The Impact of Adversarial Attacks on Federated Learning: A Survey

K Naveen Kumar, C Krishna Mohan, *Senior Member, IEEE,* Linga Reddy Cenkeramaddi, *Senior Member, IEEE*

***Abstract*—Federated learning (FL) has emerged as a powerful machine learning technique that enables the development of models from decentralized data sources. However, the decentralized nature of FL makes it vulnerable to adversarial attacks. In this survey, we provide a comprehensive overview of the impact of malicious attacks on FL by covering various aspects such as attack budget, visibility, and generalizability, among others. Previous surveys have primarily focused on the multiple types of attacks and defenses but failed to consider the impact of these attacks in terms of their budget, visibility, and generalizability. This survey aims to fill this gap by providing a comprehensive understanding of the attacks' effect by identifying FL attacks with low budgets, low visibility, and high impact. Additionally, we address the recent advancements in the field of adversarial defenses in FL and highlight the challenges in securing FL. The contribution of this survey is threefold: first, it provides a comprehensive and up-to-date overview of the current state of FL attacks and defenses. Second, it highlights the critical importance of considering the impact, budget, and visibility of FL attacks. Finally, we provide ten case studies and potential future directions towards improving the security and privacy of FL systems.**

***Index Terms*—Federated learning, adversarial attacks, impact, budget, visibility, generalizability, real-world application domains, attacks & defenses, attack status, online & offline attacks, and security challenges.**

I. Introduction

Federated Learning (FL) [1] is a decentralized machine learning (ML) paradigm [2], [3] that allows model development using data from multiple parties while preserving their data ownership. This approach is particularly valuable for sensitive data, as it ensures user privacy [4]. It was initially introduced by Google researchers in 2016 [5], proposing a framework to train ML models on decentralized data sources like mobile devices without compromising privacy. Over time, FL has gained popularity as a privacy-preserving ML technique, witnessing substantial research and real-world applications [6], [7].

Threats in FL: The decentralized nature of FL exposes it to adversarial attacks [8], [9], categorized as utility-centric and privacy-centric threats [10]. Utility-centric threats involve poisoning data or models and compromising the accuracy [11]–[13]. Privacy-centric threats [14]–[16] relate to participant data exposure, risking the leakage of sensitive information and eroding trust in the FL system. Further, the attacks are divided into two categories based on the source of an attack, namely, a causative (training time) or evasion (test time), as shown in Figure 1. In causative attacks, a malicious participant

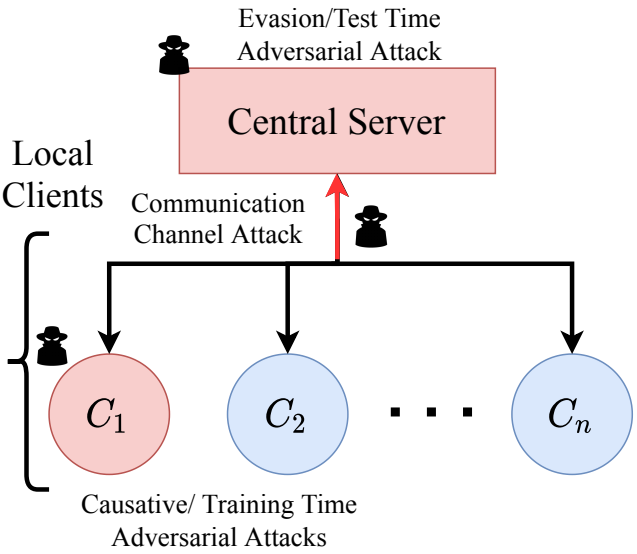


Fig. 1: Overview of federated learning with its vulnerabilities.

submits poisonous data or tampers with model updates during training that skew the model's predictions [13], [17], [18]. On the contrary, evasion attacks [19], [20] manipulate the model's predictions by modifying the input test data during inference. Both causative and evasion attacks pose significant threats to the accuracy and privacy of FL models. In addition, a communication channel attack is another threat that occurs when an adversary intercepts sensitive data or manipulates the communication between clients and server. One example of a communication channel attack is the Man-in-the-Middle (MitM) [21] attack, in which an attacker intercepts the communication channel and manipulates the data exchanged between clients and server during the training process, compromising the privacy and security of the FL model.

Furthermore, adversarial attacks on FL can take various forms, including targeted, untargeted, backdoor poisoning, model inversion, and membership inference. These attacks can consistently attack (online) or attack at the beginning of the FL process (offline). In addition, there are three attack settings based on the adversary's knowledge and capabilities: white-box, grey-box, and black-box. Under utility-centric attacks, only black-box offline data poisoning and white-box online model poisoning settings are practical for production FL, according to Shejwalkar *et al.* [11].

K Naveen Kumar and C Krishna Mohan, are with the Department of Computer Science and Engineering, Indian Institute of Technology, Hyderabad, India. (e-mails: cs19m20p000001@iith.ac.in, ckm@cse.iith.ac.in).

Linga Reddy Cenkeramaddi is with the Department of Information and Communication Technology, University of Agder, Grimstad, Norway. (e-mail: linga.cenkeramaddi@uia.no)

TABLE I: Comparison of Key Characteristics and Attributes between Existing State-of-the-Art Surveys and Our Survey in the Field of Adversarial Attacks on FL

		Scope										
S.No	**Survey Paper, Year**	**Attack vs. FL Setting**	**Attack Impact**	**Attack Budget**	**Attack Visibility**	**Attack Status**	**Attack Setting**	**Attack vs. General -izability**	**Attack Intention**	**Attack vs. Defense**	**Attack vs. Appli- cation Domains**	**Commu- nication Channel Attacks**
1	Lim *et al.* [22], 2020	✖	✔	✖	✖	✖	✖	✖	✔	✖	✔	✔
2	Zhang *et al.* [23], 2021	✖	✔	✖	✖	✖	✖	✖	✔	✔	✖	✖
3	Usynin *et al.* [10], 2021	✖	✔	✖	✖	✔	✔	✖	✔	✔	✔	✖
4	Yang *et al.* [24], 2021	✖	✔	✖	✖	✖	✖	✖	✔	✔	✔	✖
5	Bouacida *et al.* [25] , 2021	✖	✔	✖	✖	✖	✔	✖	✔	✔	✔	✔
6	Blanco *et al.* [26] , 2021	✖	✔	✖	✖	✖	✔	✖	✔	✔	✖	✔
7	Mothukuri *et al.* [27], 2021	✖	✔	✖	✖	✖	✔	✖	✔	✔	✖	✔
8	Jere *et al.* [28], 2021	✖	✔	✖	✖	✖	✔	✖	✔	✔	✖	✖
9	Yoo *et al.* [29], 2021	✖	✔	✖	✖	✖	✖	✖	✔	✔	✔	✖
10	Liu *et al.* [30], 2022	✔	✔	✖	✔	✖	✔	✖	✔	✔	✖	✔
11	Shejwalkar *et al.* [11], 2022	✖	✔	✖	✔	✔	✔	✖	✖	✔	✔	✖
12	Wang *et al.* [31], 2022	✖	✔	✖	✖	✖	✖	✖	✖	✔	✖	✖
13	Chen *et al.* [32], 2022	✔	✔	✖	✔	✖	✔	✖	✔	✔	✖	✖
14	Lyu *et al.* [33], 2022	✔	✔	✖	✖	✔	✔	✖	✔	✔	✖	✔
15	Jatain *et al.* [34], 2022	✖	✔	✖	✖	✖	✖	✖	✔	✔	✖	✖
16	Zhang *et al.* [35], 2022	✖	✔	✖	✖	✖	✖	✖	✔	✖	✖	✖
17	Qammar *et al.* [36], 2022	✖	✔	✖	✔	✖	✔	✖	✔	✔	✖	✔
18	Ghimire *et al.* [37], 2022	✖	✔	✖	✔	✖	✖	✖	✖	✔	✔	✔
19	Benmalek *et al.* [38], 2022	✖	✔	✖	✖	✖	✖	✖	✔	✔	✖	✔
20	Zhang *et al.* [39] , 2022	✖	✔	✖	✔	✖	✖	✖	✔	✔	✖	✖
21	Rodriguez *et al.* [40], 2023	✔	✔	✖	✖	✖	✔	✖	✔	✔	✖	✔
22	Xia *et al.* [41] , 2023	✖	✔	✖	✔	✖	✖	✖	✔	✔	✖	✖
23	Nair *et al.* [42], 2023	✖	✔	✖	✔	✖	✔	✖	✔	✖	✖	✔
24	Hallaji *et al.* [43], 2023	✖	✔	✖	✖	✖	✔	✖	✔	✔	✖	✔
25	Sikandar *et al.* [44], 2023	✖	✔	✖	✔	✔	✔	✖	✔	✔	✖	✖
26	**Our Survey**, 2023	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔

Limitations of Existing Surveys: Table I highlights the gaps in existing surveys and outlines the broad scope of our survey. While some previous surveys have considered the topic of attack visibility, they have not comprehensively addressed the attack impact, the attack budget, and the attack visibility as a **combined effect** on FL systems. The existing surveys primarily focused on individual aspects or a subset of the mentioned factors without providing a **united holistic analysis** of their interplay. Therefore, our survey aims to bridge this gap by providing a more comprehensive understanding of the potential influence of adversarial attacks on FL, considering the attack impact, budget, and visibility collectively. In addition, many existing surveys have not addressed the attacks' status, such as online or offline, and their implications on FL system security. Hence, we will cover various aspects such as attack impact, budget, visibility, generalizability, case studies of attacks on different real-world application domains, and the effect of adversarial attacks on FL types, such as horizontal, vertical FL, and federated transfer learning.

Furthermore, this survey paper focuses on identifying FL attacks with low budgets, low visibility, and high impact, which can be challenging to detect and prevent, leading to severe consequences for FL systems. In addition, we highlight recent advancements in adversarial defenses in FL and address

the challenges in securing FL. Our survey is a comprehensive resource for understanding the impact of malicious attacks on FL. It provides insights into current challenges and future research directions toward trustworthy FL systems for decentralized data sources.

Our Contributions: The contributions of this survey are as follows:

- **Comprehensive overview:** The survey provides a comprehensive and up-to-date overview of the current state of FL attacks and defenses, covering various aspects such as attack impact, budget, visibility, generalizability, attack status, and attack metrics with performance values.
- **Identification of FL attacks:** We focus on understanding and identifying FL attacks with low budgets, low visibility, and high impact, as such attacks can be more difficult to detect and can result in severe consequences for FL systems.
- **Categories of FL:** We discuss adversarial attacks on FL categories such as horizontal FL, vertical FL, and federated transfer learning to understand the impact.
- **Communication channel attacks:** The survey briefly discusses the attacks which target the communication channel between the FL clients and the server.
- **Adversarial defenses:** The survey discusses recent advancements in the field of adversarial defenses in FL, their categories, and metrics.
- **Real-world case studies:** This study presents ten comprehensive case studies that assess the applicability of current FL attacks and their limitations. Further, we propose potential solutions to address these limitations effectively.
- **Highlighting research gaps and potential future directions:** At the end, we highlight and identify ten research gaps and potential future directions in securing and providing privacy to the FL systems.

Paper Outline: The rest of this paper is organized as follows. Section II provides a background of FL, types of FL based on data distribution, and threat models. Section III covers the different types of adversarial attacks that can be launched against FL systems. It includes an analysis of the impact of these attacks in terms of their budget, visibility, and generalizability. Section IV discusses the different defense strategies to protect FL systems from adversarial attacks. Section V covers the different evaluation metrics used for assessing the effectiveness of malicious attacks in FL systems. Section VI discusses the possible research gaps and areas to explore in the field of FL attacks. Finally, section VII provides the summary and concluding remarks of the paper.

II. Background

1) Federated Learning: It consists of a server and n clients, each with its own local private data. The objective is to learn the global model W that exhibits good performance on the global test data. During each epoch t, the server sends the W_t model to all clients. Subsequently, each client trains it on their respective data and calculates the update. Then the server receives the individual updates and aggregates using synchronous federated weighted averaging (FedAvg) [5] to obtain W_{t+1}. This process continues iteratively until the global model converges.

2) Different Types of FL: Depending on the characteristics of the participants' data and their distribution, FL can be categorized into three types, as shown in Figure 2.

Horizontally Federated Learning (HFL) is a form of FL, where participants possess distinct data samples but overlapping data features, as depicted in Figure 2a. It is also referred to as homogeneous federated learning. For instance, two hospitals at different locations may possess data on diverse patients, yet the underlying features are similar, resulting in overlapping data characteristics. *The presence of red vertical bars in Figure 2a indicate that participants have the same features.* The primary challenge in HFL is ensuring that the aggregated model accurately represents the clients' data while simultaneously upholding privacy and security concerns [45].

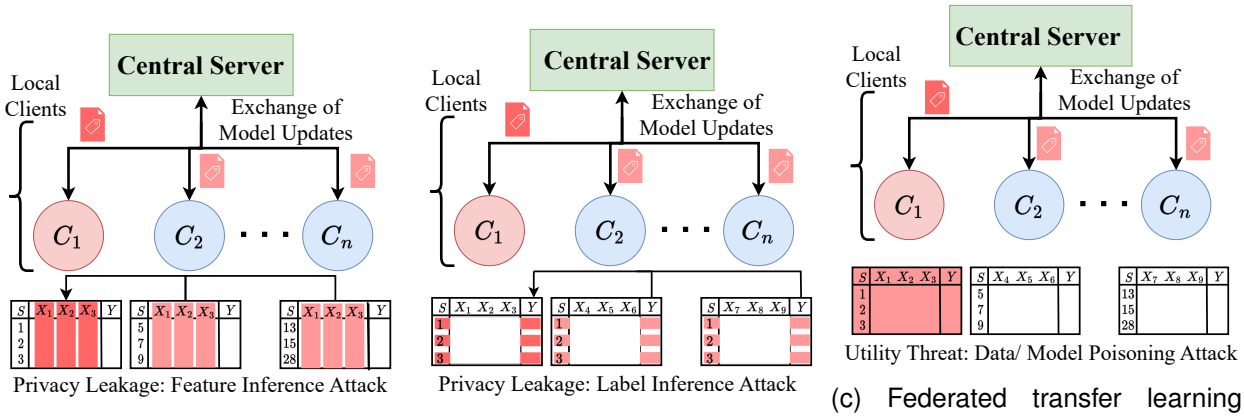
Vertically Federated Learning (VFL) encompasses scenarios where participants possess overlapping data samples but different features, as illustrated in Figure 2b. *The presence of red horizontal bars in Figure 2b signifies participants with the same overlapping samples.* For instance, in the finance industry, a bank and an e-commerce company may share common users, but the collected data from each entity may differ. The primary challenge in VFL lies in effectively aligning the data from different sources while upholding privacy and security considerations [46].

Federated Transfer Learning (FTL) integrates the concepts of FL and transfer learning [47]. It is characterized by limited overlap in data characteristics across different samples and minimal overlap of data samples among participants, as depicted in Figure 2c. *The presence of a red table in Figure 2c highlights distinct data and potential poisoning threats specific to FTL, which is discussed in subsequent sections.* Transfer learning is an ML technique that leverages knowledge acquired from one domain or task to enhance performance in another domain or task. FTL addresses the challenge of non-overlapping samples and features by training a model on a large public dataset and fine-tuning it with each participant's local data. FTL is particularly valuable in scenarios involving diverse data types, such as text, image, and speech data, which are not easily combined. The key challenge in FTL lies in ensuring the shared model's resilience to domain shifts between the public dataset and clients' local data, all while preserving privacy and security.

3) Threat Models in FL: Threat modelling [48] is a process of identifying potential threats and vulnerabilities by analyzing a system's components [49]. It involves identifying potential attackers, their motivations, capabilities, and tactics and assessing the impact of successful attacks on confidentiality, integrity, and availability. This process helps to prioritize critical threats and strengthen a system's security.

Threat modelling is essential in FL as it helps to identify potential vulnerabilities such as poisoning or model inversion attacks. Shejwalkar *et al.* [11] highlighted the crucial dimensions of the threat model in FL, as discussed below.

