



Federated learning attack surface: taxonomy, cyber defences, challenges, and future directions

Attia Qammar¹ · Jianguo Ding² · Huansheng Ning¹ 

© The Author(s), under exclusive licence to Springer Nature B.V. 2021

Abstract

Federated learning (FL) has received a great deal of research attention in the context of privacy protection restrictions. By jointly training deep learning models, a variety of training tasks can be competently performed with the help of invited participants. However, FL is concerned with a large number of attacks involving privacy and security aspects. This paper shows a federated learning workflow process and how a malicious client can exploit vulnerabilities in the FL system to attack the system. A systematic survey of existing research on the taxonomy of federated learning attack surface and the classification is presented. As with the FL attack surface, attackers compromise security, privacy, gain free incentives and abuse the Confidentiality, Integrity, and Availability (CIA) security triad. In addition, state-of-the-art defensive approaches against FL attacks are elaborated which help to protect and minimize the likelihood of attacks. FL models and tools for privacy attacks are explained, along with their best aspects and drawbacks. Finally, technical challenges and possible research guidelines are discussed as future work to build robust FL systems.

Keywords Federated learning · Security · Privacy · Attack surface · Cyber defence

1 Introduction

Federated learning is a technique that trains millions of AI models by learning from the collaboratively shared model with the help of underlying devices. Currently, FL has become a widely explored topic and has emerged as an alternative to the centralized machine learning (ML) paradigm. Conventional machine learning cannot deal with ubiquitous development due to its centralized infrastructure as it faces challenges such as data computation distribution, limited upload bandwidth, and latency constraints Samarakoon et al. (2020); Li et al. (2018). The purpose of federated learning is to train the global

✉ Huansheng Ning
ninghuansheng@ustb.edu.cn

Jianguo Ding
jianguo.ding@bth.se

¹ School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China

² Department of Computer Science, Blekinge Institute of Technology, Karlskrona, Sweden

model with numerous participants or clients without distributing individual training data. In the FL scenario, the privacy of each participant is guaranteed by localizing the training of the global model as a comparison to centralized ML where the raw data is publicly outsourced to the main server Li et al. (2020a). Around the globe, data breaches have become a major concern, with governments announcing privacy regulations for data protection such as General Data Protection Regulation (GDPR) in Europe Voigt and von dem Bussche (2017), California Privacy Rights Act (CPRA) in the United States CPRA (2020), and Personal Data Protection Act (PDPA) in Singapore Chik (2013). According to a report by Thales Aimee O’Driscoll (2021), 49% of US companies, or almost half, have suffered a data breach in the past. Likewise, Google was charged with \$57 million for data breaches by GDPR and that was a prevalent penalty in March 2020 Satariano (2019). Federated learning is famous as "privacy by design" ensuring that personal data no longer needs to be collected and stored. Primarily, FL assumes that user data is not accessible on a central server because it is isolated, confidential, and exists only in distributed devices. FL promises to provide security and privacy by complying with its customers’ data security laws House (2012). Furthermore, federated learning successfully works in various fields such as health care, finance, smart cities, transportation, visual object detection, and next-word prediction, etc. Jie et al. (2020); Chen et al. (2020); Long et al. (2020); Tan et al. (2020); Zheng et al. (2021); Cheng et al. (2020); Liu et al. (2020a); McMahan et al. (2017). For instance, recently a French rollout was revolutionized with the extension of a biomedical ML model built on FL to preserve the control of patients’ personal data Kuchler (2019). However, FL faces challenges in terms of security and privacy attacks as adversaries have been focusing on data points which are employed in the training of models. Recent studies have elucidated several vulnerabilities in the FL settings related to privacy leakage and security attacks Minghong et al. (2020); Bagdasaryan et al. (2020); Tolpegin et al. (2020); Li et al. (2019); Zhao et al. (2020); Hitaj et al. (2017); Shokri et al. (2017).

