention of gaining disproportionate influence over the global model aggregation process. Its originates from the notion of an entity masquerading as many, a concept rooted in distributed computing and peer-to-peer networks. Xie et al. (2021) and Xie et al. (2024) discuss Sybil attacks as a particularly destructive variant of poisoning or backdoor attacks, uses multiple fake identities to ehance the impact of malicious updates. Since FL lacks strong identity verification mechanisms, most especially in a cross-device architecture, attackers exploit this vulnerability to control a larger share of influence (Feng et al., 2025).

- *Data Reconstruction Attack:*

Data reconstruction attack exploits the inherent correlation between the gradients and the underlying data used for aggregation. By observing these gradients, a malicious client or external attacker can employ optimisation techniques to approximate the data samples that most likely produced the samples (Huang, Huo and Fan, 2024). Gradient inversion attacks such as Deep Leakage from Gradients (DLG) use iteration to generate inputs that, when passed through the model, yield gradients similar to those received. This approach has been shown to successfully reconstruct sensitive data such as medical images or text, even when the original data remains local to the client (Ding et al., 2024).

- *Free Rider Attack:*

This is where client without any relevant contribution towards the model training benefits from quality model update (Chen et al., 2024). The free rider enjoys benefits such as computation, energy, or bandwidth without offering any contribution but in return submit arbitrary updates to the server. Despite providing no valid input, the attacker continues to receive improved global model versions aggregated from the honest clients' updates (Chen et al., 2024). This behaviour not only exploits the resources of network but may also degrade model performance and skew fairness metrics. Free-rider attacks are difficult to detect because the FL server cannot directly verify the origin or authenticity of local updates. (Araki et al., 2016).

- *Byzantine Attacks:*

Byzantine attack occurs when one or more malicious clients intentionally deviate from the protocol to disrupt, corrupt model training or degrade system performance. Originally belonging to the *Byzantine Generals' Problem* (Lamport et al., 1982). In federated learning (FL) Byzantine nodes may submit poisoned gradients, sign-flipping attacks to skew model updates. Collude to create sybil nodes to dominate aggregation (Baruch et al., 2019) (Li et al., 2022).

I. Taxonomy of Attacks:

To effectively defend against Federated Learning threats, it is essential to categorise and analyse the different types of attacks that can compromise privacy, security, and model performance. This section discusses a taxonomy of attacks in federated learning systems.

➢ *Data-to-Model (D2M) Attacks:*

The D2M attacks are carried out by modifying the local data of all the devices that actively take part in the learning processes of the system. The main motive behind a D2M attack is that the attacker can aim at the models which are being learnt without the requirement of weights, or updates of the model (Xie et al., 2024). Common techniques used includes removing or altering data labels or adding noise that makes the global model not to cluster. As a result, the model might perform very poorly and be untrustworthy. Data poisoning falls under this taxonomy of attack (Radford et al., 2019).

➢ *Model-to-Model (M2M) Attacks:*

Model-to-model attacks refer to local model updates or weights that are intentionally misused to have an effect on the global model. So, these attacks can disrupt the learning process by causing the main server to receive incorrect updates. An example is the use of model poisoning and Sybil attacks (Radford et al., 2019).

➢ *Model-to-Data (M2D) Attacks:*

The main goal of the M2D attack is to expose certain properties or fragments of the data on which the model is trained and perform the attacks. The interaction between

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may617

models and data is the main source of this kind of attack, in which the sensitive information is being exploited (Radford et al., 2018).

- *Composite Attacks:*

Composite attacks are quite advanced and posed by the attackers to target the multiple components of the FL process. Generally, different attack vectors would be joined together under composite attacks, e.g., Data-to-Model (D2M) and Model-to-Model (M2M) might be combined to lunch the attacks to achieve their goals (Radford et al., 2018). Its involves adding triggers of specific patterns to the local training data poisoning the models update which results in backdoors attack thus the global model learns to respond to certain triggers while at the same time it appears normal with clean data (Radford et al., 2018).

- *System-Level Attacks:*

These attacks are geared towards rendering the overall functioning of the FL system inefficient, rather than simply aiming at the data or the model. This group of threats consists of attacks such as free-rider attacks (when the adversaries involved just take advantage of the system without contributing worthfully to the training process). Direct changes in the model's logic may not be performed but the reliability and collaborative efficiency of the model may be made invalid (Liu, Xu, and Wang, 2022).

J. Impact of Adversarial Attack on Federated Learning

Adversarial attacks pose significant and multifaceted threats to the reliability, privacy, and trustworthiness of federated learning (FL) systems. The decentralised nature of the federated learning system creates a fertile ground for attackers to exploit its vulnerabilities. From the review of related literature, the impact of these attacks spans across model accuracy, privacy concern, system availability, and system robustness.

- *Degradation of Model Accuracy and Utility:*

One of the most noticeable and immediate consequences of an attack is degradation of model accuracy and utility, often caused by poisoning attacks. For example, Xie et al. (2021) reveal that both data and model poisoning attacks can lead to the global model mistaking the inputs or incorporating certain biases. Moreover, the attackers operating the training data or the gradients can not only impair their model performance on cleanly set samples but also create inputs and outputs to produce what they desire (e.g., backdoors) (Yang et al., 2019). Besides, this makes FL systems less trustworthy in safety-critical areas such as healthcare, where wrong predictions can lead to lethal consequences.

- *Compromise the Core Privacy Promises:*

In addition to degrading the model's accuracy and utility, adversarial attacks compromise the core privacy promises of FL. Ge et al. (2023) show through their work that the privacy of data is endangered even if data is not shared during model training. Attackers can reverse-engineer training inputs through gradient inversion or extract membership information from model updates. The breach of privacy clearly not only undermines user trust but also leads to the non-compliance of data protection regulations such as GDPR (Truong et al., 2021).