Attacker's Objective: Adversaries in FL aim to achieve three objectives, categorized by three attributes: security violation



(a) Horizontal FL (HFL) (b) Vertical FL (VFL) (c) Federated transfer learning (FTL)

Fig. 2: Overview of different types of federated learning with potential sources of vulnerabilities and possible attacks.

(integrity or availability), attack specificity (discriminate or indiscriminate), and error specificity (specific or generic) [11]. Misclassification of test inputs can compromise the system's performance. Understanding these objectives and attributes is crucial to protect FL systems.

Attacker's Knowledge: In FL, the adversary's knowledge can be classified as white-box, grey-box, or black-box, as shown in Figure 3. In the white-box setting [17], [50], the attacker has full access to model parameters and predictions, making it vulnerable to sophisticated model poisoning attacks. In the grey-box setting [51], the attacker has partial knowledge of the system, enabling the development of targeted poisoning attacks. In the black-box setting [52], the attacker lacks access to model parameters and must infer predictions, making data poisoning attacks a concern. In addition, attackers can have full or partial knowledge of data from benign clients [13], where full knowledge allows access to all benign and compromised clients in FL, while partial knowledge only provides access to data from compromised clients.

Attacker's Capabilities: Attackers in FL have various capabilities that can be broadly classified into two categories, namely, passive and active, as shown in Figure 3. Passive attacks [53] are those where the attacker simply observes the communication between the clients and the server without interfering with it. In FL, passive attacks may involve eavesdropping on the communication channels and can be difficult to detect as they do not disrupt the normal functioning of the system. On the other hand, active attacks involve actively manipulating the data and models of the clients and the server. Active attacks in FL can take different forms, such as data poisoning, model poisoning, and inference attacks [54].

III. Adversarial Attacks on FL

Figure 3 presents a comprehensive tree diagram of various FL attacks, categorized as active and passive, with sub-classifications of utility and privacy-centric attacks. The attacks originate from vulnerability sources like communication channels, models, and data. The attack space is further divided into attack type, mode, status, and setting, providing a comprehensive understanding of attack variations. For instance, attack types include targeted, untargeted, and backdoor, while modes involve periodic shuffling or without shuffling. Attack status can be online or offline, and the setting can be white, grey, or black-box. The leaf nodes represent actual attacks, offering specific examples of attack execution. This tree diagram serves as a valuable tool to comprehend the diversity of FL attacks and aids in developing effective defenses to enhance FL system security. In subsequent sections, further explanations and corresponding attacks are detailed in Table II.

1) Source of Attacks in FL: In FL, attacks can stem from clients, the central server, or the communication channel. As insiders participating in the training phase, clients can launch causative attacks affecting model accuracy. The central server, involved in the testing phase, is susceptible to evasion attacks attempting to bypass model defenses. Additionally, outsiders intercepting model updates between clients and the central server can target the communication channel with man-in-the-middle attacks. The fifth column of Table II shows sources of various state-of-the-art attacks in FL.

Potential sources of attacks in HFL, VFL, and FTL. Figure 2 provides an overview of FL types (HFL, VFL, and FTL) and their associated vulnerabilities. In HFL, clients with distinct samples but common features are vulnerable to feature inference attacks [78]. These attacks reconstruct the private data of benign clients by leveraging gradient-based generative adversarial networks (GANs) [78]. *The arrow in Figure 2a symbolizes the adversary's acquisition of feature representations from benign clients' training data and subsequent reconstruction of the entire private dataset.* On the contrary, in VFL, clients' data samples have overlapping characteristics with varying features (Figure 2b), making them vulnerable to label inference attacks [73]. The locally trained bottom model and gradients introduce potential privacy concerns, allowing a malicious participant to infer private labels [73] by analyzing the sign of received gradients. *The arrow in Figure 2b signifies the adversary's ability to infer the labels of the victims' data, ultimately leading to privacy breaches.* Further, in FTL, participants with distinct samples and features face utility threats. Adversaries exploit this to launch data poisoning attacks [11], [12] or manipulate model layers, resulting in

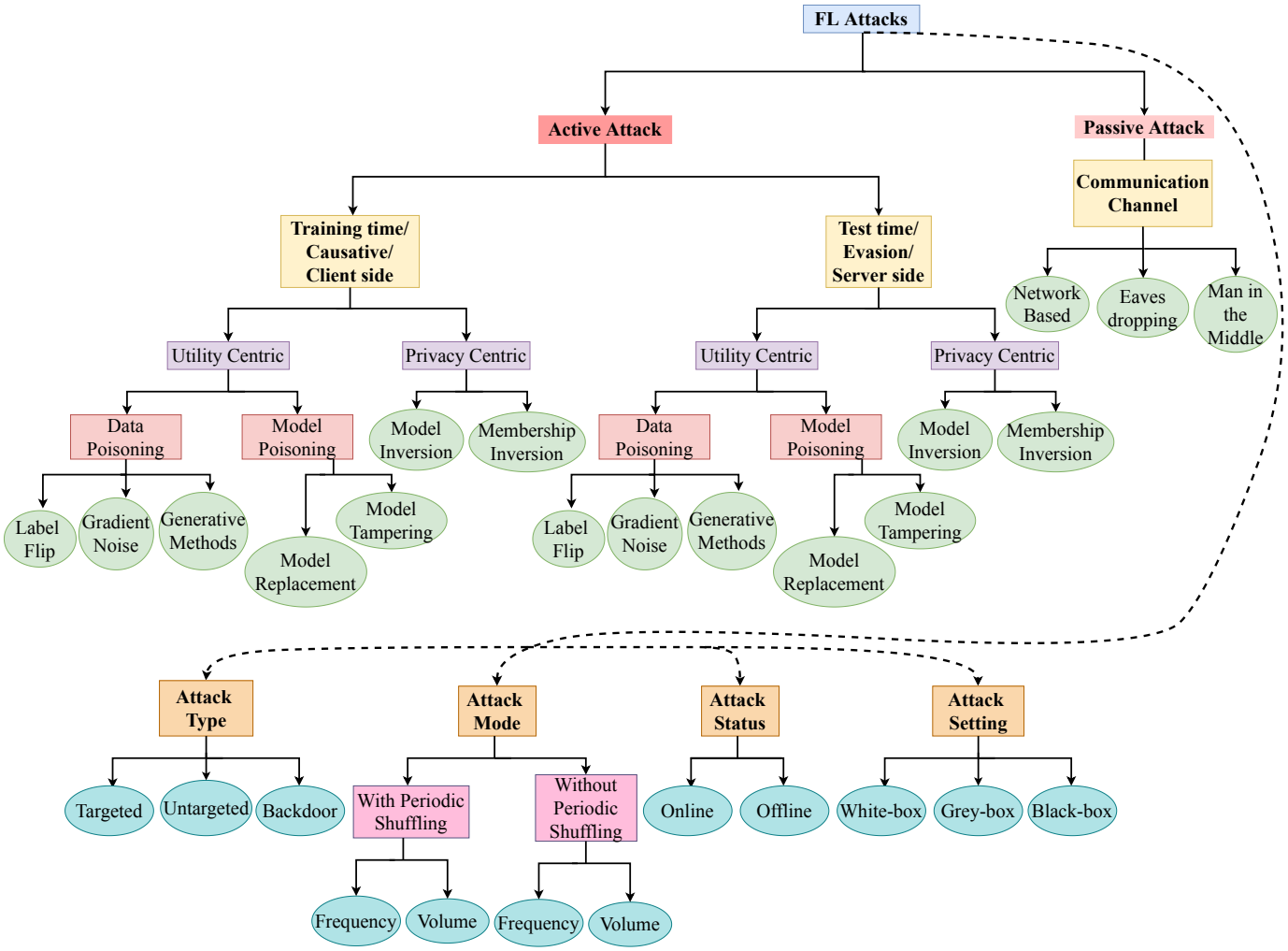


Fig. 3: A Comprehensive tree diagram of various types of attacks in FL, vulnerability sources, attack space including attack type, mode, status, and setting, with the actual attacks represented at the leaf nodes.

disruptive updates that compromise the utility of benign clients (Figure 2c). These attacks disrupt the intended purpose of the shared model and hinder the seamless transfer of knowledge from pre-trained models. *The presence of a red table in Figure 2c represents the injection of poisoned data that compromises the utility of benign clients.* The second column of Table II presents a list of various attacks targeting HFL, VFL, and FTL in FL. However, it is worth noting that attacks specifically targeting FTL systems are limited in number, indicating a gap in research and development in this area.

In summary, understanding FL types and associated vulnerabilities helps to develop targeted security measures.

2) Attack Impact (AIm) on FL: The impact of an attack in FL refers to the reduction in the accuracy of the global model caused by the attack. For utility-centric attacks, it is measured as the reduction in the test accuracy of the global model after the attack. To quantify the impact, A_θ represents the maximum accuracy achieved by the global model over all FL training rounds, and A^*_θ represents the maximum accuracy of the model under the given attack. The attack impact (I_θ) is then calculated as $|A_\theta - A^*_\theta|$ [11]. Privacy-centric attacks, aiming to perform membership inference or model inversion, have different impact metrics like mean squared error (MSE), peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and learned perceptual image patch similarity (LPIPS). *To assess the impact of various attacks, we assign three labels, i.e., low (L), medium (M), and high (H), based on their performance (more details are in the evaluation section V).* Evaluating attack impact is crucial in understanding attack severity and devising effective mitigation strategies. The sixth column of Table II displays the attack impact labels for different utility-centric and privacy-centric attacks in FL.

3) Attack Budget (AB) in FL: The attack budget in FL is critical in determining the resources an attacker employs to compromise the system. It encompasses the utility budget, impacting the global model's accuracy, and the privacy budget, compromising clients' data privacy. In general, the utility budget can consider factors like the number of malicious clients, the amount of poisoned data, the number of corrupted model layers in model poisoning attacks, the usage of surrogate models, and additional overhead models like GANs. On the other hand, the privacy budget parameter ϵ is considered while adding differential privacy [79], [80] as protection to FL systems. However, in this paper, we refer to

TABLE II: Comparative Analysis of State-of-the-Art Attacks with respect to their Attack Characteristics, Target Tasks, and Training Datasets. DP: data poisoning, MP: model poisoning, MsI: membership inference, MI: model inversion, CI*: category inference, AR*: attribute reconstruction, DDoS: Distributed Denial of Service, C: Client, S: Server, L: Low, M: Medium, H: High, On: online, Off: offline, W: white-box, G: grey-box, B: black-box, T: targeted, Semi T*: semi targeted, UT: untargeted, Bd: backdoor, U: utility-centric, and P: privacy-centric. '-' indicate not applicable or missing values.

S. No	Attack, Year	Type of FL	Point of Attack: DP/ MP MsI/ MI	Attack Source	Impact: L/M/H	Budget: L/M/H	Visibility: L/M/H	Attack Generalizibility	Status: On/Off	Setting: W/G/B	T/ UT/ Bd	Task	Dataset(s)	Intention: U/P	Communication Channel
1	DLF [11], 2022	HFL	DP	C	L	H	H	Yes	Off	B	UT	Image Classification	FEMNIST, CIFAR10, Purchase	U	No
2	SLF [11], 2022	HFL	DP	C	L	H	H	Yes	Off	B	UT	Image Classification	FEMNIST, CIFAR10, Purchase	U	No
3	PGA [11], 2022	HFL	MP	C	H	H	H	Yes	On	W	UT	Image Classification	FEMNIST, CIFAR10, Purchase	U	No
4	LIE [17], 2019	HFL	MP	C	M	L	L	No	On	W	Bd	Image Classification	MNIST, CIFAR10, CIFAR100	U	No
5	Yang *et al.* [55], 2023	VFL	MsI	C	H	L	L	Yes	On	B	-	Feature Prediction	Bank, Drive, Credit	P	No
6	DYN-OPT [50], 2021	HFL	MP	C	M	L	L	Yes	On	W	UT	Image Classification	MNIST, CIFAR10, Purchase, FEMNIST	U	No
7	ALGANs [56], 2022	HFL	MsI	C	H	H	-	No	On	W	-	Image Classification	MNIST, CIFAR10	P	No
8	PoisonGAN [12], 2020	HFL	DP	C	M	H	H	No	On	W	Bd	Image Classification	MNIST, CIFAR10 Fashion-MNIST,	U	No
9	LR-BA [57], 2023	VFL	DP	C	H	L	L	No	Off	W	Bd	Classification	NUS-WIDE, CIFAR10, CIFAR100, CINIC-10, BHI, Yahoo	U	No
10	AGIC [58], 2022	HFL	MI	S	M	L	-	Yes	On	W	-	Image Classification	CIFAR10, CIFAR100, ImageNet	P	No
11	ADA [59], 2022	HFL	MP	C	H	L	L	No	On	W, B	Semi T*	Image Classification	CIFAR10, MNIST, Fashion-MNIST	U	No
12	HCGLA [60], 2023	HFL	MI	S	H	L	-	No	On	W	-	Image Classification	CIFAR10, CIFAR100, MNIST, FMNIST, ImageNet, CelebA, LFW, PubFace, GoogleImages	P	No
13	Sybil [61], 2021	HFL	MP	C	M	H	H	Yes	On	W	UT	Image Classification	CIFAR10, MNIST	U	No
14	Category Privacy [62], 2023	HFL	CI*	S and/ or C	H	H	L	No	On	W	-	Image, Text Classification	CIFAR10, MNIST, AG_news, DBPedia	P	No
15	CAFE [14], 2021	VFL	MI	S	H	H	-	No	On	W	-	Image Classification	CIFAR10, MNIST, Linnaeus	P	No
16	Graph-Fraudster [63], 2022	VFL	DP	C	H	L	L	Yes	On	B	T	Graph Classification	Cora, Cora_ML, Citeseer, Pol.Blogs, Pubmed	P	No
17	ARA [64], 2022	HFL	AR*	S	H	M	-	Yes	On	W	-	Classification	Genome, Purchase, Location, Cancer	P	No
18	RL [65], 2022	HFL	MP	C	H	M	L	No	On	W,B	UT	Image Classification	MNIST, Fashion-MNIST, EMNIST, CIFAR10	U	No
19	Eavesdropping [66], 2021	HFL	Eavesdropping	-	H	H	L	-	On	W	-	Logistic Regression	LEAF Synthetic, Adult	P	Yes
20	Edge-case backdoor [67], 2020	HFL	DP and MP	C	H	L	L	Yes	On	W, B	Bd	Image, Sentiment Classification, Next word, Prediction	CIFAR10, EMNIST, ImageNet, Reddit Sentiment140	U	No
21	DBA [68], 2020	HFL	DP	C	H	H	H	No	On	W	Bd	Image Classification	CIFAR10, MNIST, LOAN, Tiny ImageNet	U	No
22	Stealthy Attack [69], 2019	HFL	MP	C	H	L	L	Yes	On	W	T	Image Classification	Fashion-MNIST, UCI Adult Censes	U	No
23	Model Replacement Attack [70], 2020	HFL	MP	C	H	L	L	No	On	W	Bd	Image Classification, Word Prediction	CIFAR10, Reddit	U	No
24	DDoS [71], 2000	HFL, VFL, and FTL	DDoS	-	H	L	-	-	On	-	-	Router Service	ns-2 network Simulator	-	Yes
25	Jamming [72], 2006	HFL, VFL, and FTL	Jamming	-	H	L	-	-	On	-	-	-	-	-	Yes
26	Label Inference [73], 2022	VFL	MsI	C	H	L	L	No	On	W	-	Image Classification, Text Representation, Predicting AD Click-through Rate,	CIFAR10, CIFAR100, CINIC-10, Yahoo Answers, Criteo, and Breast Histopathology Images	P	No
27	Internal Evasion [74], 2023	HFL	DP	C	M	H	L	Yes	On	G	UT	Image Classification, Text Representation	CIFAR10, Celeba, Fake News	U	No
28	Zhang *et al.* [75], 2023	HFL	MI	S and/ or C	H	L	-	No	On	G	-	Image Classification	MNIST, Fashion-MNIST, KMNIST	P	No
29	Feature Inference [76], 2021	VFL	MI	C	H	L	-	No	On	W	-	Classification	Bank marketing, Credit card, Drive Diagnosis, News popularity	P	No
30	Cheng *et al.* [77], 2022	FTL	DP	C	H	L	H	No	Off	B	T	Intrusion Detection	NSL-KDD, UNSW-NB15	U	No

the privacy budget from an attacker's perspective, where it is the number of resources used by the attacker to compromise the privacy of the individual client's data. We employed a comprehensive approach, considering *multiple parameters like total clients, clients selected per round for aggregation, number of malicious clients, local and global epochs, to assign labels, namely, low (L), medium (M), high (H) representing attack resource intensity and complexity.* Assigning labels based on these parameters enables us to gauge the attack's resource intensity and complexity. For instance, an attack involving a larger number of clients, a higher number of clients selected per round, and a greater count of malicious clients indicates a multi-client attack strategy with significant resource investment. Such attacks were assigned the "H" label. Additionally, the number of local and global epochs serve as a tie-breaker when two attacks had similar values for the previously mentioned parameters. The method with a higher number of local and global epochs was considered to require more iterations and computations to converge, indicating a more significant resource commitment. This consideration of local and global epochs as a tie-breaker further enhances the accuracy and granularity of our labelling approach. The

seventh column of Table II displays the attack budget label for various state-of-the-art utility and privacy-centric attacks in FL. Further comprehensive details and quantification used for labelling are in the evaluation section V of our study.