In previous years, federated learning research has attracted many scholars and the growth of publications has been exponential. In Fig. 1a, the total publications are portrayed with respect to different years. The keyword “Federated Learning” is searched on the top scientific research databases such as IEEE, ACM, Springer, and ScienceDirect by applying the year-wise filter and it shows approximately 5x proliferation in research papers. In the year 2020, more papers were published compared to previous years. Apart from that, till August-2021, the graph shows the highest publication growth rate which creates an opportunity for hype. Likewise, Fig. 1b demonstrates the total publications with the keyword “Federated Learning” and “Attacks” quest from the aforementioned top scientific research

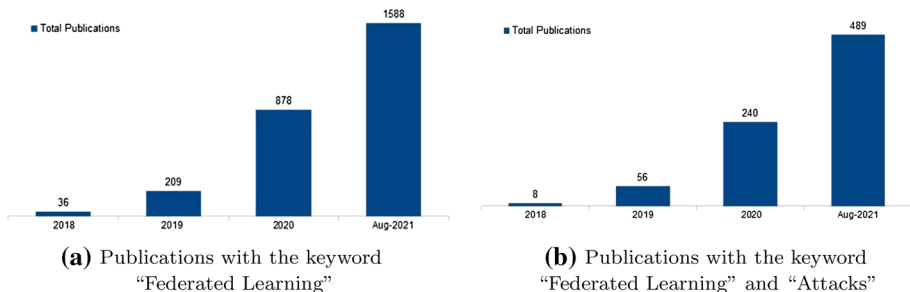


Fig. 1 Total publications in federated learning domain

databases. Correspondingly, as in Fig. 1a, it also shows the highest ratio of publications in the year 2021 till August. However, the publication's growth velocity is very less in previous years. Hence, the statistics attest a great adoption rate of federated learning security research. In earlier surveys Li et al. (2020b); Lyu et al. (2020), researchers highlighted the threat landscape of federated learning, including unique characteristics and security attacks. However, the authors did not discuss the limitations of existing privacy-preserving and security approaches in FL. Furthermore, defensive approaches to federated learning are not addressed in the context of privacy and security attacks. This paper discusses the FL environment with its vulnerable attack surface, FL tools and models against security and privacy attacks. Likewise, this paper attempts to fill the gap by delving into the attacks, defensive approaches, shortcomings, challenges, and future directions.

The remaining of the paper is organized as follows. Section 2 discusses the background knowledge of federated learning with its workflow and types. Section 3 goes over federated learning attack taxonomy with poisoning, inference, free-rider, and CIA security triad attacks. Section 4 introduces the state-of-the-art techniques to secure the federated learning attack surface. Section 5 illustrates the difficulties in federated learning security and privacy, as well as suggestions of some future directions to attain protection. Finally, sect. 6 collects concluding remarks as a conclusion.

2 Background

This section provides the background understanding of how federated learning works in a continuous iterative method, and each iteration improves the performance of the model. Undoubtedly, FL promises to provide demanding protections against attacks, but adversaries are existed to destroy privacy and security as highlighted in this section. Furthermore, data partitioning in FL is explained with its three types and real-world examples.

2.1 Federated learning working process

Federated learning is an innovative branch of artificial intelligence (AI) that opens new horizons in machine learning. FL makes it possible to train a model without the need to transfer or store the training data into a central server. In this way, information can be shared between clients and servers without compromising privacy.

In Fig. 2, the working process of federated learning is illustrated. Typically, the implementation of FL involves three subsequent steps:

1. *Initial model updates:* In the first phase, work typically starts by training a global model, located on a central server, which serves as a baseline model just to start from it. From the beginning of the training procedure, its main parameters are initiated and then the global model is broadcasted as a machine learning model to the clients (the range of clients may belong to "N clients") in the federated learning environment.
2. *Local model updates:* In the second phase, every client at its end locally trains the model with the help of private data. Afterward, all clients send their local model updates to the central FL server without compromising privacy. Periodically, the global model learns from the local models and is improved over time. Furthermore, at any stage, the FL server can add or remove clients during the training process.