- *System-Level Disruptions:*

Adversarial influence exposes the system to other types of disruptions as well. An example of such attacks is the so-called Sybil attack dealt with by Xie et al. (2024). It is a type of attack that enables a single malicious attacker to change the global learning update by creating various fake clients. This has the effect of changing the aggregation process and of the underlying defences that are based on the idea of honest-majority participation (Zhang et al., 2024).

- *Persistent and Undetected Corruption of the Global Model:*

Perhaps the most challenging of adversary attack is that the use of clean-label poisoning by stealthy attacker to circumvent the existing defences mechanism, which result in a long-term and unnoticed corruption of the golbal model (Kasyap & Tripathy, 2024). These attacks have a long-term impact by embedding malicious behaviours that only manifest under specific conditions, making them difficult to detect and reverse once deployed (Benmalek, Benrekia and Challal, 2022).

III. MITIGATION STRATEGIES

Although Federated Learning (FL) offers significant advantage in terms of data privacy and decentralized model training, it can be attacked and cause the efficacy and security of the system to be compromised. Thus, to prevent FL systems from being compromised, it is crucial to employ suitable defence strategies that not only neutralize but also eliminate these threats. The identified strategies are intended for the identification of attacks, the preservation of data privacy and the assurance of the continuous operation of the global model. The focus of these mitigation strategy is to introduce a number of approaches to protect the FL infrastructures from a variety of possible attacks.

- *Data Poisoning Attacks:*

Data poisoning attacks can be mitigated by deploying robust aggregation algorithms, such as Krum, Trimmed Mean, and Median. These algorithms are commonly employed to resist poisoned updates by minimising the impact of adversaries during model aggregation (Xie et al., 2021). These algorithms ensure that malicious updates with extreme values do not take control over the global model. Again, client behaviour auditing strategies, which monitor the consistency and statistical properties of local updates over multiple rounds, can help identify clients that frequently submit suspicious or harmful updates (Kasyap & Tripathy, 2024). Data sanitisation techniques at the local level, such as adversary detection or label consistency checks.

- *Model Poisoning Attacks:*

Anomaly detection systems are crucial to detect and mitigate adversary actors that inject model updates malicious data. Anomaly detection systems calculate the similarity of each client's data to other participating devices and signal that which are quite different (Du et al., 2023). By enforcing the

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may617

use of differential privacy (DP) at the time of local training updates are regulated so no single client can change the global model (Zhou et al., 2024). Gradient clipping and normalisation methods are some more tools that can help as they restrict the maximum possible update and also prevent poisoned updates from highly contaminating aggregation.

➢ *Backdoor Attacks:*

Backdoor attacks can be mitigated by the deployment of server-side validation techniques that test models using supporting datasets containing potential triggers. By introducing synthetic trigger patterns during validation, servers can detect unusual misclassification behaviour (Xie et al., 2024). Injecting Gaussian noise into the model aggregation process diminishes the precision necessary for backdoor triggers to operate effectively (Kasyap and Tripathy, 2024). Adversarial training is another promising method where the model is intentionally exposed to intentionally crafted backdoor inputs during training to build robustness against such triggers (Li et al., 2023).

➢ *Membership Inference Attacks:*

Regularisation techniques such as dropout, weight decay, and label smoothing during model training can reduce overfitting, thereby minimising the model's ability to remember individual data points (Liman et al., 2024). Output interference strategies, like confidence masking or adding noise to model predictions, further obscure whether a sample was part of the training set (Wu et al., 2024). Limiting the number of client updates or participation frequency can also decrease the amount of information exposed about client datasets (Ribero, Vikalo and de Veciana, 2025).

➢ *Sybil Attacks:*

Preventing Sybil attacks requires robust client authentication protocols, such as using federated identities or cryptographic certificates to verify the legitimacy of participating clients (Xie et al., 2024). Another approach is weighting client updates based on their historical trustworthiness, giving lower influence to newly joined or unknown clients until their reliability is established (Douceur, 2002). Participation auditing, such as analysing device metadata, submission patterns, and consistency, can also detect if multiple identities originate from a single attacker (Zhang et al., 2024).

➢ *Free-Rider Attacks:*

To mitigates free-rider attacks, proof-of-contribution mechanisms can be implemented which requiring participating devices to demonstrate meaningful local training work, such as solving computational puzzles or achieving acceptable model improvements, before receiving global updates (Wang et al., 2024). Randomised client sampling and periodic validation against trusted additional datasets can detect clients whose updates are stale or randomly generated (Xie et al., 2021).

➢ *Clean-Label Poisoning Attacks:*

Clean-label attacks can be identified by anomaly detection systems, which identifies samples that, while correctly labelled, behave unusually in the latent feature space compared to clean data (Yin et al. 2024). The next step is to apply transfer of robust training to check if local updates do not jeopardize the performance of downstream tasks and thus detect poisoned representations well in advance. Another way to secure the models is to use models with soft labels that are generated by models whose performance has been checked and trained using this data before. Defence through distillation can result in the model being less sensitive to slight adversarial perturbations and hence making the system better protected (Gong et al. 2020).

➢ *Applications of Federated Learning*

The decentralized nature of federated system has made it possible for its applications sectors such as healthcare, finance and IoT systems where privacy and security are critical. FL has empowered corporations to generate powerful machine learning models that are incapable of breaching user privacy or any regulation.

- **Healthcare:** Federated Learning enable healthcare institutions train a machine learning model without sharing patient-related data thus, achieving the goal of securing the privacy of and not transferring user data. It is possible for hospitals to work together and use AI to identify diseases like cancer, diabetes, and heart conditions through analysing medical pictures or patient data without any personal data transfer (Brisimi et al., 2018).