In summary, we acknowledge that every attack method operates within a highly dynamic range of multi-dimensional parameter values, and comparing them on a similar basis can be challenging. Nonetheless, we tried our best to estimate their attack budgets fairly through our labelling approach. It is important to note that the attack budget is inherently subjective and varies based on the attacker's intentions and available resources.

4) Attack Visibility (AV) in FL: Attack visibility is crucial in FL security, affecting attack detection and prevention. For utility-centric attacks, visibility lies in detecting malicious model updates, which can be challenging if attackers introduce stealthy model poisoning attacks or by poisoning the data in a way that is difficult to detect through accuracy and weight update analysis by the server. In this case, attack visibility lies in whether the model update from a malicious client is far away from benign updates. *We evaluate attack visibility for utility-centric attacks with labels, i.e., low (L), medium (M), high (H), based on the attacker's efforts to avoid detection.* Privacy-centric attacks focus on data reconstruction or membership inference without noticeable perturbations in the model, making attack visibility less relevant. Assessing attack visibility aids in developing countermeasures against potential attacks. For attacks with low visibility, enhanced detection measures may be necessary. The eighth column of Table II displays attack visibility labels for state-of-the-art utility-centric attacks, which does not apply to privacy-centric attacks in FL.

To summarize, we comprehensively evaluate FL attacks by considering three crucial dimensions: attack impact, budget, and visibility. These dimensions encompass the tradeoffs an attacker must weigh to launch a successful attack. However, the attacker's goal is to create maximum impact with minimal resources and low visibility, these dimensions remain subjective and depend on the attacker's perspective and multi-dimensional factors. We aim to provide a fair and appropriate understanding of the state-of-the-art attacks' tradeoffs despite subjectivity. Our approach recognizes FL systems' vulnerability to attacks, including those with low budget and visibility but significant impact. By considering these dimensions, we gain insight into the attacks' nuances and develop stronger defenses. The evaluation section V quantifies and justifies these labels based on relevant literature analysis.

5) Attack Generalizability in FL: Attack generalizability in FL refers to an attacker's ability to compromise the privacy or utility of an FL system, regardless of its specific configuration. FL systems can vary in device types, client selection, aggregation methods, etc., making generalizability essential for attackers to launch large-scale attacks. In utility-centric attacks, generalizability means an attack can degrade the performance of FL models with similar characteristics. For example, a data poisoning or model stealing attack that can successfully compromise one FL model may also be able to compromise another FL model with similar architectures. Similarly, in privacy-centric attacks, it refers to compromising participants' data across different FL configurations or models. Techniques like data validation, model validation, and differential privacy are crucial in designing robust FL systems to resist attacks with high generalizability. By studying generalizability in FL, we protect the privacy and utility of FL models against large-scale threats. The ninth column of Table II displays the attack generalizability for various state-of-the-art utility and privacy-centric attacks in FL.

6) Attack Status: Attack status in FL refers to whether an attack is continuous during every communication round (online) or occurs only at the beginning of the FL process (offline). Online attacks involve the adversary repeatedly poisoning compromised clients, leading to a persistent impact on the attack budget and visibility. For example, model poisoning attacks in FL continuously poison the updated global model in each round, resulting in lasting effects. On the other hand, offline attacks poison only at the start of the FL process, such as static label flipping data poisoning attacks, with low budget and visibility. Understanding the attack status is crucial in estimating the attacker's intentions for high impact, low budget, and low visibility online attacks. By comprehending the attacker's behaviour, appropriate defense mechanisms can be implemented to safeguard the FL system from online and offline attacks. The tenth column of Table II displays the state-of-the-art FL attack status.

7) Attack Setting: The attacker's knowledge influences the attack setting in FL and can be divided into three categories: white-box, grey-box, and black-box, as discussed in Section II-3. The black-box offline data poisoning attack setting and white-box online model poisoning setting are closest to production deployment FL according to [11]. Understanding the attacker's knowledge setting is crucial in estimating the attacker's budget, impact, and visibility. The eleventh column of Table II presents the knowledge setting for various state-of-the-art FL attacks. Our observation indicates a notable research gap in grey-box FL attacks, with limited studies focusing on this area.

8) Attack Type: In FL, utility-centric attacks can be categorized into targeted, untargeted, and backdoor attacks based on the attacker's objectives. In targeted attacks, the goal is to manipulate the model's prediction for a specific target class. In contrast, untargeted attacks aim to degrade the model's overall performance by causing incorrect predictions for any class. Backdoor attacks involve adding a trigger or pattern to the training data, leading the model to make incorrect predictions when the trigger is present in the input data. The twelfth column of Table II displays the attack types for various state-of-the-art utility-centric FL attacks. Understanding these attack types is vital for designing robust FL systems and developing effective defense mechanisms. By studying their nature and impact, we can identify weaknesses, evaluate tradeoffs between security and performance, and implement preventive measures to enhance the security of FL systems.

9) Attack Mode: The area of FL attacks can be visualized from the perspective of attack modes, specifically with and without periodic shuffling, as shown in Figure 3. Periodic shuffling refers to the attacker dynamically switching between

malicious client IDs, making it challenging for the detection mechanism to identify the malicious clients. This parameter is characterized by factors such as frequency (number of global epochs in which the adversary maintains the same IDs) and volume (number of malicious clients at any given time). Considering this parameter is crucial because periodic shuffling allows the attacker to confuse the server's detection mechanism by switching IDs, significantly impacting the overall performance of FL. By understanding this aspect, effective defense mechanisms can be developed to counter such attacks and improve the security of FL systems.

10) Common Tasks in FL under Attack: The tasks targeted by various state-of-the-art FL attacks are shown in the thirteenth column of Table II. While most attack papers focus on image classification tasks, a few also consider other tasks such as regression, text, sentiment classification, and anomaly detection. However, there remains a research gap in attacking higher computer vision tasks like detection, tracking, and segmentation within the domain of FL. Addressing this gap would lead to a more comprehensive understanding of potential vulnerabilities and risks associated with FL across various domains. By identifying and mitigating such vulnerabilities, more robust and secure FL systems can be designed to protect sensitive data and provide reliable and accurate outcomes.

IV. Defense Strategies

The defense strategies in FL are of different categories. Table III presents a comprehensive overview of existing defenses for utility and privacy-centric attacks systematically grouped based on categories, sources, attacks they defend, and the used metrics.

1) Defenses against Utility-centric FL Attacks:

- **Adversarial Training:** This defense strategy is primarily implemented at the client side and aims to defend against data poisoning attacks. The local model at the clients is trained on adversarial examples, which are maliciously crafted input data designed to fool the model and cause incorrect predictions. By training the model on these examples, it can learn to be more robust to such attacks. Examples of this strategy include FAT [81], FedDynAT [82], and GALP [83].
- **Model Pruning:** Model pruning is a defense strategy applied against model poisoning attacks in FL and can be performed on both the server and client side, depending on the specific approach. On the server side, pruning algorithms can be applied during the aggregation process to remove unnecessary parameters or connections from the global model. Similarly, clients can perform local model pruning before sending updates to the server, reducing the size and focusing on relevant features for their data. Overall, it reduces model complexity and computational requirements, enhancing communication, robustness, generalizability, and inference efficiency. Examples include PruneFL [84] and Network Pruning [85].
- **Byzantine Robust Aggregation Techniques:** These are typically implemented on the server side and used to protect against model and data poisoning attacks where some clients send malicious updates to the server. These techniques are used as aggregation algorithms at the server instead of FedAvg [5] to filter out malicious updates and ensure that only valid updates are used in the model. Some of them include Krum [86], Multi-krum [86], Trimmed Mean [87], Median [87], Bulyan [88], RLR [89], ZeKoC [90], ShieldFL [91], Adaptive Model Averaging [92], and FLTrust [93].
- **Regularization:** Regularization is a defense strategy applicable to both server and client sides in FL, protecting against data poisoning and model poisoning attacks. On the server side, regularization techniques are applied during model update aggregation to prevent overfitting and enhance model generalizability. Common methods include L1 or L2 regularization, which introduces penalty terms in the loss function during model training. It controls the shared model's complexity and promotes learning robust patterns from local data. On the client side, participants employ regularization techniques like dropout, batch normalization, or weight decay during their model training. This combats overfitting on local data, contributing to overall FL robustness. Examples include LSR [94], ConTre [95], and Su *et al.* [96].
- **Detect and Remove:** This method is typically implemented on the server side. It involves detecting and removing malicious updates from the model. This is done by monitoring the updates and flagging any that are suspicious or do not conform to the expected pattern. Examples include LoMar [97], Federated Reverse [98], DDaBA [99], SecFedNIDS [100], PA-SM [101], Density-based anomaly detection [102], BaFFLe [103], KPCA [104], and Mahalnobix distance-based detection [105].
- **Robust Client Selection Techniques:** These are used to select clients that are most trustworthy and reliable and are implemented on the server side. This helps to ensure that only valid updates are used in the aggregation and reduces the risk of using malicious updates. Examples include Ensemble FL [106], DivFL [107], and FedClean [108].
- **Data and Update Analysis:** This defense strategy can be implemented at both the server and client sides in FL. On the server side, it involves scrutinizing aggregated client updates to detect suspicious patterns or inconsistencies. Statistical and ML techniques, such as outlier detection and anomaly detection, are applied to identify potential malicious activity or data manipulation. On the client side, data and update analysis include preprocessing, data quality checks, validation, and cleaning to ensure data integrity before training. Clients also validate their local updates before sharing them with the server. Examples include SIREN [109], DeepSight [110], Biscotti [111], and FL-Defender [112].

2) Defenses against Privacy-centric FL attacks:

- **Homomorphic Encryption:** This powerful defense in FL ensures privacy protection by allowing computations to be performed on encrypted data. Homomorphic encryption enables data to remain encrypted while still enabling

TABLE III: Comparison and Overview of Existing Defenses for Utility and Privacy-Centric Attacks in FL. U: utility-centric, P: privacy-centric, DP: data poisoning, MP: model poisoning, S: server, and C: client.

Attack Intention	Defense Category (Source of Defense)	Defense Strategies	Can Defend (Point of Attack: Attacks)	Metrics (Used)
U	Adversarial Training **(C)**	FAT [81], FedDynAT [82], GALP [83]	**DP:** DLF [11], SLF [11], PoisonGAN [12], AT2FL [18], DBA [68], Edge-case backdoor [67]	Accuracy, True Positive Rate, Precision, Recall, F1 Score, Attack Success Rate
	Model Pruning **(S and/or C)**	PruneFL [84], Network Pruning [85]	**MP:** PGA [11], LIE [17], STAT-OPT [13], DYN-OPT [50], RL [113] ADA [59], Sybil [61], Stealthy Poisoning [69], Model Replacement [70], Random Reports [114]	
	Byzantine Robust Aggregation Techniques **(S)**	Krum [86], Multi-krum [86], Trimmed Mean [87], Median [87], Bulyan [88], RLR [89], ZeKoC [90], ShieldFL [91], Adaptive Model Averaging [92], AutoGM [115], FLTrust [93]	**DP:** DLF [11], SLF [11], PoisonGAN [12], AT2FL [18], DBA [68] Edge-case backdoor [67], **MP:** PGA [11], LIE [17], STAT-OPT [13], DYN-OPT [50], RL [113] ADA [59], Sybil [61], Stealthy Poisoning [69], Model Replacement [70], Random Reports [114]	
	Regularization **(S and/or C)**	LSR [94], ConTre [95], Su *et al.* [96]		
	Detect and Remove **(S)**	FLDetector [116], EIFFeL [117], LoMar [97], Federated Reverse [98], DDaBA [99], SecFedNIDS [100], Wan *et al.* [102], PA-SM [101], BaFFLe [103], KPCA [104], Lee *et al.* [105], Li *et al.* [113]		
	Robust Client Selection **(S)** Techniques	Ensemble FL [106], FedClean [108], DivFL [107], Auction [118], FEDGS [119]		
	Data and Update Analysis **(S and/or C)**	FL-WBC [120], SIREN [109], DeepSight [110], Biscotti [111], FL-Defender [112], SparseFed [121]		
P	Homomorphic Encryption **(C)**	DCAE [46], PEFL [122], Batchcrypt [123], Dropout-tolerance [124]	ALGANs [56], GAN based membership inference [125] AGIC [58], HCGLA [60], CAFE [14], GRNN [15], Evasdropping [66], User-level Privacy Leakage [16], Feature Inference [76]	MSE, PSNR, SSIM, Accuracy, Privacy Loss, LPIPS
	Knowledge Distillation **(C)**	FEDGEN [126], FedKD [127], FedFTG [128], PATE [129]		
	Secure Multi-party Computation **(C)**	AMPC [130], SMPAI [131], Byrd *et al.* [132]		
	Trusted Execution Environments **(C)**	Flatee [133], Chen *et al.* [134], PPFL [135]		
	Split Learning **(C)**	SplitLearn [136], FSL [137], Splitfed [138]		
	Perturbing Gradients **(C)**	OtA-FL [139], Yang *et al.* [140], Local Random Perturbation [141], Soteria [142], DgstNN [143], Haung *et al.* [144], DtC [145]		
	Differential Privacy **(C)**	NbAFL [146], Hu *et al.* [147], Triastcyn *et al.* [79], PixelDP [148], 2DP-FL [149]		

useful computations, making it challenging for attackers to access sensitive information. Examples include DCAE [46] and PEFL [122].

- **Knowledge Distillation:** It involves transferring knowledge from a larger, complex model to a smaller, simpler one. By doing so, the smaller model can make accurate predictions without processing as much data, reducing the need to share large amounts of data between the client and server. Examples include FEDGEN [126], FedKD [127], FedFTG [128], PATE [129].
- **Secure Multi-party Computation:** This technique protects privacy by enabling multiple parties to jointly compute a function on their inputs without revealing individual data. It ensures a secure process without disclosing any sensitive information. Examples include AMPC [130], SMPAI [131], and Byrd *et al.* [132].
- **Trusted Execution Environments (TEE):** By providing a secure environment for running sensitive applications, TEEs protect against tampering or reverse engineering attacks. This protects against privacy attacks that attempt to steal or analyze data. Examples include Flatee [133], Chen *et al.* [134], PPFL [135].
- **Split Learning:** It divides the model between the client and the server, allowing sensitive data to be kept on the client side while enabling the model to be updated in a distributed manner. This approach protects against attacks that attempt to steal or analyze data. Examples include SplitLearn [136], FSL [137], and Splitfed [138].
- **Perturbing Gradients:** This technique protects susceptible data from unauthorized access by adding random noise to the gradients used during model training. By doing so, it hinders attackers from using gradient-based

attacks to reverse engineer the model or steal sensitive data. Examples include OtA-FL [139], Yang *et al.* [140], Local Random Perturbation [141], Soteria [142], DgstNN [143], Haung *et al.* [144], and DtC [145].

- **Differential Privacy:** This technique is employed in highly sensitive FL scenarios to protect data from unauthorized access. By introducing noise to the model update before sending it to the server, differential privacy safeguards the privacy of the data used for model training, preventing attackers from learning sensitive information about individual data points. It serves as a defense against privacy attacks that aim to steal or analyze data. Examples include NbAFL [146], Hu *et al.* [147], Triastcyn *et al.* [79], PixelDP [148], and 2DP-FL [149].

V. Evaluation

In this section, we assess the latest attacks based on the tradeoff dimensions of attack impact, visibility, and budget. Additionally, we present the commonly used datasets, FL, and attack parameters, as well as the metrics and their respective values in the subsequent sections.

1) Datasets: The datasets commonly employed for utility-centric attacks in FL include CIFAR10, MNIST, Fashion-MNIST, Purchase, FMNIST, EMNIST, and Reddit. Meanwhile, MNIST, CIFAR10, CIFAR100, ImageNet, and LFW are the datasets commonly used for privacy-centric attacks in FL, as indicated in Table II.

2) Metrics: As outlined in Section V, the difference between the global test accuracy (GTA) before and after the attack is utilized to evaluate the impact of utility-centric attacks in FL. Furthermore, other metrics such as error rate, RFA distance, FoolsGold weight, L_2 distance, and model weight values are also employed. In contrast, the evaluation of privacy-centric attacks in FL is carried out using metrics such as accuracy, precision, recall rate, F1 score, MSE, PSNR, SSIM, and LPIPS. Table IV and Table V display all the aforementioned metrics.

3) Tradeoff Evaluation of Utility-centric Attacks in FL: Figure 4 shows a comparative analysis of utility-centric attacks in FL, considering the tradeoff between attack budget, visibility, and impact. Ideally, attackers aim for low budget, low visibility, and high impact. However, there is a tradeoff among these dimensions, where a higher budget can lead to greater impact even with lower visibility. On the other hand, high visibility can result in a higher detection rate and may lead to lower impact. Balancing these dimensions is crucial for attackers to achieve their objectives effectively.

Attacker Perspective Label Assignment and Quantification: Assigning and quantifying labels for the attack budget, visibility, and impact of utility-centric attacks in Table II (Section III) is a complex task due to the diverse and dynamic parameter settings each attack operates under, as shown in Table IV. Our approach aims to provide a generalized assessment of attacks from the attacker's perspective, offering a comprehensive view of their relative standings concerning impact, budget, and visibility.

The attack impact labels are systematically assigned by designating 'L' (low) if the impact is below 30%, 'M' (medium) for impacts between 30% to 50%, and 'H' (high) for impacts exceeding 50%. However, exceptions are considered, and additional metrics are analyzed for accurate label assignments. For example, the LIE attack, with an impact above 50%, is labeled 'M' due to its operation under lower client numbers and a relatively simple CNN architecture. This careful consideration of specific attack characteristics allows for appropriate impact label assignments, providing a nuanced assessment of each attack's potential influence.

The attack budget labels are based on an analysis of various parameters, including total clients, clients selected per round, malicious clients, local and global epochs, as discussed in Section III-3. For example, the DLF attack [11] involves 1000 total clients, 25 clients per round, 100 malicious clients, and 2 local and 1000 global epochs, resulting in a 'H' (high) attack budget label. The RL attack [65] with 100 total clients, 10 clients per round, 20 malicious clients, and 20 local and 1000 global epochs, receives a 'M' (medium) attack budget label. The Edge-case backdoor attack [67], requiring 200 total clients, 10 clients per round, 10 malicious clients, and 2 local and 500 global epochs, is assigned a 'L' (low) attack budget label in Table II. Through this approach, we can effectively gauge the resource intensity and complexity of each attack, enabling appropriate labels to indicate the level of budget required for its execution.

The label assignment for attack visibility involves two categories. Attacks designed with defense mechanisms receive an 'L' (low) visibility label, while those without such considerations are labeled 'H' (high) visibility as they lack specific detection avoidance measures. This comprehensive label assignment and quantification process allow us to assess utility-centric attacks from the attacker's perspective, gaining valuable insights into their impact, budget requirements, and visibility in the context of our research.

The quantification of attack budget, visibility, and impact in Figure 4 is based on a [0, 1] scale, representing low, medium, and high levels. Attacks labeled as low are assigned a value of 0.1 for each parameter (budget, visibility, and impact), while medium and high labels are assigned values of 0.5 and 0.9, respectively. Minor adjustments of $\pm$ 0.05 were made for similar attacks based on their evaluation in Table IV. Notably, the Edge-case attack [67] and ADA [59] achieved close to ideal attack states in Figure 4, posing potential risks and necessity in evaluating against FL defenses.

Attack Impact: Table IV presents the impact of utility-centric attacks in FL, ranging from 3% to 76.14% for data poisoning attacks and 3.4% to 89.97% for model poisoning attacks across various datasets and threat models. It offers a structured classification based on datasets, data vs model poisoning, threat model parameters, and best vs worst case attack impact w.r.t different numbers of malicious clients. The table provides readers with a comprehensive overview of attack impact values achieved by state-of-the-art attacks under different settings. *We emphasize that our focus of comparison lies in tradeoff analysis rather than declaring the best attack based on only attack impact, as attack budget, visibility, and other factors also play an important role in evaluating the efficiency of the attack.*

TABLE IV: Comparison and Evaluation of Commonly Used Datasets in FL: Analysis of Utility-centric Attacks and their Impact on Global Test Accuracy (GTA) and Other Metrics. DP: data poisoning, MP: model poisoning, MC: number of malicious clients, α: Dirichlet distribution parameter, LE: local epochs, and GE: global epochs. '-' indicate not applicable or missing values.

Dataset	Point of Attack	Attack	Total Clients	Clients Selected Per Round	MC	α	Aggregation Technique	LE	GE	Model	GTA Before Attack $\mathbf{A}_\theta$(%)	Best and Worst Case GTA After Attack, MC: $\mathbf{A}^*_\theta$ (%)	Best and Worst Case Attack Impact, MC: $\mathbf{I}_\theta = \|A_\theta - \mathbf{A}^*_\theta\|$ (%)	Other Metrics
CIFAR10 [151]	DP	DLF [11]	1000	25	1, 10, 100	1	FedAvg, Norm-bound, Multi-Krum, Trimmed Mean	2	1000	VGG-9	86.6	**Best** 100: 81.6 **Worst** 1: 83.6	**Best** 100: 5 **Worst** 1: 3	-
		SLF [11]	1000	25	1, 10, 100	1	FedAvg, Norm-bound, Multi-Krum, Trimmed Mean	2	1000	VGG-9	86.6	**Best** 100: 82.6 **Worst** 1: 84.6	**Best** 100: 4 **Worst** 1: 2	-
		PoisonGAN [12]	100	10	5, 10, 15, 20	-	FedAvg	20	300	ResNet-18	95.43	**Best** 20: 32.25 **Worst** 5: 65	**Best** 20: 63.18 **Worst** 5: 30.43	-
		Edge-case backdoor [67]	200	10	1, 2, 4, 10	0.5	FedAvg, Krum, NDC Multi-Krum, RFA	2	500	VGG-9	77.68	**Best** 10: 50 **Worst** 2: 60	**Best** 10: 27.68 **Worst** 2: 17.68	-
		DBA [68]	100	10	4	0.5	FedAvg, Multi-Krum, Bulyan	6	400	ResNet-18	84	**Best = Worst** 4: 81	**Best = Worst** 4: 3	RFA Distance, FoolsGold Weight
	MP	PGA [11]	1000	25	1, 10, 100	1	FedAvg, Norm-bound, Multi-Krum, Trimmed Mean	2	1000	VGG-9	86.6	**Best** 100: 21.6 **Worst** 10: 81.6	**Best** 100: 65 **Worst** 10: 5	-
		LIE [17]	51	51	12	-	FedAvg, Krum, Bulyan, Trimmed Mean	5	50	CNN	59.6	**Best = Worst** 12: 7.3	**Best = Worst** 12: 52.3	-
		DYN-OPT [50]	50	50	10	-	Krum, Bulyan, Multi-Krum, Trimmed Mean, AFA, FangTMean	1	1200	VGG-11, Alexnet	67.6	**Best = Worst** 10: 30.3	**Best = Worst** 10: 35.3	-
		ADA [59]	100	10	1	-	FedAvg	1	50	4 layer CNN	71.1	**Best = Worst** 1: 58.1	**Best = Worst** 1: 13	-
		Sybil [61]	100	10	20	-	FedAvg, Krum, Trimmed Mean	5	50	MLP	41	**Best = Worst** 20: 9	**Best = Worst** 20: 32	Error Rate
		RL [65]	100	10	20	0.5	FedAvg, Krum, Clipping Median	20	1000	ResNet-18	60	**Best = Worst** 20: 10	**Best = Worst** 20: 50	-
		Model Replacement Attack [70]	100	10	1, 2, 5, 10, 20, 50, 100	0.9	FedAvg	2	100	ResNet-18	85	**Best** 100: 40 **Worst** 1: 80	**Best** 100: 45 **Worst** 1: 5	-
MNIST [152]	DP	PoisonGAN [12]	33	10	3, 4, 5, 6	-	FedAvg	20	300	6 layer CNN	96.41	**Best** 6: 52.87 **Worst** 3: 80	**Best** 6: 43.54 **Worst** 3: 16.41	-
		DBA [68]	100	10	4	0.5	FedAvg, Multi-Krum, Bulyan	10	400	2 layer CNN	91.57	**Best = Worst** 4: 38	**Best = Worst** 4: 53.57	RFA Distance, FoolsGold Weight
	MP	LIE [17]	51	51	12	-	FedAvg, Krum, Bulyan, Trimmed Mean	5	50	CNN	96.1	**Best = Worst** 12: 36.9	**Best = Worst** 12: 59.1	-
		STAT-OPT [13]	100	100	20	0.5	FedAvg, Krum, Median, Trimmed Mean	10	2000	DNN	90	**Best = Worst** 20: 30	**Best = Worst** 20: 60	Error Rate
		DYN-OPT [50]	100	100	20	-	Krum, Bulyan, Multi-Krum, Trimmed Mean, AFA, FangTMean	1	500	Fully Connected (FC) Network	96.5	**Best = Worst** 20: 94	**Best = Worst** 20: 2.5	-
		ADA [59]	100	10	1	-	FedAvg	1	50	4 layer CNN	98.9	**Best = Worst** 1: 83.2	**Best = Worst** 1: 15.7	-
		Sybil [61]	100	10	20	-	FedAvg, Krum, Trimmed Mean	5	50	CNN	99.97	**Best = Worst** 20: 10	**Best = Worst** 20: 89.97	Error Rate
		RL [65]	100	10	20	0.5	FedAvg, Krum, Clipping Median	20	1000	3 layer CNN	91	**Best = Worst** 20: 10	**Best = Worst** 20: 81	-
Fashion-MNIST [153]	DP	PoisonGAN [12]	33	10	3, 4, 5, 6	-	FedAvg	20	300	6 layer CNN	96	**Best** 6: 75 **Worst** 3: 90	**Best** 6: 21 **Worst** 3: 6	-
		STAT-OPT [13]	100	100	20	0.5	FedAvg, Krum, Median, Trimmed Mean	10	2000	DNN	84	**Best = Worst** 20: 9	**Best = Worst** 20: 75	Error Rate
	MP	ADA [59]	100	10	1	-	FedAvg	1	50	4 layer CNN	81.1	**Best = Worst** 1: 69.5	**Best = Worst** 1: 11.6	-
		RL [65]	100	10	20	0.5	FedAvg, Krum, Clipping Median	20	1000	3 layer CNN	80	**Best = Worst** 20: 10	**Best = Worst** 20: 70	-
		Stealthy Attack [69]	10, 100	10	1	-	FedAvg, Krum, Median	5	40, 50	3 layer CNN	91	**Best = Worst** 1: 15	**Best = Worst** 1: 76	L_2 Distance Weight Values
Purchase [154]	DP	DLF [11]	5000	25	1, 10, 100	1	FedAvg, Norm-bound, Multi-Krum, Trimmed Mean	5	500	Fully Connected, (FC) Network	81.2	**Best** 100: 56.2 **Worst** 1: 78.2	**Best** 100: 25 **Worst** 1: 3	-
		SLF [11]	5000	25	1, 10, 100	1	FedAvg, Norm-bound, Multi-Krum, Trimmed Mean	5	500	Fully Connected, (FC) Network	81.2	**Best** 100: 61.2 **Worst** 1: 78.2	**Best** 100: 20 **Worst** 1: 3	-
	MP	PGA [11]	5000	25	1, 10, 100	1	FedAvg, Norm-bound, Multi-Krum, Trimmed Mean	5	500	Fully Connected, (FC) Network	81.2	**Best** 100: 48.2 **Worst** 1: 80.2	**Best** 100: 33 **Worst** 1: 1	-
		DYN-OPT [50]	80	80	16	-	Krum, Bulyan, Multi-Krum, Trimmed Mean, AFA, FangTMean	1	1000	Fully Connected (FC) Network	91.7	**Best = Worst** 16: 88.3	**Best = Worst** 16: 3.4	-
FMNIST [155]	DP	DLF [11]	34000	50	1, 10, 100	1	FedAvg, Norm-bound, Multi-Krum, Trimmed Mean	5	500	LeNet	82.4	**Best** 100: 42.4 **Worst** 1: 79.4	**Best** 100: 40 **Worst** 1: 3	-
		SLF [11]	34000	50	1, 10, 100	1	FedAvg, Norm-bound, Multi-Krum, Trimmed Mean	5	500	LeNet	82.4	**Best** 100: 52.4 **Worst** 1: 77.4	**Best** 100: 30 **Worst** 1: 5	-
	MP	PGA [11]	34000	50	1, 10, 100	1	FedAvg, Norm-bound, Multi-Krum, Trimmed Mean	5	500	LeNet	82.4	**Best** 100: 57.4 **Worst** 1: 81.4	**Best** 100: 25 **Worst** 1: 1	-
		DYN-OPT [50]	3400	60	680	-	Krum, Bulyan, Multi-Krum, Trimmed Mean, AFA, FangTMean	1	1500	CNN	84.6	**Best = Worst** 680: 7.6	**Best = Worst** 680: 77	-
EMNIST [156]	DP	Edge-case backdoor [67]	3383	30	17, 34, 68, 170	0.5	FedAvg, Krum, NDC Multi-Krum, RFA	5	500	LeNet	88	**Best** 170: 20 **Worst** 34: 80	**Best** 170: 68 **Worst** 34: 8	-
	MP	RL [65]	100	10	20	0.5	FedAvg, Krum, Clipping Median	20	1000	3 layer CNN	80	**Best = Worst** 20: 10	**Best = Worst** 20: 70	-
Reddit [5], [70]	DP	Edge-case backdoor [67]	80000	100	400, 800, 1600, 4000	0.5	FedAvg, Krum, NDC Multi-Krum, RFA	2	600	LSTM	18.86	**Best** 4000: 1 **Worst** 400: 5	**Best** 4000: 17.86 **Worst** 400: 13.86	-
	MP	Model Replacement Attack [70]	83293	100	8, 41, 83, 166, 415, 832, 1664, 4160, 8320	0.9	FedAvg	2	100	LSTM	19	**Best** 8320: 1 **Worst** 8: 14.25	**Best** 8320: 18 **Worst** 8: 4.75	-

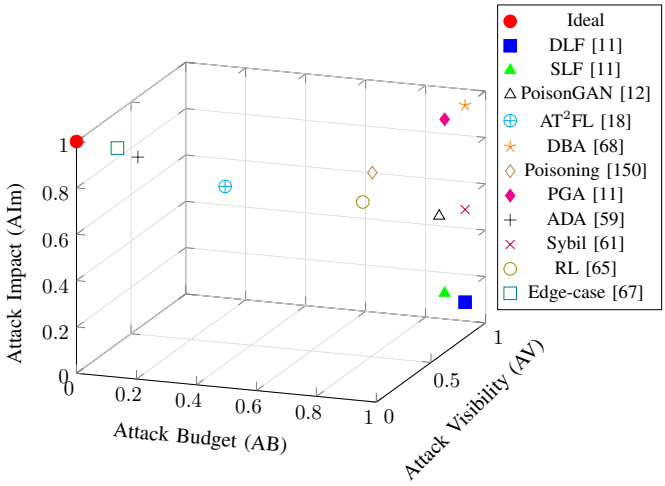


Fig. 4: Comparative 3D scatter plot analysis of utility-centric attacks with respect to attack budget (AB), attack visibility (AV), and attack impact (AIm).