- **Finance:** By using Federated Learning systems, banks and financial institutions can find and stop illegal activities in the financial system through the detection of fraud such as suspicious transaction behaviors, and without the necessity to keep all personal financial information in one place. Each bank can teach its models with transaction data and at the end only distribute the updates that were learned instead of the whole data (Yang et al. 2019).

- **IoT (Internet of Things):**Federated Learning (FL) can empower IoT devices in their training of local models by utilizing the decentralized data from the sensors, and also by customizing user interactions without the need to exchange sensitive data with the central server. This can be validated with an opportunity in which a device at home, such as a thermostat or voice assistant, can use FL to enable the user to have better experiences by feeding on the data from individual usages directly without disposing the personal information to the cloud. A smart thermostat is one of the cases where it can record the constant indoor temperature, hence, control itself by setting the most preferred temperature without forwarding any user's data to the cloud which is sensitive (Aggarwal, 2024).

IV. METHODOLOGY

A. Study Design

The study utilized a qualitative and exploratory methodology to explore attack vectors and mitigation strategies existing in Federated Learning (FL) systems. We

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may617

kicked off the studies by reviewing related literature, we extracted knowledge from peer-reviewed journals, conference, and benchmark technical reports to be able to know and also group the different types of adversarial threats in Federated Learning. After the classification, we designed an attack model framework which is built on standardized threat modeling practices and includes threat assumptions, attack goals, and attacker capabilities. A threat matrix is established to examine the impact and probability of each attack type in a systematic way, thus the vulnerabilities are identified most effectively.

The methodology also includes a comparative analysis of existing defence mechanisms such as secure aggregation, anomaly detection, robust aggregation functions (e.g., Krum, Trimmed Mean), and Trusted Execution Environments (TEEs). To evaluate these mechanisms, we analyse their theoretical security guarantees, computational efficiency, and effectiveness under different attack scenarios as reported in empirical studies. Visual models, including workflow diagrams and attack taxonomies, are generated using tools like Mermaid.js and LaTeX to illustrate the Federated Learning process and threat surfaces.

Furthermore, we conduct a simplified mathematical analysis using federated averaging to simulate local model updates and aggregation, demonstrating the potential manipulation by malicious clients and how weighted or simple averaging affects the global model. The findings are synthesized to highlight existing research gaps and to formulate recommendations.

B. Literature Search and Selection Procedures

In order to carry out a thorough study of the present issues concerning the security in Federated Learning (FL), a structured literature search strategy was deployed. The main objective was to locate academic papers that cover the topic of attacks and defense methods in the frameworks of FL, but with a preference for empirical, theoretical, and implementation-based investigations. We accessed prominent academic databases like IEEE Xplore, SpringerLink, ScienceDirect, ACM Digital Library, and arXiv. In addition, Google Scholar was also used as an augmented search tool for collecting unpublished information sources (grey literature) and the very newest papers. It all started with running a search through the databases using a list of keywords and logical operators: "Federated Learning," "FL attacks," "model poisoning," "Byzantine attacks," "gradient inversion," "Sybil attacks," "FL security," "secure aggregation," and "trusted execution environments." Were the document types limited to include only journal articles, conference proceedings, and preprints between 2018 and 2025. Only the documents that passed the checks for acceptance were considered valid for the search.

As shown in the PRISMA-style flowchart, we found an initial pool of 250 records. After the removal of 40 repeated articles, 210 unique articles were left. Via a screening process which comprised of the titles and abstracts to assess relevance, 100 records were eliminated as either irrelevant or redundant. The remaining 110 full-text articles were examined for eligibility based on, for example, the study focus, clear methodology, and relevance to federated learning security. Right from these, 40 articles were found to be not suitable for the inclusion criteria and thus, 65 articles were the ones to be included in the final synthesis.

➢ *Inclusion and Exclusion Criteria*

- *Inclusion Criteria were:*

✓ Peer-reviewed journal and conference papers.
✓ Publications between 2018–2025.
✓ Studies focusing on attacks and countermeasures in FL systems.

- *Exclusion Criteria were:*

✓ Non-English publications.
✓ Articles lacking technical or empirical evidence.
✓ Duplicate records and irrelevant studies based on title/abstract review.

In addition, a snowballing method was used to review the bibliographies of the key articles to locate additional studies that are relevant to the review. The studies that were chosen were then categorized to the four main themes of attacks, defenses, theories, and practical applications. Consequently, this systematic search of the literature imparted a comprehensive understanding of the security of Federated Learning and paved the way for research gaps, attack taxonomies, and defense evaluations which are benefits of this study.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may617