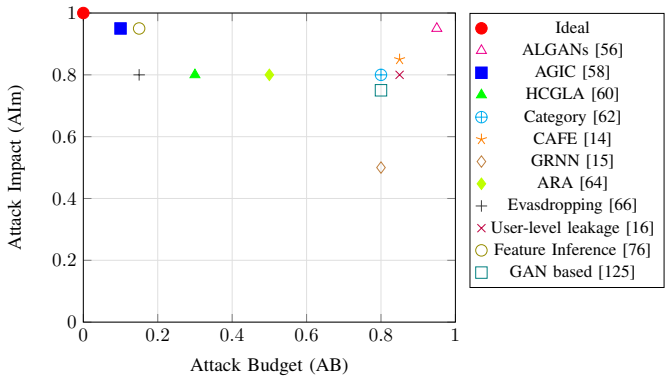


Fig. 5: Comparative analysis of privacy-centric attacks with respect to attack budget (AB) and attack impact (AIm).

4) Tradeoff Evaluation of Privacy-centric Attacks in FL: Figure 5 presents a two-dimensional comparative analysis of privacy-centric attacks in FL, focusing on the attack budget and impact dimensions since the attack visibility is not particularly relevant. The assignment of labels and quantification within the range of [0, 1] for privacy-centric attacks follows a similar approach to utility-centric attacks, as previously discussed. Also, minor adjustments of $\pm$ 0.05 are made for attacks with similar labels based on evaluation with respect to Table V. After an extensive literature review, we identified AGIC [58] and feature inference attack [76] approach as the ideal attack state, demonstrating significant effectiveness and efficiency in compromising privacy. AGIC utilizes approximate gradient inversion to efficiently reconstruct images from model or gradient updates across multiple epochs, while feature inference attack proposes a method based on logistic regression and decision tree models and extends to complex models like neural networks and random forest models. These observations highlight the importance of considering various attack characteristics to evaluate the robustness and severity.

Attack Impact: Table V presents the impact of privacy-centric attacks in FL using commonly used metrics such as Accuracy, PSNR, SSIM, MSE, and F1 score, among others. The impact values vary across datasets and attacks, depending on factors like threat models, attackers' capabilities, and objectives. The table provides a structured classification of attack impact data based on different datasets, points of attack, privacy attack metrics, threat model parameters, and impact. *Our intention with this table is to offer readers a comprehensive overview of different datasets and parameter values used in the literature by various state-of-the-art privacy-centric attacks and their corresponding attack impact values rather than declaring the best attack.*

5) Other FL Parameters: Table IV and Table V present different FL parameter values for utility-centric and privacy-centric attacks, respectively.

Clients' number: The number of clients used in FL attacks varies widely, ranging from 10 to 83293, with 10 to 100 client updates selected by the server for aggregation per round in utility-centric attacks. The number of malicious clients varies from 1 to 8320, depending on their capabilities and objectives. For privacy-centric attacks, the total number of clients typically ranges from 4 to 100, with all clients being used for every round. As the server is the primary source of the attack in privacy-centric scenarios, there are no malicious clients. However, we observed that some papers lack details regarding the total number of clients and the number of clients selected per round. The number of clients directly influences the attack budget, visibility, and, consequently, the impact.

Effect of Non-IID on FL Attacks (Dirichlet Distribution [159]): The Dirichlet distribution is commonly used in FL to model data distribution across clients. By adjusting the parameter α, the density of independently and identically distributed (IID) data splits among clients can be controlled, influencing the non-IID nature of the data distribution. Higher values of α result in a more uniform data distribution among clients, reducing variation in their data distributions. Conversely, lower values of α lead to a more concentrated or skewed data distribution, introducing heterogeneity and non-IIDness among clients. Properly adjusting α enables FL systems to account for the inherent heterogeneity in client data, which is crucial in real-world FL scenarios.

The Dirichlet distribution plays a crucial role in assessing the impact of FL utility attacks. When combined with attack strategies, highly non-IID client datasets can reduce global model accuracy before and during the attack, intensifying the overall impact. However, the attack's impact is not solely determined by non-IIDness. Other factors and parameters, such as the total number of clients, clients selected per round, the presence of malicious clients, and local and global training epochs, collectively influence the magnitude of the attack's impact.

Non-IID data in FL has both positive and negative implications for privacy. On the positive side, the varying data characteristics among clients make it harder for attackers to perform accurate inference attacks or reconstruct sensitive information. However, highly distinct non-IID data distributions could also potentially reveal specific individuals or groups

TABLE V: Comparison and Evaluation of Commonly Used Datasets in FL: Analysis of Privacy-Centric Attacks with respect to Metrics used and their Values. MsI: membership inference, MI: model inference, CI: category inference, LE: local epochs, and GE: global epochs. '-' indicate not applicable or missing values.

Dataset	Point of Attack	Attack	Source of Attack	Total Clients	Clients Selected Per Round	Malicious Clients	Aggregation Technique	LE	GE	Model	Metrics Used	Metric Values
MNIST [152]	MsI	ALGANs [56]	C	100	100	1	FedAvg	10	-	CNN	Accuracy, Recall Rate, F1 Score	98.35%, 88.31%, 93%
		GAN based MsI [125]	C	100	100	1	FedAvg	10	100	CNN	Accuracy, Recall Ratio, F1 Score	97.9%, 87.9%, 93%
	MI	HCGLA [60]	S	-	-	-	FedAvg	1	1000	LeNet	MSE, PSNR, SSIM	3.6 $\times 10^-6$, 102.99, 0.99
		CAFE [14]	S	4	4	-	FedAvg	1	20000	CNN	PSNR	43.15
			S	16	16	-	FedAvg	1	20000	CNN	PSNR	39.28
		GRNN [15]	S	-	-	-	FedAvg	1	2000	ResNet-18	MSE, PSNR, SSIM	0.1, 39.48, -
		User-level Privacy Leakage [16]	S	10	10	-	FedAvg	1	300	CNN	Inception Score, Accuracy	0.61 $\pm$ 0.05, 99%
	CI	Category Privacy [62]	S and/ or C	1000	10	1	FedAvg+Mask	2	500	CNN	Precision, Recall, F1 Score	83.1%, 85.7%, 84.4%
CIFAR10 [151]	MsI	ALGANs [56]	C	100	100	1	FedAvg	10	-	CNN	Accuracy, Recall Rate, F1 Score	93.52%, 85.45%, 89%
		Label Inference [73]	C	2	2	1	FedAvg	-	-	ResNet-18	Top-1 Accuracy	63.42%
	MI	AGIC [58]	S	-	-	-	FedAvg	8	10000	ResNet20-4	PSNR, SSIM, LPIPS	16.18, 0.53, 0.11
		HCGLA [60]	S	-	-	-	FedAvg	1	1000	ResNet-18	MSE, PSNR, SSIM	2.5 $\times 10^-3$, 77.66, 0.98
		CAFE [14]	S	4	4	-	FedAvg	1	8000	CNN	PSNR	33.93
			S	16	16	-	FedAvg	1	8000	CNN	PSNR	28.39
	CI	Category Privacy [62]	S and/ or C	1000	10	1	FedAvg+Mask	2	500	CNN	Precision, Recall, F1 Score	98.1%, 99.2%, 98.6%
CIFAR100 [151]	MsI	Label Inference [73]	C	2	2	1	FedAvg	-	-	ResNet-18	Top-5 Accuracy	47%
	MI	AGIC [58]	S	-	-	-	FedAvg	8	10000	ResNet20-4	PSNR, SSIM, LPIPS	19.13, 0.67, 0.05
		HCGLA [60]	S	-	-	-	FedAvg	1	1000	ResNet-18	MSE, PSNR, SSIM	4.6 $\times 10^-3$, 77.50, 0.96
		GRNN [15]	S	-	-	-	FedAvg	1	2000	LeNet	MSE, PSNR, SSIM	0.25, 38.44, 0.95
ImageNet [157]	MI	AGIC [58]	S	-	-	-	FedAvg	8	10000	ResNet-50	PSNR, SSIM, LPIPS	10.69, 0.18, 0.78
		HCGLA [60]	S	-	-	-	FedAvg	1	1000	LeNet	MSE, PSNR, SSIM	4.4 $\times 10^-3$, 73.44, 0.91
LFW [158]	MI	HCGLA [60]	S	-	-	-	FedAvg	1	1000	ResNet-18	MSE, PSNR, SSIM	0.5 $\times 10^-3$, 81.86, 0.99
		GRNN [15]	S	-	-	-	FedAvg	1	2000	LeNet	MSE, PSNR	0.28 38.17

within the dataset. Attackers might exploit these differences to identify clients with unique characteristics or sensitive attributes. Privacy implications of non-IID data depend on factors like the specific attack scenario, utilization of privacy protection mechanisms (e.g., differential privacy techniques), and the attacker's knowledge and capabilities. Non-IID data alone does not guarantee privacy protection but can add complexity for attackers in certain contexts.

Table IV and Table V show the use of various α values in state-of-the-art attacks and their impact. Most FL utility attacks use α values of 0.5, 0.9, and 1, while privacy-centric attacks typically employ α as 1, indicating an IID data distribution. However, there is a significant research gap in evaluating attacks under different degrees of non-IID to understand their effects on attack impact in real-world settings.

Aggregation Techniques: The widely used federated averaging (FedAvg) [5] for aggregating client updates in FL is complemented by several byzantine-robust aggregation techniques discussed in Section IV and outlined in Table IV and Table V. The choice of aggregation technique is crucial in understanding the attack's robustness, as non-robust aggregation methods may lead to high impact, low budget, and poor detection of visible attacks on the server.

Local and Global Epochs: Our literature analysis reveals a wide range of values for local and global epochs in utility-centric attacks, i.e., 1 to 20 and 40 to 2000, respectively, as shown in Table IV). Similarly, for privacy-centric attacks, the range of local epochs is from 1 to 10, and the range of global epochs is from 100 to 20000, as shown in Table V. This analysis provides insights into the attacker's level of effort and budget required to poison the model and leak the data. Hence, the number of local and global epochs is a critical consideration in the design of FL attacks.

Model Architectures: In Section III-10, we discussed that image classification is a prevalent task in FL attack papers, where model architectures such as VGG-9, ResNet-18, ResNet-50, ResNet20-4, Alexnet, customized CNNs, and fully connected networks (MLP) are commonly used. For other tasks, LSTM architectures were observed (as shown in Table IV and Table V). The choice of model architecture is crucial as

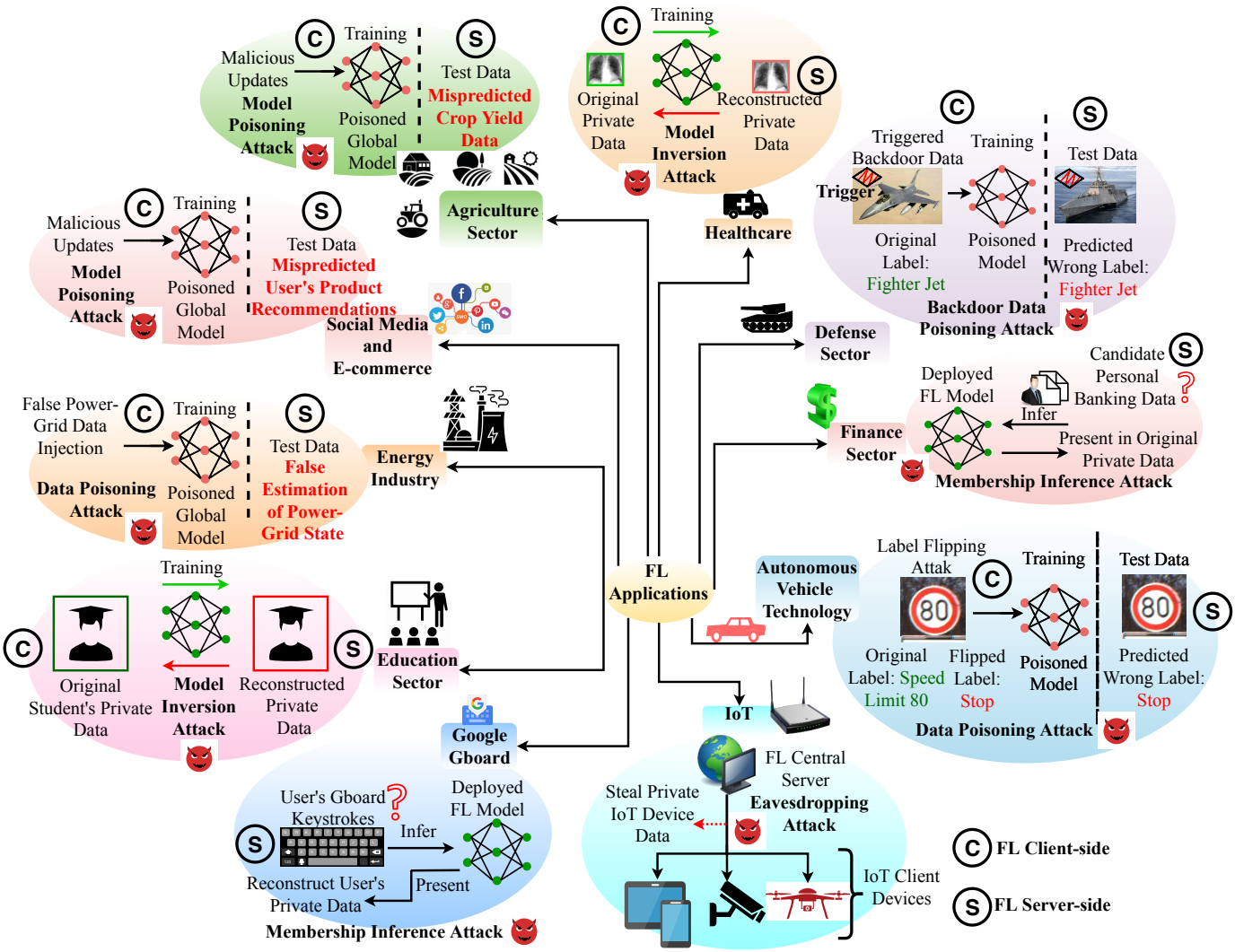


Fig. 6: Overview of FL application domains and corresponding most threatening attacks.

it impacts the attack budget, visibility, and impact. Attacking simpler architectures may have a low budget, high impact, and low visibility, but the base performance without an attack becomes less practical for real-world usage. On the other hand, attacking deeper standard neural networks requires a higher budget and leads to a lower impact due to the difficulty of poisoning them, making the attack less visible. Therefore, attackers must carefully choose a model that provides better base accuracy and aim to attack it with a low budget and low visibility to achieve a higher impact.

6) Case Studies: Exploring the Application of FL across Diverse Domains with Possible Threats: Figure 6 and Table VI offer an overview of FL applications, how FL facilitates their operations, threats to their security and privacy, and research gaps that need to be addressed. Research gaps include privacy-preserving FL, enhancing online learning security, and investigating attack impacts. Despite FL's potential, addressing risks through awareness and regulatory frameworks is essential for mitigating vulnerabilities in FL across diverse domains.

VI. Research Gaps and Potential Future Research Directions

This study extensively reviews the literature on attacks from the attacker's perspective in various FL systems. We identify open challenges and propose potential directions for future research in Table VII. These research gaps present promising opportunities to address the current security and privacy limitations in FL. The areas of focus include novel attacks and defenses considering diverse FL parameters, attack budgets, visibility, impact, and generalizability. Additionally, vulnerabilities in different FL components such as communication protocols, learning algorithms, hyperparameters, compression, and fairness need exploration. Moreover, developing practical and cost-effective defenses is crucial in ensuring FL's security and privacy in real-world applications.

VII. Conclusion

This survey provided a comprehensive overview of the impact of malicious attacks on FL, covering various aspects such as attack impact, budget, visibility, generalizability, real-world applications, attacks and defenses, and attack status. We

TABLE VI: Summary of Case Studies on FL Attacks, Limitations, and Proposed Solutions MI: model inversion, DP: data poisoning, MsI: membership inference, MP: model poisoning, AB: attack budget, AV: attack visibility, and AIm: attack impact.

Case Study ID	Application Domain (Most Threatening FL Attack as shown in Figure 6)	Overview: Rationale for Applying FL	Existing Relevant Attacks from Table II	Our Survey Findings: Limitations in terms of AB, AV, and AIm as per Tables IV, V, and Figures 4, 5	Proposed Solutions
1	Healthcare **(MI)**	FL facilitates collaborative improvement of **personalized medical models while ensuring patient privacy through collaborative model training** [6].	AGIC [58], HCGLA [60], CAFE [14], Zhang *et al.* [75], Feature Inference [76]	- The attacks **have not been applied** to domain-specific datasets. - Evaluation conducted on a **relatively small number** of total clients. - Varied high AIm, low AB, but **no explicit discussion** on AV.	- Apply attack methods to datasets like **Camelyon17 [160], QDS [161]**. - **Scale evaluation** to large number of clients. - **Discuss on AV** to understand robustness against defenses.
2	Education Sector **(MI)**	FL empowers **the education sector** to build **personalized learning models** by enabling secure data collection from diverse students and **privately training** models that **predict individual learning outcomes [161].**			
3	Defense Sector **(DP)**	FL fosters the collaborative development of robust models for critical **defense sector tasks, including target object detection, tracking among military organizations** while safeguarding sensitive data privacy [162].	DLF [11], SLF [11], PoisonGAN [12], LR-BA [57], Graph-Fraudster [63], DBA [68], Internal Evasion [74], Cheng *et al.* [77],	- **Lack of attacks** on domain-specific datasets. - Limited evaluation of attack methods with **varying degrees of non-IID.** - Most attacks have **high AB** leading to **increased AV** from the attacker's perspective.	- Evaluate attacks on **DesertSim [163], Cityscapes [164], Ningxia [165],** for tasks, including **object detection and tracking.** - Investigate the AIm under **varying degrees of non-IID** by using different α values. - Study attack methods with **limited attack data (low AB)** and explore strategies to sustain AIm.
4	Autonomous Vehicle Technology (AVT) **(DP)**	FL is used in **AVT** to develop models for tasks such as **object detection, path planning, etc.,** where different autonomous vehicles collaborate while maintaining data privacy [166].			
5	Energy Industry **(DP)**	FL enables **the energy industry** to develop predictive private models for **energy demand, consumption, and renewable energy forecasting** by leveraging data from diverse sources such as smart meters, IoT devices, and weather sensors [167].			
6	Finance Sector **(MsI)**	**Banks and other financial institutions** can use FL to build predictive models for **credit scoring, fraud detection, and risk management,** while maintaining data privacy [7].	Yang *et al.* [55], ALGANs [56], Label Inference [73]	- Attacks applied to **finance domain** datasets, not **Google Gboard** domain. - Limited to scenarios with a **single malicious client at the client side.**	- Assess attacks on **Logs data [168]** for **keyboard next-word prediction.** - Consider scenarios with **curious and malicious server.**
7	Google Gboard **(MsI)**	Google Gboard is a keyboard application that utilizes **personalized FL to improve its prediction** capabilities of different users ensuring data privacy [168].			
8	IoT **(Eavesdropping)**	In the **IoT domain,** FL is used to develop models for various tasks such as **predictive maintenance, anomaly detection,** etc., by privately collaborating with different IoT devices without sharing sensitive data [169].	Xu *et al.* [66], Yuan *et al.* [170]	**Research gap -** Limited work **on eavesdropping attacks** in FL, particularly on **IoT datasets.**	Explore and **develop eavesdropping attacks** tailored for FL on **IoT datasets** like **TON_IoT [171]** to address the research gap.
9	Social Media and E-commerce **(MP)**	FL is used in **social media and e-commerce to develop private customer product recommendation systems** without compromising user privacy [172].	PGA [11], LIE [17], DYN-OPT [50], ADA [59], RL [65], DBA [68], Stealthy Attack [69], Model Replacement Attack [70], Sybil [61]	- **Applied to E-commerce datasets,** no attacks on Agriculture domain datasets. - Limited evaluation of attacks with **varying degrees of non-IID** and have **high AB,** mainly targeting classifiers, **neglecting** high-end task models.	- Explore attacks on Agriculture domain datasets like **CropDeep [173]**. - Evaluating attacks with **varying degrees** of **non-IID** and lower AB, while also **targeting high-end task models** like object detection and segmentation.
10	Agriculture Sector **(MP)**	FL can **revolutionize the agriculture sector** by enabling models for **crop yield prediction, pest control, and weather forecasting**. It leverages distributed data while ensuring privacy and security improving accuracy and decision-making [174].			

also highlighted the importance of considering the impact, budget, visibility, and generalizability of FL attacks in the design and deployment of FL systems. We identified the research gaps and potential future directions for improving the security and privacy of FL systems. These gaps include developing novel attacks and defenses that balance low attack budget, low visibility, high impact, and high generalizability, investigating the performance of attacks on various real-world application domains beyond image classification, developing defenses that can operate in online practical attack settings close to FL production deployments, exploring low vs high concentrated data in privacy-centric FL attacks using the Dirichlet distribution, and investigating the impact of attacks and defenses on different types of FL. We also discussed the recent advancements in adversarial defenses in FL and highlighted the challenges in securing FL. Overall, this survey paper serves as a valuable resource for researchers, practitioners, and students in the field of FL security and provides insights into the current challenges and future directions for research in this area.

ACKNOWLEDGEMENT

This work was supported by the Indo-Norwegian Collaboration in Autonomous Cyber-Physical Systems (INCAPS) project: 287918 of the International Partnerships for Excellent Education, Research and Innovation (INTPART) program from the Research Council of Norway.

REFERENCES

[1] S. Hong and J. Chae, "Communication-efficient randomized algorithm for multi-kernel online federated learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 9872–9886, 2021.

[2] D. Qiao, G. Liu, S. Guo, and J. He, "Adaptive federated learning for non-convex optimization problems in edge computing environment," *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 5, pp. 3478–3491, 2022.

[3] D. Qiao, S. Guo, D. Liu, S. Long, P. Zhou, and Z. Li, "Adaptive federated deep reinforcement learning for proactive content caching in edge computing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 12, pp. 4767–4782, 2022.

[4] S. Zhou and G. Y. Li, "Federated learning via inexact admm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–10, 2023.

[5] 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*. PMLR, 2017, pp. 1273–1282.

[6] J. Li, Y. Meng, L. Ma, S. Du, H. Zhu, Q. Pei, and X. Shen, "A federated learning based privacy-preserving smart healthcare system," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 3, 2021.

[7] A. Imteaj and M. H. Amini, "Leveraging asynchronous federated learning to predict customers financial distress," *Intelligent Systems with Applications*, vol. 14, p. 200064, 2022.

TABLE VII: Comprehensive Analysis of Open Challenges and Future Directions with Research Gaps for Attacks in FL.

S.No	Future Research Area	Research Gap Statement	Corresponding Technical Terms & Concepts	Relevant References
1	Unique evasion, FTL based, and multi-task learning attacks in FL	Lack of **unique evasion attacks tailored for FL settings,** with existing literature predominantly focusing on poisoning, model inversion, and membership inference attacks. Addressing this gap enhances FL system security and robustness. Further, there is a need to explore potential vulnerabilities and security implications of **transfer learning-based FL systems and multi-task learning** in federated settings.	In FL, **evasion attacks** manipulate test data and pose security risks. **Transfer learning** enhances model generalizability by transferring knowledge from pre-trained models. **Multi-task learning** improves model performance by jointly training on multiple related tasks.	[19], [20], [74], [175], [47], [176]
2	Novel attacks with balanced tradeoffs	Research is needed to develop **novel attacks that strike a balance** between **low AB, low AV, high AIm, high generalizability, and practicality,** specifically tailored for online attack scenarios. Similarly, there is a need for **novel defenses that consider tradeoffs** between budget, visibility, and the **defender's perspective.**	Section III provided insights into the significance of AB, AV, and AIm from the **attacker's perspective.** Similarly, we can analyze the **defender's perspective.**	[172], [175], [17], [50]
3	Low-cost Defenses w.r.t defender's perspective			
4	Attack extension to other high-end tasks	Research gaps exist in extending attacks to **high-end tasks (e.g., detection)** beyond image classification in FL. Also, limited research has been conducted on **application-specific FL attacks,** as discussed in Table VI.	Object detection, segmentation, tracking, next word prediction, domain-specific datasets.	[77], [171], [166], [173]
5	Application specific attacks in FL			
6	Robustness against sybil and hybrid attacks	Sybil attacks involve the creation of multiple fake identities to control the FL system. The development of **robust defenses against Sybil attacks** in FL settings requires further investigation.	Sybil attacks pose a significant threat to the **integrity and security of FL, disrupting the model training** process and compromising model accuracy and privacy.	[61], [177]
7	Studying the effect of client selection, non-IID on FL security	Current research lacks a **comprehensive understanding of client selection's impact** on FL security and the influence of data concentration **(degree of non-IID)** in privacy-centric FL attacks using the Dirichlet distribution [159] as discussed in Section V-5. Further investigation is needed to address these gaps and enhance the understanding of vulnerabilities in FL process.	**Client selection** in FL involves choosing a **subset of clients' updates** for aggregation. Selection criteria depend on **client's performance, data quality,** and **communication capabilities.**	[118], [119], [77], [107], [95], [149]
8	Developing attacks that 1. Exploit vulnerabilities in communication protocols 2. Target the model's hyperparameters 3. Exploit model compression	Existing research lacks **comprehensive exploration of communication protocol vulnerabilities, hyperparameter attacks, and model compression risks** in FL. There is a need for targeted investigations to identify potential threats and develop attacks exploiting these vulnerabilities. Additionally, **effective defense mechanisms** should be devised to mitigate communication -related risks, hyperparameter manipulation, and model compression vulnerabilities in FL systems for enhancing the robustness and privacy of FL in real-world scenarios.	**-** FL involves **communication** between the **central server and participating clients** to exchange model updates. Communication protocols define **how** data and model updates are **transmitted securely and efficiently.** **- Hyperparameters** control the learning process in FL, governing how clients train their models and how the central server aggregates these updates. **-Model compression** in FL aims to reduce the size of ML models, making them more lightweight and efficient for communication and storage, while maintaining model performance.	[5], [60], [22], [178]
9	Explainability and interpretability in FL attacks	The lack of **interpretability and explainability in FL** models raises concerns about their trustworthiness. To address this, there is a research gap in developing methods to enhance the transparency and interpretability of FL attacks **while maintaining their accuracy and performance.**	In FL, **model transparency and interpretability** are crucial for understanding model behavior and building trust. Explaining how FL models make predictions is essential, especially in sensitive domains like healthcare where model **output impacts critical decisions.**	[179], [180]
10	Hardware-based Attacks in FL	**Memory leaks, side-channel attacks, and hardware trojans** pose potential threats, and addressing this research gap would enhance the security and resilience of FL systems. Also, there is a need to explore and develop hardware-based defenses that can **effectively safeguard various attacks targeting the underlying hardware.**	The **underlying hardware** in FL encompasses the **physical components, devices,** and **infrastructure** involved in the system. This includes **servers, client devices** (e.g., smartphones, laptops), **network connections, storage systems,** and other hardware elements essential for the training process.	[181]

[8] W. Peng, R. Liu, R. Wang, T. Cheng, Z. Wu, L. Cai, and W. Zhou, "Ensemblefool: A method to generate adversarial examples based on model fusion strategy," *Computers & Security*, vol. 107, p. 102317, 2021.

[9] S. He, R. Wang, T. Liu, C. Yi, X. Jin, R. Liu, and W. Zhou, "Type-i generative adversarial attack," *IEEE Transactions on Dependable and Secure Computing*, 2022.

[10] D. Usynin, A. Ziller, M. Makowski, R. Braren, D. Rueckert, B. Glocker, G. Kaissis, and J. Passerat-Palmbach, "Adversarial interference and its mitigations in privacy-preserving collaborative machine learning," *Nature Machine Intelligence*, vol. 3, no. 9, pp. 749–758, 2021.

[11] V. Shejwalkar, A. Houmansadr, P. Kairouz, and D. Ramage, "Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 1354–1371.

[12] 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 Internet of Things Journal*, vol. 8, no. 5, pp. 3310–3322, 2020.

[13] M. Fang, X. Cao, J. Jia, and N. Z. Gong, "Local model poisoning attacks to byzantine-robust federated learning," in *Proceedings of the 29th USENIX Conference on Security Symposium*, 2020, pp. 1623–1640.

[14] X. Jin, P.-Y. Chen, C.-Y. Hsu, C.-M. Yu, and T. Chen, "Cafe: Catastrophic data leakage in vertical federated learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 994–1006, 2021.

[15] H. Ren, J. Deng, and X. Xie, "Grnn: generative regression neural network—a data leakage attack for federated learning," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 4, pp. 1–24, 2022.

[16] Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, and H. Qi, "Beyond inferring class representatives: User-level privacy leakage from federated learning," in *IEEE INFOCOM 2019-IEEE conference on computer communications*. IEEE, 2019, pp. 2512–2520.

[17] G. Baruch, M. Baruch, and Y. Goldberg, "A little is enough: Circumventing defenses for distributed learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[18] G. Sun, Y. Cong, J. Dong, Q. Wang, L. Lyu, and J. Liu, "Data poisoning attacks on federated machine learning," *IEEE Internet of Things Journal*, vol. 9, no. 13, pp. 11 365–11 375, 2021.

[19] S. Wang, R. Sahay, and C. G. Brinton, "How potent are evasion attacks for poisoning federated learning-based signal classifiers?" *arXiv preprint arXiv:2301.08866*, 2023.

[20] T. Kim, S. Singh, N. Madaan, and C. Joe-Wong, "pfeddef: Characterizing evasion attack transferability in federated learning," *Software Impacts*, p. 100469, 2023.

[21] D. Javeed, U. MohammedBadamasi, C. O. Ndubuisi, F. Soomro, and M. Asif, "Man in the middle attacks: Analysis, motivation and prevention," *International Journal of Computer Networks and Communications Security*, vol. 8, no. 7, pp. 52–58, 2020.

[22] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Communications Surveys Tutorials*, vol. 22, no. 3, pp. 2031–2063, 2020.

[23] J. Zhang, M. Li, S. Zeng, B. Xie, and D. Zhao, "A survey on security and privacy threats to federated learning," in *2021 International Conference on Networking and Network Applications (NaNA)*. IEEE, 2021, pp. 319–326.

[24] M. Yang, Y. He, and J. Qiao, "Federated learning-based privacy-preserving and security: Survey," in *2021 Computing, Communications and IoT Applications (ComComAp)*. IEEE, 2021, pp. 312–317.

[25] N. Bouacida and P. Mohapatra, "Vulnerabilities in federated learning," *IEEE Access*, vol. 9, pp. 63 229–63 249, 2021.