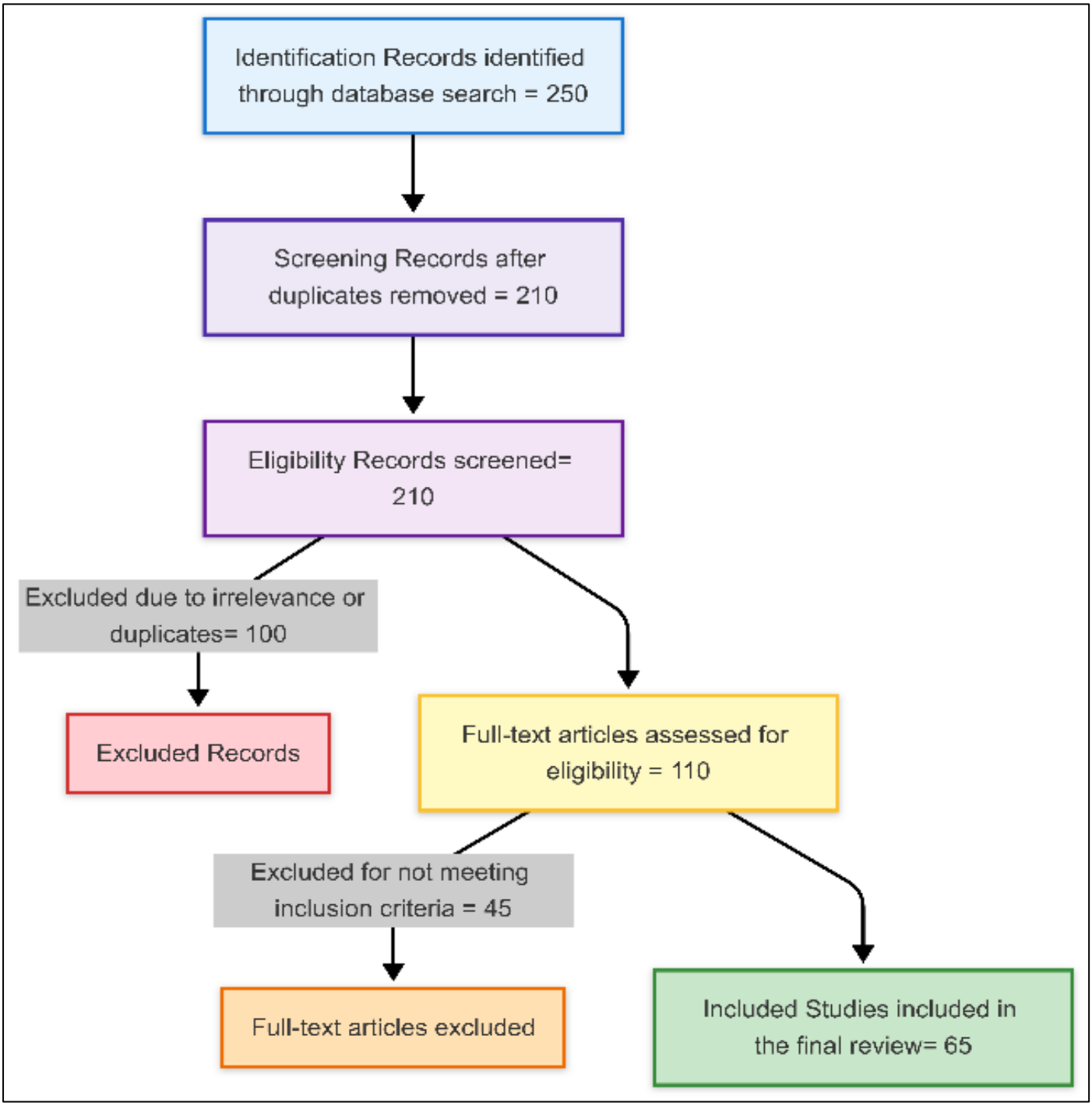

Fig 8 Literature Search and Selection Procedures

V. FINDINGS

The analysis of recent literature on federated learning (FL) reveals that while FL offers substantial benefits in preserving data privacy and enabling distributed model training, it is significantly vulnerable to a diverse range of adversarial attacks. The key findings of this study are centred on the types of attacks, their operational mechanisms, and the corresponding gaps in existing defence strategies.

- We observed that poisoning attacks both data and model-based represent the most pervasive threats in FL environments. These attacks exploit the lack of direct oversight in local training by allowing malicious participating clients to inject corrupted data or manipulate gradient updates. Model poisoning, in particular, can be subtle yet highly disastrous, leading to backdoor models that maintain overall performance while misclassifying inputs containing specific triggers.
- The study finds that privacy-centric attacks, such as gradient inversion and membership inference, present a serious challenge to FL's core privacy promise. Even when raw data is never shared, gradient updates and intermediate feature maps can leak sensitive information about the training data. Du et al. (2023) demonstrated that attackers could reconstruct private inputs from shared

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may617

gradients, especially in the absence of privacy-preserving mechanisms like differential privacy.

- The stealth-based attacks such as clean-label poisoning and free-rider behaviours emerge as particularly difficult to detect due to their ability to blend seamlessly with normal client operations. These attacks often circumvent standard anomaly detection measures by maintaining consistency in labels and statistical properties while still compromising the integrity or fairness of the global model. Moreover, the study highlights the systemic vulnerability to Sybil attacks, where a single adversary controls multiple clients. This amplifies malicious influence in model aggregation, especially in open FL systems where client authentication is minimal or absent.
- The review finds that existing defence mechanisms, while effective in isolation, are often insufficient when facing composite or multi-phase attacks. There is a clear need for integrated defence frameworks that combine robust aggregation, privacy-preserving updates, behavioural monitoring, and adaptive security protocols to protect against a broad array of adversarial strategies.

These findings highlight the importance of rethinking federated learning architecture and trust models, moving toward a more secure-by-design approach that anticipates both active and passive adversarial behaviours.

VI. DISCUSSION

The findings of this study emphasis a critical issue in federated learning (FL): while FL is designed to enhance privacy and decentralization, these very features also broaden its vulnerability landscape. This duality has significant implications for how FL systems are conceptualized, deployed, and secured, particularly in sensitive areas such as healthcare, finance, and IoT systems.

The prevalence of poisoning attacks, as documented in the literature study, suggests that FL's dependence on unverified local computations is still a serious security vulnerability. FL assigns trust to distributed clients, many of whom may be malicious or compromised, in contrast to centralised solutions where data pipelines can be managed and observed. The insufficiency of traditional validation or aggregation approaches, which commonly overlook deviously written updates, is further exposed by the stealth and sophistication of model poisoning and backdoor attacks. This requires incorporating more sophisticated anomaly detection systems that take into account the contextual behaviour of updates over time as well as their statistical distribution.

The results further refute the generally accepted but inaccurate view that FL automatically ensures data privacy. Attacks that use model transparency to extract sensitive data, such as gradient inversion and membership inference, demonstrate that issues with privacy exist even in cases where raw data is never transmitted. This emphasises the necessity of implementing more robust formal privacy guarantees, like secure aggregation or differential privacy, as essential parts of FL systems rather than as add-ons.

The report also highlights the challenge to defend against composite or multifaceted attacks. For example, Sybil attacks and clean-label poisoning combine to create a powerful threat that threatens model availability, fairness, and integrity concurrently. The way these threats are now addressed in the defence system frequently results in brittle systems that are simple for adaptive adversaries to get around. This indicates that comprehensive, layered security systems that incorporate behavioural, statistical, and cryptographic protections are essential.

Equally important is the socio-technical consideration of FL deployment. The assumption of honest clients and an honest-but-curious server may not hold in real-world implementations, particularly in open or large-scale federations. Therefore, revisiting FL's trust model potentially by introducing reputation systems, authenticated participation, and trust score adjustments could play a crucial role in mitigating long-term systemic risks.

VII. RECOMMENDATIONS

Based on the comprehensive analysis of existing literature on federated learning (FL) attacks and the evaluation of current defence mechanisms, we make the following recommendations to enhance the security and resilience of future FL systems:

- **Integration of Multi-layered Defence Mechanisms***:* Single-point defence system such as robust aggregation or differential privacy are insufficient against sophisticated attacks. It is recommended that federated learning frameworks incorporate multi-layered security mechanisms that combine cryptographic protections, anomaly detection, behavioural auditing, and adaptive trust management. Integrating these layers can better defend against composite threats like model poisoning combined with Sybil attacks.
- **Mandatory Privacy-Preserving Techniques**: Privacy-preserving systems like differential privacy, secure aggregation, and homomorphic encryption need to be not only considered as optional but rather deeply integrated into FL system architectures for guarantees against gradient inversion and membership inference threats, as illustrated by Du et al. (2023).
- **Dynamic Client Trust and Reputation Systems**: To mitigate Sybil attacks and free-rider behaviours, it is recommended to implement client reputation scoring and trust evaluation mechanisms. Clients should earn influence in model aggregation based on consistent, verified, and honest participation rather than through static system parameters.
- **Development of Standardized Testing and Validation Protocols***:*Standard adversarial testing protocols that are uniform should be formed to periodically assess the durability of federated models that are secured against recognized attack types, such as the poisoning, backdoor, and clean-label attacks. This proactive approach can identify vulnerabilities before real-world exploitation.
- **Promotion of Explainability and Transparency in FL**: Enhancing the interpretability of client contributions and

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may617

model behaviour through explainable AI (XAI) techniques can improve anomaly detection and foster greater trust in federated systems.

- **Investment in Lightweight Security Solutions:** Given the resource constraints of many federated clients (e.g., mobile devices, IoT), future research should prioritize the development of computationally efficient security mechanisms that do not compromise model performance or training scalability.
- **Re-examination of FL Deployment Models***:* Practitioners must revisit and strengthen assumptions about trust, client authentication, and participation policies before deploying FL systems, especially in open or heterogeneous environments where adversarial participation is likely.

VIII. CONCLUSION

This study critically examined the diverse spectrum of attacks in federated learning (FL) and its corresponding defence mechanisms, drawing insights from recent scholarly works. The review of related literature revealed that FL, while promoting privacy and decentralisation, introduces many different vulnerabilities to poisoning, inference, and stealth-based attacks. These threats exploit the lack of centralised oversight and the openness of communication protocols, which compromise both the integrity and confidentiality of model training. Key findings emphasised the limitations of existing defence mechanisms, which are often effective in isolation but insufficient against composite or adaptive attacks. The study further highlighted the need for integrated, multi-layered defence frameworks that combine privacy-preserving techniques, robust aggregation, and behavioural validation to secure FL deployments. Moreover, the discussion emphasis the necessity of revising FL's trust assumptions and incorporating dynamic reputation systems to strengthen its resilience. To conclude, securing federated learning is not a matter of isolated technical patches but requires a holistic, secure-by-design approach that anticipates adversarial innovation and prioritises long-term trustworthiness.

REFERENCES

[1]. Aggarwal, M., Khullar, V., Rani, S., Thomas André Prola, Shyama Barna Bhattacharjee, Sarowar Morshed Shawon and Goyal, N. (2024). Federated Learning on Internet of Things: Extensive and Systematic Review. *Computers, materials & continua/Computers, materials & continua (Print)*, 0(0), pp.1–10. doi:https://doi.org/10.32604/cmc.2024.049846.

[2]. Aljanabi, M., Omran, A.H., Mijwil, M.M., Mostafa , A., El-kenawy, E.-S.M., Yousif Mohammed, S. and Ibrahim, A. (2023). Data poisoning: issues, challenges, and needs. doi:https://doi.org/10.1049/icp.2024.0951.

[3]. Almutairi, S. and Barnawi, A. (2023). Federated learning vulnerabilities, threats and defenses: A systematic review and future directions. *Internet of Things*, 24, pp.100947–100947. doi:https://doi.org/10.1016/j.iot.2023.100947.

[4]. Araki, T., Furukawa, Y., Lindell, Y., Nof, A. and Ohara, K. (2016). High-Throughput Semi-Honest Secure Three-Party Computation with an Honest Majority. doi:https://doi.org/10.1145/2976749.2978331.

[5]. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D. and Shmatikov, V. (2019). How To Backdoor Federated Learning. *arXiv:1807.00459 [cs]*. [online] Available at: https://arxiv.org/abs/1807.00459.