[26] A. Blanco-Justicia, J. Domingo-Ferrer, S. Martínez, D. Sánchez,

A. Flanagan, and K. E. Tan, "Achieving security and privacy in federated learning systems: Survey, research challenges and future directions," *Engineering Applications of Artificial Intelligence*, vol. 106, p. 104468, 2021.

[27] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, "A survey on security and privacy of federated learning," *Future Generation Computer Systems*, vol. 115, pp. 619–640, 2021.

[28] M. S. Jere, T. Farnan, and F. Koushanfar, "A taxonomy of attacks on federated learning," *IEEE Security Privacy*, vol. 19, no. 2, pp. 20–28, 2021.

[29] J. H. Yoo, H. Jeong, J. Lee, and T.-M. Chung, "Federated learning: Issues in medical application," in *Future Data and Security Engineering: 8th International Conference, FDSE 2021, Virtual Event, November 24–26, 2021, Proceedings 8*. Springer, 2021, pp. 3–22.

[30] P. Liu, X. Xu, and W. Wang, "Threats, attacks and defenses to federated learning: issues, taxonomy and perspectives," *Cybersecurity*, vol. 5, no. 1, pp. 1–19, 2022.

[31] Z. Wang, Q. Kang, X. Zhang, and Q. Hu, "Defense strategies toward model poisoning attacks in federated learning: A survey," in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2022, pp. 548–553.

[32] Y. Chen, Y. Gui, H. Lin, W. Gan, and Y. Wu, "Federated learning attacks and defenses: A survey," in *2022 IEEE International Conference on Big Data (Big Data)*, 2022, pp. 4256–4265.

[33] L. Lyu, H. Yu, X. Ma, C. Chen, L. Sun, J. Zhao, Q. Yang, and S. Y. Philip, "Privacy and robustness in federated learning: Attacks and defenses," *IEEE transactions on neural networks and learning systems*, 2022.

[34] D. Jatain, V. Singh, and N. Dahiya, "A contemplative perspective on federated machine learning: Taxonomy, threats & vulnerability assessment and challenges," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 9, pp. 6681–6698, 2022.

[35] K. Zhang, X. Song, C. Zhang, and S. Yu, "Challenges and future directions of secure federated learning: a survey," *Frontiers of computer science*, vol. 16, pp. 1–8, 2022.

[36] A. Qammar, J. Ding, and H. Ning, "Federated learning attack surface: taxonomy, cyber defences, challenges, and future directions," *Artificial Intelligence Review*, pp. 1–38, 2022.

[37] B. Ghimire and D. B. Rawat, "Recent advances on federated learning for cybersecurity and cybersecurity for federated learning for internet of things," *IEEE Internet of Things Journal*, vol. 9, no. 11, pp. 8229–8249, 2022.

[38] M. Benmalek, M. A. Benrekia, and Y. Challal, "Security of federated learning: attacks, defensive mechanisms, and challenges," *Revue des Sciences et Technologies de l'Information-Série RIA: Revue d'Intelligence Artificielle*, vol. 36, no. 1, pp. 49–59, 2022.

[39] J. Zhang, H. Zhu, F. Wang, J. Zhao, Q. Xu, H. Li *et al.*, "Security and privacy threats to federated learning: Issues, methods, and challenges," *Security and Communication Networks*, vol. 2022, 2022.

[40] N. Rodríguez-Barroso, D. Jiménez-López, M. V. Luzón, F. Herrera, and E. Martínez-Cámara, "Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges," *Information Fusion*, vol. 90, pp. 148–173, 2023.

[41] G. Xia, J. Chen, C. Yu, and J. Ma, "Poisoning attacks in federated learning: A survey," *IEEE Access*, 2023.

[42] A. K. Nair, E. D. Raj, and J. Sahoo, "A robust analysis of adversarial attacks on federated learning environments," *Computer Standards & Interfaces*, p. 103723, 2023.

[43] E. Hallaji, R. Razavi-Far, and M. Saif, *Federated and Transfer Learning: A Survey on Adversaries and Defense Mechanisms*. Cham: Springer International Publishing, 2023, pp. 29–55. [Online]. Available: https://doi.org/10.1007/978-3-031-11748-0_3

[44] H. S. Sikandar, H. Waheed, S. Tahir, S. U. Malik, and W. Rafique, "A detailed survey on federated learning attacks and defenses," *Electronics*, vol. 12, no. 2, p. 260, 2023.

[45] J. Guo, H. Li, F. Huang, Z. Liu, Y. Peng, X. Li, J. Ma, V. G. Menon, and K. K. Igorevich, "Adfl: A poisoning attack defense framework for horizontal federated learning," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 10, pp. 6526–6536, 2022.

[46] T. Zou, Y. Liu, Y. Kang, W. Liu, Y. He, Z. Yi, Q. Yang, and Y.-Q. Zhang, "Defending batch-level label inference and replacement attacks in vertical federated learning," *IEEE Transactions on Big Data*, 2022.

[47] Y. Liu, Y. Kang, C. Xing, T. Chen, and Q. Yang, "A secure federated transfer learning framework," *IEEE Intelligent Systems*, vol. 35, no. 4, pp. 70–82, 2020.

[48] W. Xiong and R. Lagerström, "Threat modeling–a systematic literature review," *Computers & security*, vol. 84, pp. 53–69, 2019.

[49] B. Li, L. Fan, H. Gu, J. Li, and Q. Yang, "Fedipr: Ownership verification for federated deep neural network models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4521–4536, 2023.

[50] V. Shejwalkar and A. Houmansadr, "Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning," in *NDSS*, 2021.

[51] T. Kim, S. Singh, N. Madaan, and C. Joe-Wong, "pfeddef: Defending grey-box attacks for personalized federated learning," *arXiv preprint arXiv:2209.08412*, 2022.

[52] K. N. Kumar, C. Vishnu, R. Mitra, and C. K. Mohan, "Black-box adversarial attacks in autonomous vehicle technology," in *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, Oct 2020, pp. 1–7.

[53] O. Zari, C. Xu, and G. Neglia, "Efficient passive membership inference attack in federated learning," *arXiv preprint arXiv:2111.00430*, 2021.

[54] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 739–753.

[55] R. Yang, J. Ma, J. Zhang, S. Kumari, S. Kumar, and J. J. Rodrigues, "Practical feature inference attack in vertical federated learning during prediction in artificial internet of things," *IEEE Internet of Things Journal*, 2023.

[56] Y. Xie, B. Chen, J. Zhang, and W. Li, "Algans: Enhancing membership inference attacks in federated learning with gans and active learning," in *2022 IEEE International Symposium on Product Compliance Engineering-Asia (ISPCE-ASIA)*. IEEE, 2022, pp. 1–6.

[57] Y. Gu and Y. Bai, "Lr-ba: Backdoor attack against vertical federated learning using local latent representations," *Computers & Security*, vol. 129, p. 103193, 2023.

[58] J. Xu, C. Hong, J. Huang, L. Y. Chen, and J. Decouchant, "Agic: Approximate gradient inversion attack on federated learning," in *2022 41st International Symposium on Reliable Distributed Systems (SRDS)*. IEEE, 2022, pp. 12–22.

[59] Y. Sun, H. Ochiai, and J. Sakuma, "Semi-targeted model poisoning attack on federated learning via backward error analysis," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 1–8.

[60] H. Yang, M. Ge, K. Xiang, and J. Li, "Using highly compressed gradients in federated learning for data reconstruction attacks," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 818–830, 2023.

[61] Y. Jiang, Y. Li, Y. Zhou, and X. Zheng, "Sybil attacks and defense on differential privacy based federated learning," in *2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE, 2021, pp. 355–362.

[62] J. Gao, B. Hou, X. Guo, Z. Liu, Y. Zhang, K. Chen, and J. Li, "Secure aggregation is insecure: Category inference attack on federated learning," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 1, pp. 147–160, 2023.

[63] J. Chen, G. Huang, H. Zheng, S. Yu, W. Jiang, and C. Cui, "Graph-fraudster: Adversarial attacks on graph neural network-based vertical federated learning," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 2, pp. 492–506, 2022.

[64] C. Chen, L. Lyu, H. Yu, and G. Chen, "Practical attribute reconstruction attack against federated learning," *IEEE Transactions on Big Data*, pp. 1–1, 2022.

[65] H. Li, X. Sun, and Z. Zheng, "Learning to attack federated learning: A model-based reinforcement learning attack framework," in *Advances in Neural Information Processing Systems*, 2022.

[66] C. Xu and G. Neglia, "What else is leaked when eavesdropping federated learning?" in *CCS workshop Privacy Preserving Machine Learning (PPML)*, 2021.

[67] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J.-y. Sohn, K. Lee, and D. Papailiopoulos, "Attack of the tails: Yes, you really can backdoor federated learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 070–16 084, 2020.

[68] C. Xie, K. Huang, P.-Y. Chen, and B. Li, "Dba: Distributed backdoor attacks against federated learning," in *International conference on learning representations*, 2020.

[69] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Analyzing federated learning through an adversarial lens," in *International Conference on Machine Learning*. PMLR, 2019, pp. 634–643.

[70] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 2938–2948.

[71] F. Lau, S. H. Rubin, M. H. Smith, and L. Trajkovic, "Distributed denial of service attacks," in *Smc 2000 conference proceedings. 2000 ieee international conference on systems, man and cybernetics.'cybernetics evolving to systems, humans, organizations, and their complex interactions'(cat. no. 0*, vol. 3. IEEE, 2000, pp. 2275–2280.

[72] W. Xu, K. Ma, W. Trappe, and Y. Zhang, "Jamming sensor networks: attack and defense strategies," *IEEE network*, vol. 20, no. 3, pp. 41–47, 2006.

[73] C. Fu, X. Zhang, S. Ji, J. Chen, J. Wu, S. Guo, J. Zhou, A. X. Liu, and T. Wang, "Label inference attacks against vertical federated learning," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 1397–1414.

[74] T. Kim, S. Singh, N. Madaan, and C. Joe-Wong, "Characterizing internal evasion attacks in federated learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023, pp. 907–921.

[75] C. Zhang, H. Liang, Y. Li, T. Wu, L. Zhu, and W. Zhang, "Stealing secrecy from outside: A novel gradient inversion attack in federated learning," in *2022 IEEE 28th International Conference on Parallel and Distributed Systems (ICPADS)*. IEEE, 2023, pp. 282–288.

[76] X. Luo, Y. Wu, X. Xiao, and B. C. Ooi, "Feature inference attack on model predictions in vertical federated learning," in *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 2021, pp. 181–192.

[77] Y. Cheng, J. Lu, D. Niyato, B. Lyu, J. Kang, and S. Zhu, "Federated transfer learning with client selection for intrusion detection in mobile edge computing," *IEEE Communications Letters*, vol. 26, no. 3, pp. 552–556, 2022.

[78] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep models under the gan: information leakage from collaborative deep learning," in *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, 2017, pp. 603–618.

[79] A. Triastcyn and B. Faltings, "Federated learning with bayesian differential privacy," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 2587–2596.

[80] N. Truong, K. Sun, S. Wang, F. Guitton, and Y. Guo, "Privacy preservation in federated learning: An insightful survey from the gdpr perspective," *Computers & Security*, vol. 110, p. 102402, 2021.

[81] G. Zizzo, A. Rawat, M. Sinn, and B. Buesser, "Fat: Federated adversarial training," *arXiv preprint arXiv:2012.01791*, 2020.

[82] D. Shah, P. Dube, S. Chakraborty, and A. Verma, "Adversarial training in communication constrained federated learning," *arXiv preprint arXiv:2103.01319*, 2021.

[83] E. Hallaji, R. Razavi-Far, M. Saif, and E. Herrera-Viedma, "Label noise analysis meets adversarial training: A defense against label poisoning in federated learning," *Knowledge-Based Systems*, p. 110384, 2023.

[84] Y. Jiang, S. Wang, V. Valls, B. J. Ko, W.-H. Lee, K. K. Leung, and L. Tassiulas, "Model pruning enables efficient federated learning on edge devices," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[85] S. Liu, G. Yu, R. Yin, and J. Yuan, "Adaptive network pruning for wireless federated learning," *IEEE Wireless Communications Letters*, vol. 10, no. 7, pp. 1572–1576, 2021.

[86] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," *Advances in neural information processing systems*, vol. 30, 2017.

[87] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5650–5659.

[88] R. Guerraoui, S. Rouault *et al.*, "The hidden vulnerability of distributed learning in byzantium," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3521–3530.

[89] M. S. Ozdayi, M. Kantarcioglu, and Y. R. Gel, "Defending against backdoors in federated learning with robust learning rate," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, 2021, pp. 9268–9276.

[90] Z. Chen, P. Tian, W. Liao, and W. Yu, "Zero knowledge clustering based adversarial mitigation in heterogeneous federated learning," *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 2, pp. 1070–1083, 2020.

[91] Z. Ma, J. Ma, Y. Miao, Y. Li, and R. H. Deng, "Shieldfl: Mitigating model poisoning attacks in privacy-preserving federated learning," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 1639–1654, 2022.

[92] L. Muñoz-González, K. T. Co, and E. C. Lupu, "Byzantine-robust federated machine learning through adaptive model averaging," *arXiv preprint arXiv:1909.05125*, 2019.

[93] X. Cao, M. Fang, J. Liu, and N. Z. Gong, "Fltrust: Byzantine-robust federated learning via trust bootstrapping," *arXiv preprint arXiv:2012.13995*, 2020.

[94] X. Jiang, S. Sun, Y. Wang, and M. Liu, "Towards federated learning against noisy labels via local self-regularization," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 862–873.

[95] Z. Chen, Z. Wu, X. Wu, L. Zhang, J. Zhao, Y. Yan, and Y. Zheng, "Contractible regularization for federated learning on non-iid data," in *2022 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2022, pp. 61–70.

[96] T. Su, M. Wang, and Z. Wang, "Federated regularization learning: An accurate and safe method for federated learning," in *2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS)*. IEEE, 2021, pp. 1–4.

[97] X. Li, Z. Qu, S. Zhao, B. Tang, Z. Lu, and Y. Liu, "Lomar: A local defense against poisoning attack on federated learning," *IEEE Transactions on Dependable and Secure Computing*, 2021.

[98] C. Zhao, Y. Wen, S. Li, F. Liu, and D. Meng, "Federatedreverse: A detection and defense method against backdoor attacks in federated learning," in *Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security*, 2021, pp. 51–62.

[99] N. Rodríguez-Barroso, E. Martínez-Cámara, M. V. Luzón, and F. Herrera, "Dynamic defense against byzantine poisoning attacks in federated learning," *Future Generation Computer Systems*, vol. 133, pp. 1–9, 2022.

[100] Z. Zhang, Y. Zhang, D. Guo, L. Yao, and Z. Li, "Secfednids: Robust defense for poisoning attack against federated learning-based network intrusion detection system," *Future Generation Computer Systems*, vol. 134, pp. 154–169, 2022.

[101] S. Lu, R. Li, W. Liu, and X. Chen, "Defense against backdoor attack in federated learning," *Computers & Security*, vol. 121, p. 102819, 2022.

[102] W. Wan, J. Lu, S. Hu, L. Y. Zhang, and X. Pei, "Shielding federated learning: A new attack approach and its defense," in *2021 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2021, pp. 1–7.

[103] S. Andreina, G. A. Marson, H. Möllering, and G. Karame, "Baffle: Backdoor detection via feedback-based federated learning," in *2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2021, pp. 852–863.

[104] D. Li, W. E. Wong, W. Wang, Y. Yao, and M. Chau, "Detection and mitigation of label-flipping attacks in federated learning systems with kpca and k-means," in *2021 8th International Conference on Dependable Systems and Their Applications (DSA)*. IEEE, 2021, pp. 551–559.

[105] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," *Advances in neural information processing systems*, vol. 31, 2018.

[106] X. Cao, J. Jia, and N. Z. Gong, "Provably secure federated learning against malicious clients," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 8, 2021, pp. 6885–6893.

[107] R. Balakrishnan, T. Li, T. Zhou, N. Himayat, V. Smith, and J. Bilmes, "Diverse client selection for federated learning via submodular maximization," in *International Conference on Learning Representations*, 2022.