[6]. Benmalek, M., Benrekia, M.A. and Challal, Y. (2022). Security of Federated Learning: Attacks, Defensive Mechanisms, and Challenges. *Revue d'Intelligence Artificielle*, 36(1), pp.49–59. doi:https://doi.org/10.18280/ria.360106.

[7]. Betul Yurdem, Murat Kuzlu, Mehmet Kemal Gullu, Ferhat Ozgur Catak and Tabassum, M. (2024). Federated Learning: Overview, Strategies, Applications, Tools and Future Directions. *Heliyon*, [online] 10(19), pp.e38137–e38137. doi:https://doi.org/10.1016/j.heliyon.2024.e38137.

[8]. Betul Yurdem, Murat Kuzlu, Mehmet Kemal Gullu, Ferhat Ozgur Catak and Tabassum, M. (2024). Federated Learning: Overview, Strategies, Applications, Tools and Future Directions. *Heliyon*, [online] 10(19), pp.e38137–e38137. doi:https://doi.org/10.1016/j.heliyon.2024.e38137.

[9]. Bhatti, D.M.S., Ali, M., Yoon, J. and Choi, B.J. (2025). Efficient Collaborative Learning in the Industrial IoT Using Federated Learning and Adaptive Weighting Based on Shapley Values. *Sensors*, 25(3), p.969. doi:https://doi.org/10.3390/s25030969.

[10]. Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečný, J., Mazzocchi, S., McMahan, H.B., Van Overveldt, T., Petrou, D., Ramage, D. and Roselander, J. (2019). *Towards Federated Learning at Scale: System Design*. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.1902.01046.

[11]. Brisimi, T.S., Chen, R., Mela, T., Olshevsky, A., Paschalidis, I.Ch. and Shi, W. (2018). Federated learning of predictive models from federated Electronic Health Records. *International Journal of Medical Informatics*, [online] 112, pp.59–67. doi:https://doi.org/10.1016/j.ijmedinf.2018.01.007.

[12]. Cao, X., Fang, M., Liu, J. and Gong, N.Z. (2022). FLTrust: Byzantine-robust Federated Learning via Trust Bootstrapping. *arXiv:2012.13995 [cs]*. [online] Available at: https://arxiv.org/abs/2012.13995 [Accessed 8 Aug. 2022].

[13]. Chen, C., Liu, J., Tan, H., Li, X., Wang, K.I-Kai., Li, P., Sakurai, K. and Dou, D. (2024a). Trustworthy Federated Learning: Privacy, Security, and Beyond. *arXiv (Cornell University)*. doi:https://doi.org/10.48550/arxiv.2411.01583.

[14]. Chen, D., Yu, N., Zhang, Y. and Fritz, M. (2020). GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models. *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, [online] pp.343–362. doi:https://doi.org/10.1145/3372297.3417238.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may617

[15]. Chen, J., Li, M., Liu, T., Zheng, H., Du, H. and Cheng, Y. (2024b). Rethinking the defense against free-rider attack from the perspective of model weight evolving frequency. *Information Sciences*, 668, pp.120527–120527. doi:https://doi.org/10.1016/j.ins.2024.120527.

[16]. Criado, M.F., Casado, F.E., Iglesias, R., Regueiro, C.V. and Barro, S. (2022). Non-IID data and Continual Learning processes in Federated Learning: A long road ahead. *Information Fusion*, 88, pp.263–280. doi:https://doi.org/10.1016/j.inffus.2022.07.024.

[17]. Ding, X., Liu, Z., You, X., Li, X. and Vasilakos, A.V. (2024). Improved gradient leakage attack against compressed gradients in federated learning. *Neurocomputing*, [online] 608, p.128349. doi:https://doi.org/10.1016/j.neucom.2024.128349.

[18]. Douceur, J.R. (2002). The Sybil Attack. *Peer-to-Peer Systems*, pp.251–260. doi:https://doi.org/10.1007/3-540-45748-8_24.

[19]. Enthoven, D. and Al-Ars, Z. (2021). An Overview of Federated Deep Learning Privacy Attacks and Defensive Strategies. pp.173–196. doi:https://doi.org/10.1007/978-3-030-70604-3_8.

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

[21]. Ge, L., Li, H., Wang, X. and Wang, Z. (2023). A review of secure federated learning: privacy leakage threats, protection technologies, challenges and future directions. *Neurocomputing*, [online] p.126897. doi:https://doi.org/10.1016/j.neucom.2023.126897.

[22]. Gong, Y., Wang, S., Yu, T., Jiang, X. and Sun, F. (2024). Improving adversarial robustness using knowledge distillation guided by attention information bottleneck. *Information Sciences*, [online] 665, p.120401. doi:https://doi.org/10.1016/j.ins.2024.120401.

[23]. Guo, P., Zeng, S., Chen, W., Zhang, X., Ren, W., Zhou, Y. and Qu, L. (2024). *A New Federated Learning Framework Against Gradient Inversion Attacks*. [online] arXiv.org. Available at: https://arxiv.org/abs/2412.07187 [Accessed 20 Apr. 2025].

[24]. Hu, K., Gong, S., Zhang, Q., Seng, C., Xia, M. and Jiang, S. (2024). An overview of implementing security and privacy in federated learning. *Artificial intelligence review*, 57(8). doi:https://doi.org/10.1007/s10462-024-10846-8.

[25]. Huang, L., Joseph, A.D., Nelson, B., Rubinstein, B.I.P. and Tygar, J.D. (2011). Adversarial machine learning. *Proceedings of the 4th ACM workshop on Security and artificial intelligence - AISec '11*. doi:https://doi.org/10.1145/2046684.2046692.