[108] A. Kumar, V. Khimani, D. Chatzopoulos, and P. Hui, "Fedclean: A defense mechanism against parameter poisoning attacks in federated learning," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4333–4337.

[109] H. Guo, H. Wang, T. Song, Y. Hua, Z. Lv, X. Jin, Z. Xue, R. Ma, and H. Guan, "Siren: Byzantine-robust federated learning via proactive alarming," in *Proceedings of the ACM Symposium on Cloud Computing*, 2021, pp. 47–60.

[110] P. Rieger, T. D. Nguyen, M. Miettinen, and A.-R. Sadeghi, "Deepsight: Mitigating backdoor attacks in federated learning through deep model inspection," *arXiv preprint arXiv:2201.00763*, 2022.

[111] M. Shayan, C. Fung, C. J. Yoon, and I. Beschastnikh, "Biscotti: A blockchain system for private and secure federated learning," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 7, pp. 1513–1525, 2020.

[112] N. M. Jebreel and J. Domingo-Ferrer, "Fl-defender: Combating targeted attacks in federated learning," *Knowledge-Based Systems*, vol. 260, p. 110178, 2023.

[113] S. Li, Y. Cheng, W. Wang, Y. Liu, and T. Chen, "Learning to detect malicious clients for robust federated learning," *arXiv preprint arXiv:2002.00211*, 2020.

[114] A. Gouissem, K. Abualsaud, E. Yaacoub, T. Khattab, and M. Guizani, "Federated learning stability under byzantine attacks," in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2022, pp. 572–577.

[115] S. Li, E. Ngai, and T. Voigt, "Byzantine-robust aggregation in federated learning empowered industrial iot," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 2, pp. 1165–1175, 2021.

[116] Z. Zhang, X. Cao, J. Jia, and N. Z. Gong, "Fldetector: Defending federated learning against model poisoning attacks via detecting malicious clients," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 2545–2555.

[117] A. Roy Chowdhury, C. Guo, S. Jha, and L. van der Maaten, "Eiffel: Ensuring integrity for federated learning," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022, pp. 2535–2549.

[118] Y. Deng, F. Lyu, J. Ren, H. Wu, Y. Zhou, Y. Zhang, and X. Shen, "Auction: Automated and quality-aware client selection framework for efficient federated learning," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 8, pp. 1996–2009, 2021.

[119] Z. Li, Y. He, H. Yu, J. Kang, X. Li, Z. Xu, and D. Niyato, "Data heterogeneity-robust federated learning via group client selection in industrial iot," *IEEE Internet of Things Journal*, vol. 9, no. 18, pp. 17 844–17 857, 2022.

[120] J. Sun, A. Li, L. DiValentin, A. Hassanzadeh, Y. Chen, and H. Li, "Fl-wbc: Enhancing robustness against model poisoning attacks in federated learning from a client perspective," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 613–12 624, 2021.

[121] A. Panda, S. Mahloujifar, A. N. Bhagoji, S. Chakraborty, and P. Mittal, "Sparsefed: Mitigating model poisoning attacks in federated learning with sparsification," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 7587–7624.

[122] X. Liu, H. Li, G. Xu, Z. Chen, X. Huang, and R. Lu, "Privacy-enhanced federated learning against poisoning adversaries," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 4574–4588, 2021.

[123] C. Zhang, S. Li, J. Xia, W. Wang, F. Yan, and Y. Liu, "Batchcrypt: Efficient homomorphic encryption for cross-silo federated learning," in *Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC 2020)*, 2020.

[124] L. Zhang, J. Xu, P. Vijayakumar, P. K. Sharma, and U. Ghosh, "Homomorphic encryption-based privacy-preserving federated learning in iot-enabled healthcare system," *IEEE Transactions on Network Science and Engineering*, 2022.

[125] J. Zhang, J. Zhang, J. Chen, and S. Yu, "Gan enhanced membership inference: A passive local attack in federated learning," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*. IEEE, 2020, pp. 1–6.

[126] Z. Zhu, J. Hong, and J. Zhou, "Data-free knowledge distillation for heterogeneous federated learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 878–12 889.

[127] C. Wu, F. Wu, L. Lyu, Y. Huang, and X. Xie, "Communication-efficient federated learning via knowledge distillation," *Nature communications*, vol. 13, no. 1, p. 2032, 2022.

[128] L. Zhang, L. Shen, L. Ding, D. Tao, and L.-Y. Duan, "Fine-tuning global model via data-free knowledge distillation for non-iid federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 174–10 183.

[129] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," *arXiv preprint arXiv:1610.05755*, 2016.

[130] C. Zhang, S. Ekanut, L. Zhen, and Z. Li, "Augmented multi-party computation against gradient leakage in federated learning," *IEEE Transactions on Big Data*, 2022.

[131] V. Mugunthan, A. Polychroniadou, D. Byrd, and T. H. Balch, "Smpai: Secure multi-party computation for federated learning," in *Proceedings of the NeurIPS 2019 Workshop on Robust AI in Financial Services*, 2019.

[132] D. Byrd and A. Polychroniadou, "Differentially private secure multi-party computation for federated learning in financial applications," in *Proceedings of the First ACM International Conference on AI in Finance*, 2020, pp. 1–9.

[133] A. Mondal, Y. More, R. H. Rooparaghunath, and D. Gupta, "Poster: Flatee: Federated learning across trusted execution environments," in *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2021, pp. 707–709.

[134] Y. Chen, F. Luo, T. Li, T. Xiang, Z. Liu, and J. Li, "A training-integrity privacy-preserving federated learning scheme with trusted execution environment," *Information Sciences*, vol. 522, pp. 69–79, 2020.

[135] F. Mo, H. Haddadi, K. Katevas, E. Marin, D. Perino, and N. Kourtellis, "Ppfl: privacy-preserving federated learning with trusted execution environments," in *Proceedings of the 19th annual international conference on mobile systems, applications, and services*, 2021, pp. 94–108.

[136] S. Otoum, N. Guizani, and H. Mouftah, "On the feasibility of split learning, transfer learning and federated learning for preserving security in its systems," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[137] V. Turina, Z. Zhang, F. Esposito, and I. Matta, "Federated or split? a performance and privacy analysis of hybrid split and federated learning architectures," in *2021 IEEE 14th International Conference on Cloud Computing (CLOUD)*. IEEE, 2021, pp. 250–260.

[138] C. Thapa, P. C. M. Arachchige, S. Camtepe, and L. Sun, "Splitfed: When federated learning meets split learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 8, 2022, pp. 8485–8493.

[139] J. Liao, Z. Chen, and E. G. Larsson, "Over-the-air federated learning with privacy protection via correlated additive perturbations," in *2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2022, pp. 1–8.

[140] X. Yang, Y. Feng, W. Fang, J. Shao, X. Tang, S.-T. Xia, and R. Lu, "An accuracy-lossless perturbation method for defending privacy attacks in federated learning," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 732–742.

[141] J. Wang, S. Guo, X. Xie, and H. Qi, "Protect privacy from gradient leakage attack in federated learning," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 2022, pp. 580–589.

[142] J. Sun, A. Li, B. Wang, H. Yang, H. Li, and Y. Chen, "Soteria: Provable defense against privacy leakage in federated learning from representation perspective," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9311–9319.

[143] H. Lee, J. Kim, R. Hussain, S. Cho, and J. Son, "On defensive neural networks against inference attack in federated learning," in *ICC 2021-IEEE International Conference on Communications*. IEEE, 2021, pp. 1–6.

[144] Y. Huang, S. Gupta, Z. Song, K. Li, and S. Arora, "Evaluating gradient inversion attacks and defenses in federated learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 7232–7241, 2021.

[145] Y. Kaya, S. Hong, and T. Dumitras, "On the effectiveness of regularization against membership inference attacks," *arXiv preprint arXiv:2006.05336*, 2020.

[146] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020.

[147] R. Hu, Y. Guo, H. Li, Q. Pei, and Y. Gong, "Personalized federated learning with differential privacy," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 9530–9539, 2020.

[148] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, "Certified robustness to adversarial examples with differential privacy," in *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019, pp. 656–672.

[149] Z. Xiong, Z. Cai, D. Takabi, and W. Li, "Privacy threat and defense for federated learning with non-iid data in aiot," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 2, pp. 1310–1321, 2021.

[150] T. D. Nguyen, P. Rieger, M. Miettinen, and A.-R. Sadeghi, "Poisoning attacks on federated learning-based iot intrusion detection system," in *Proc. Workshop Decentralized IoT Syst. Secur.(DISS)*, 2020, pp. 1–7.

[151] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[152] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[153] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[154] W. C. DMDave, Todd B, "Acquire valued shoppers challenge," 2014. [Online]. Available: https://kaggle.com/competitions/acquire-valued-shoppers-challenge

[155] S. Caldas, S. M. K. Duddu, P. Wu, T. Li, J. Konečnỳ, H. B. McMahan, V. Smith, and A. Talwalkar, "Leaf: A benchmark for federated settings," *arXiv preprint arXiv:1812.01097*, 2018.

[156] G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik, "Emnist: Extending mnist to handwritten letters," in *2017 international joint conference on neural networks (IJCNN)*. IEEE, 2017, pp. 2921–2926.
[157] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.
[158] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.
[159] T. Minka, "Estimating a dirichlet distribution," 2000.
[160] P. Bandi, O. Geessink, Q. Manson, M. Van Dijk, M. Balkenhol, M. Hermsen, B. E. Bejnordi, B. Lee, K. Paeng, A. Zhong *et al.*, "From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge," *IEEE transactions on medical imaging*, vol. 38, no. 2, pp. 550–560, 2018.
[161] Y.-W. Chu, S. Hosseinalipour, E. Tenorio, L. Cruz, K. Douglas, A. Lan, and C. Brinton, "Mitigating biases in student performance prediction via attention-based personalized federated learning," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 3033–3042.
[162] N. Razmi, B. Matthiesen, A. Dekorsy, and P. Popovski, "Ground-assisted federated learning in leo satellite constellations," *IEEE Wireless Communications Letters*, vol. 11, no. 4, pp. 717–721, 2022.
[163] S. D. Vanstone, "Synthetically generated image dataset for military relevant machine learning experiments," in *Automatic Target Recognition XXXIII*, vol. 12521. SPIE, 2023, pp. 56–68.
[164] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.
[165] T. Cebecauer and M. Suri, "Typical meteorological year data: Solargis approach," *Energy Procedia*, vol. 69, pp. 1958–1969, 2015.
[166] Y. Li, X. Tao, X. Zhang, J. Liu, and J. Xu, "Privacy-preserved federated learning for autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 8423–8434, 2021.
[167] X. Zhang, F. Fang, and J. Wang, "Probabilistic solar irradiation forecasting based on variational bayesian inference with secure federated learning," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 11, pp. 7849–7859, 2020.
[168] A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage, "Federated learning for mobile keyboard prediction," *arXiv preprint arXiv:1811.03604*, 2018.
[169] J. He, S. Guo, D. Qiao, and L. Yi, "Hetefl: Network-aware federated learning optimization in heterogeneous mec-enabled internet of things," *IEEE Internet of Things Journal*, vol. 9, no. 15, pp. 14 073–14 086, 2022.
[170] X. Yuan, X. Ma, L. Zhang, Y. Fang, and D. Wu, "Beyond class-level privacy leakage: Breaking record-level privacy in federated learning," *IEEE Internet of Things Journal*, vol. 9, no. 4, pp. 2555–2565, 2021.
[171] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and A. Anwar, "Ton_iot telemetry dataset: A new generation dataset of iot and iiot for data-driven intrusion detection systems," *Ieee Access*, vol. 8, pp. 165 130–165 150, 2020.
[172] J. Li, T. Cui, K. Yang, R. Yuan, L. He, and M. Li, "Demand forecasting of e-commerce enterprises based on horizontal federated learning from the perspective of sustainable development," *Sustainability*, vol. 13, no. 23, p. 13050, 2021.
[173] Y.-Y. Zheng, J.-L. Kong, X.-B. Jin, X.-Y. Wang, T.-L. Su, and M. Zuo, "Cropdeep: The crop vision dataset for deep-learning-based classification and detection in precision agriculture," *Sensors*, vol. 19, no. 5, p. 1058, 2019.
[174] P. Kumar, G. P. Gupta, and R. Tripathi, "Pefl: Deep privacy-encoding-based federated learning framework for smart agriculture," *IEEE Micro*, vol. 42, no. 1, pp. 33–40, 2021.
[175] O. Suciu, R. Marginean, Y. Kaya, H. Daume III, and T. Dumitras, "When does machine learning {FAIL}? generalized transferability for evasion and poisoning attacks," in *27th USENIX Security Symposium (USENIX Security 18)*, 2018, pp. 1299–1316.
[176] E. Hallaji, R. Razavi-Far, and M. Saif, "Federated and transfer learning: A survey on adversaries and defense mechanisms," in *Federated and Transfer Learning*. Springer, 2022, pp. 29–55.
[177] X. Xiao, Z. Tang, C. Li, B. Xiao, and K. Li, "Sca: sybil-based collusion attacks of iiot data poisoning in federated learning," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 3, pp. 2608–2618, 2022.
[178] S. M. Shah and V. K. Lau, "Model compression for communication efficient federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
[179] S. Salim, B. Turnbull, and N. Moustafa, "A blockchain-enabled explainable federated learning for securing internet-of-things-based social media 3.0 networks," *IEEE Transactions on Computational Social Systems*, 2021.
[180] Z. Qin, L. Yang, Q. Wang, Y. Han, and Q. Hu, "Reliable and interpretable personalized federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 422–20 431.
[181] J. Zhang, X. Cheng, W. Wang, L. Yang, J. Hu, and K. Chen, "{FLASH}: Towards a high-performance hardware acceleration architecture for cross-silo federated learning," in *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, 2023, pp. 1057–1079.

K Naveen Kumar received a B.Tech degree in Computer Science Engineering from the Indian Institute of Information Technology Vadodara (IIITV), Gujarat, India, in 2018. Received M.Tech degree from Indian Institute of Technology Hyderabad (IITH), India in CSE, in 2020. He is pursuing a PhD at IIT Hyderabad in the Department of Computer Science Engineering. Research interests include adversarial attacks and defenses for federated learning and adversarial machine learning.

C Krishna Mohan (Senior Member, IEEE) received the B.Sc.Ed. degree from the Regional Institute of Education, India, in 1988, the M.C.A. degree from the S. J. College of Engineering, India, in 1991, the M.Tech. degree in system analysis and computer applications from the National Institute of Technology Surathkal, India, in 2000, and a PhD degree in computer science and engineering from the Indian Institute of Technology Madras, India, in 2007. He is currently a Professor at the Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad, India. His research interests include video content analysis, pattern recognition, and neural networks.

Linga Reddy Cenkeramaddi (Senior Member, IEEE) received the master's degree in electrical engineering from the Indian Institute of Technology Delhi (IIT Delhi), New Delhi, India, in 2004, and the Ph.D. degree in electrical engineering from the Norwegian University of Science and Technology (NTNU), Trondheim, Norway, in 2011. He worked on mixed-signal circuit design at Texas Instruments. He worked on radiation imaging for an atmosphere-space interaction monitor (ASIM mission to the International Space Station) at the University of Bergen, Bergen, Norway, from 2010 to 2012. He is currently the Leader of the Autonomous and Cyber-Physical Systems (ACPS) Research Group and a Professor with the University of Agder, Grimstad, Norway. Several of his master's students received the Best Master Thesis Awards in information and communication technology (ICT). He has coauthored over 120 research publications that have been published in prestigious international journals and standard conferences. His main scientific interests include cyber-physical systems, autonomous systems, and wireless embedded systems. Dr. Cenkeramaddi is a member of the editorial boards of various international journals and the technical program committees of several IEEE conferences. He is the Principal Investigator and a Co-Principal Investigator of many research grants from the Norwegian Research Council.