[26]. Huang, Y., Chu, L., Zhou, Z., Wang, L., Liu, J., Pei, J. and Zhang, Y. (2021). Personalized Cross-Silo Federated Learning on Non-IID Data. *arXiv:2007.03797 [cs, stat]*. [online] Available at: https://arxiv.org/abs/2007.03797.

[27]. Huang, Y., Gupta, S., Song, Z., Arora, S. and Li, K. (2024). Evaluating gradient inversion attacks and defenses. *Federated Learning*, pp.105–122. doi:https://doi.org/10.1016/b978-0-44-319037-7.00014-4.

[28]. Huang, Y., Huo, Z. and Fan, Y. (2024). DRA: A data reconstruction attack on vertical federated k-means clustering. *Expert Systems with Applications*, 250, pp.123807–123807. doi:https://doi.org/10.1016/j.eswa.2024.123807.

[29]. Jebreel, Najeeb Moharram, Domingo-Ferrer, J., Sánchez, D. and Blanco-Justicia, A. (2022). *Defending against the Label-flipping Attack in Federated Learning*. [online] arXiv.org. Available at: https://arxiv.org/abs/2207.01982 [Accessed 7 Apr. 2025].

[30]. Kasyap, H. and Tripathy, S. (2024). Beyond data poisoning in federated learning. *Expert Systems with Applications*, 235, pp.121192–121192. doi:https://doi.org/10.1016/j.eswa.2023.121192.

[31]. Lamport, L. (1983). The Weak Byzantine Generals Problem. *Journal of the ACM*, [online] 30(3), pp.668–676. doi:https://doi.org/10.1145/2402.322398.

[32]. Lenaerts-Bergmans, B. (2024). *What Is Data Poisoning? | CrowdStrike*. [online] Crowdstrike.com. Available at: https://www.crowdstrike.com/en-us/cybersecurity-101/cyberattacks/data-poisoning/.

[33]. Li, Z., Huang, X., Li, Y. and Chen, G. (2023). A comparative study of adversarial training methods for neural models of source code. *Future Generation Computer Systems*, 142, pp.165–181. doi:https://doi.org/10.1016/j.future.2022.12.030.

[34]. Liang, T., Glossner, J., Wang, L., Shi, S. and Zhang, X. (2021). Pruning and quantization for deep neural network acceleration: A survey. *Neurocomputing*, 461, pp.370–403. doi:https://doi.org/10.1016/j.neucom.2021.07.045.

[35]. Liman, M.D., Osanga, S.I., Alu, E.S. and Zakariya, S. (2024). Regularization Effects in Deep Learning Architecture. *Journal of the Nigerian Society of Physical Sciences*, p.1911. doi:https://doi.org/10.46481/jnsps.2024.1911.

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

[37]. Lutho Ntantiso, Bagula, A.B., Ajayi, O. and Ngongo, F.K. (2023). *A Review of Federated Learning: Algorithms, Frameworks & Applications*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/369417303_A_Review_of_Federated_Learning_Algorithms_Frameworks_Applications [Accessed 6 May 2025].

[38]. Lyu, L., Yu, H. and Yang, Q. (2020a). Threats to Federated Learning: A Survey. *arXiv:2003.02133 [cs, stat]*. [online] Available at: https://arxiv.org/abs/2003.02133.

[39]. Lyu, L., Yu, H. and Yang, Q. (2020b). *Threats to Federated Learning: A Survey*. [online] Available at: https://arxiv.org/pdf/2003.02133.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may617

[40]. McMahan, H.B., Blaise, Ramage, D., Moore, E. and Hampson, S. (2023). *Communication-Efficient Learning of Deep Networks from Decentralized Data*. [online] alphaXiv. Available at: https://www.alphaxiv.org/abs/1602.05629 [Accessed 30 Apr. 2025].

[41]. Naik, D. and Naik, N. (2024). An Introduction to Federated Learning: Working, Types, Benefits and Limitations. *Advances in intelligent systems and computing*, pp.3–17. doi:https://doi.org/10.1007/978-3-031-47508-5_1.

[42]. Nanayakkara, S.I., Pokhrel, S.R. and Li, G. (2024). Understanding global aggregation and optimization of federated learning. *Future Generation Computer Systems*, 159, pp.114–133. doi:https://doi.org/10.1016/j.future.2024.05.009.

[43]. Qayyum, A., Janjua, M.U. and Qadir, J. (2022). Making federated learning robust to adversarial attacks by learning data and model association. *Computers & Security*, 121, p.102827. doi:https://doi.org/10.1016/j.cose.2022.102827.

[44]. Qi, P., Chiaro, D., Guzzo, A., Ianni, M., Fortino, G. and Piccialli, F. (2024). Model aggregation techniques in federated learning: A comprehensive survey. *Future Generation Computer Systems*, 150, pp.272–293. doi:https://doi.org/10.1016/j.future.2023.09.008.

[45]. Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*. [online] Available at: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf.

[46]. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. [online] Available at: https://storage.prod.researchhub.com/uploads/papers/2020/06/01/language-models.pdf.

[47]. Ribero, M., Vikalo, H. and de Veciana, G. (2025). Federated Learning at Scale: Addressing Client Intermittency and Resource Constraints. *IEEE Journal of Selected Topics in Signal Processing*, 19(1), pp.60–73. doi:https://doi.org/10.1109/jstsp.2024.3430118.

[48]. Sagar, S., Li, C.-S., Loke, S.W. and Choi, J. (2023). *Poisoning Attacks and Defenses in Federated Learning: A Survey*. [online] arXiv.org. Available at: https://arxiv.org/abs/2301.05795 [Accessed 7 Apr. 2025].

[49]. Sheller, M.J., Edwards, B., Reina, G.A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R.R. and Bakas, S. (2020). Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, [online] 10(1), p.12598. doi:https://doi.org/10.1038/s41598-020-69250-1.

[50]. Struppek, L., Hintersdorf, D., Friedrich, F., Brack, M., Schramowski, P. and Kersting, K. (2023). Class Attribute Inference Attacks: Inferring Sensitive Class Information by Diffusion-Based Attribute Manipulations. *arXiv (Cornell University)*. doi:https://doi.org/10.48550/arxiv.2303.09289.

[51]. Sun, T., Li, D. and Wang, B. (2023). Decentralized Federated Averaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, [online] 45(4), pp.4289–4301. doi:https://doi.org/10.1109/TPAMI.2022.3196503.

[52]. Truong, N., Sun, K., Wang, S., Guitton, F. and Guo, Y. (2021). Privacy preservation in federated learning: An insightful survey from the GDPR perspective. *Computers & Security*, [online] 110, p.102402. doi:https://doi.org/10.1016/j.cose.2021.102402.

[53]. Verde, L., Marulli, F. and Marrone, S. (2021). Exploring the Impact of Data Poisoning Attacks on Machine Learning Model Reliability. *Procedia Computer Science*, [online] 192, pp.2624–2632. doi:https://doi.org/10.1016/j.procs.2021.09.032.

[54]. Vungarala, S.K. (2023). *Stochastic gradient descent vs Gradient descent — Exploring the differences*. [online] Medium. Available at: https://medium.com/@seshu8hachi/stochastic-gradient-descent-vs-gradient-descent-exploring-the-differences-9c29698b3a9b.

[55]. Wang, B., Li, H., Liu, X. and Guo, Y. (2024). FRAD: Free-Rider Attacks Detection Mechanism for Federated Learning in AIoT. *IEEE Internet of Things Journal*, 11(3), pp.4377–4388. doi:https://doi.org/10.1109/jiot.2023.3298606.

[56]. Wei, Q. and Rao, G. (2024). EPFL-DAC: Enhancing Privacy in Federated Learning with Dynamic Aggregation and Clipping. *Computers & Security*, 143, p.103911. doi:https://doi.org/10.1016/j.cose.2024.103911.

[57]. Wu, X., Chen, Y., Yu, H. and Yang, Z. (2024). Privacy-preserving federated learning based on noise addition. *Expert Systems with Applications*, [online] 267, p.126228. doi:https://doi.org/10.1016/j.eswa.2024.126228.

[58]. Xie, X., Hu, C., Ren, H. and Deng, J. (2024). A survey on vulnerability of federated learning: A learning algorithm perspective. *Neurocomputing*, 573, pp.127225–127225. doi:https://doi.org/10.1016/j.neucom.2023.127225.

[59]. Xu, Z., Zhang, Y., Andrew, G., Choquette, C., Kairouz, P., Mcmahan, B., Rosenstock, J. and Zhang, Y. (2023). Federated Learning of Gboard Language Models with Differential Privacy. *arXiv (Cornell University)*. doi:https://doi.org/10.18653/v1/2023.acl-industry.60.

[60]. Yang, H., Wang, Z., Chou, B., Xu, S., Wang, H., Wang, J. and Zhang, Q. (2019). *An Empirical Study of the Impact of Federated Learning on Machine Learning Model Accuracy*. [online] Arxiv.org. Available at: https://arxiv.org/html/2503.20768v1 [Accessed 5 May 2025].

[61]. Yang, M., Cheng, H., Chen, F., Liu, X., Wang, M. and Li, X. (2023). Model poisoning attack in differential privacy-based federated learning. *Information Sciences*, [online] 630, pp.158–172. doi:https://doi.org/10.1016/j.ins.2023.02.025.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may617

[62]. Yin, Y., Chen, H., Gao, Y., Sun, P., Wu, L., Li, Z., & Liu, W. (2024). Feature-based Full-target Clean-label Backdoor Attacks. Applied Intelligence.

[63]. Zeng, D., Wu, Z., Liu, S., Pan, Y., Tang, X. and Xu, Z. (2024). Understanding Generalization of Federated Learning: the Trade-off between Model Stability and Optimization. *arXiv (Cornell University)*. doi:https://doi.org/10.48550/arxiv.2411.16303.

[64]. Zhang, C., Yang, S., Mao, L. and Ning, H. (2024a). Anomaly detection and defense techniques in federated learning: a comprehensive review. *Artificial intelligence review*, 57(6). doi:https://doi.org/10.1007/s10462-024-10796-1.

[65]. Zhang, W., Yu, C., Meng, Z., Shen, S. and Zhang, K. (2024b). Explore Patterns to Detect Sybil Attack during Federated Learning in Mobile Digital Twin Network. *ICC 2022 - IEEE International Conference on Communications*, pp.3969–3974. doi:https://doi.org/10.1109/icc51166.2024.10622975.

[66]. Zhang, X., Chen, C., Xie, Y., Chen, X., Zhang, J. and Xiang, Y. (2023). A survey on privacy inference attacks and defenses in cloud-based Deep Neural Network. *Computer Standards & Interfaces*, 83, p.103672. doi:https://doi.org/10.1016/j.csi.2022.103672.

[67]. Zhu, L., Liu, Z. and Han, S. (2019). Deep Leakage from Gradients. *arXiv (Cornell University)*. doi:https://doi.org/10.48550/arxiv.1906.08935.

[68]. Zi, B., Agrawal, A., Coburn, C., Asghar, H.J., Bhaskar, R., Kaafar, M.A., Webb, D. and Dickinson, P. (2021). On the (In)Feasibility of Attribute Inference Attacks on Machine Learning Models. *arXiv (Cornell University)*, pp.232–251. doi:https://doi.org/10.1109/eurosp51992.2021.00025.