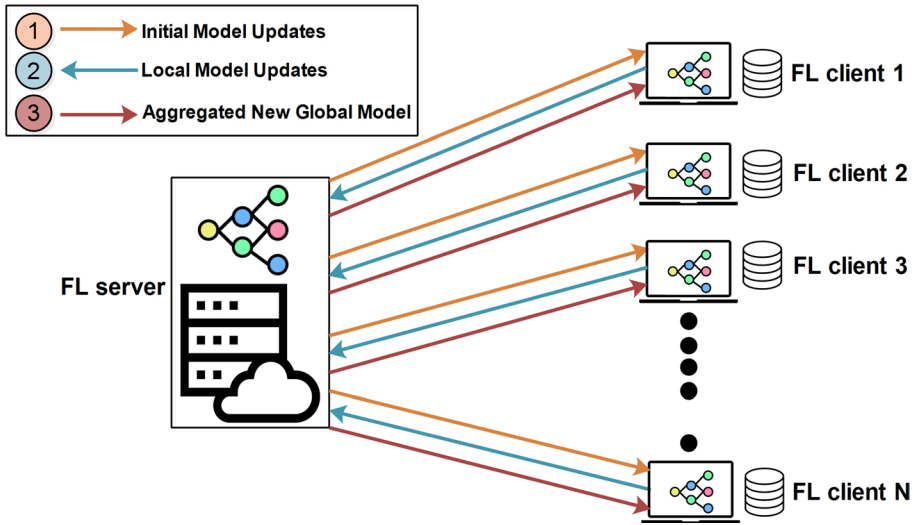


Fig. 2 General working process of federated learning

3. *Aggregated global model:* In the third or final phase, the aggregated fully trained global model is broadcasted to all clients. The aggregated global model is generated from the aggregation of local models trained at the client level. FL enters an iterative state, and with each iteration, the models placed in the central server update themselves and become more personalized.

The total number of clients are supposed to be N and n_i refers to the number of samples from each client. At time t , FL server selects some clients $k.N(0 < k \leq 1)$ randomly from all clients and further asks them to update the global model with their local dataset. According to McMahan et al. (2017), federated learning mathematically expressed in Eq. (1) with the Federated Averaging (FegAvg) algorithm.

$$w_t + 1 = w_t + \eta \cdot \frac{\sum_{i=1}^{kN} n_i \Delta w_{t+1}^i}{\sum_{i=1}^{kN} n_i} \tag{1}$$

In above Eq. (1) w_t shows the previous model updates at time t , whereas $w_t + 1$ is the updated model at time $t + 1$. Moreover, Δw_{t+1}^i presents the updated changes in global parameters by client i . Finally, η is the learning rate of global FL model.

Federated learning is an uninterrupted iterative training process that repeats the second and third phases as a means to maintain updated models of global machine learning for all clients. In this context, the FL server coordinates with the entire federation of participating clients to synchronize the models on a regular basis. Apart from that, in federated learning, the aggregation algorithm plays a crucial role that is used to bind local model updates of the participated clients in the learning round Nilsson et al. (2018). In this context, Google introduced the FegAvg algorithm McMahan et al. (2017), which was created on the basis of the Stochastic Gradient Descent (SGD) algorithm. Similarly, another algorithm named as SMC-Avg Bonawitz et al. (2016) was presented that truly lies on the notion of Secure Multiparty Computation (SMC) algorithm. Last but not least, a modified version

of FedAvg was proposed as FedProx to deal with the dissimilarity in the FL environment Li et al. (2018). In the work of Wang et al. (2020), the authors have proved that Federated Match Averaging (FedMA) algorithm performs well in comparison with FedAvg and FedProx algorithms, just in few rounds of iterations. Moreover, federated learning employs security protocols such as secure aggregation and Differential Privacy (DP) to ensure the clients' information privacy as well as to protect them from attacks Bonawitz et al. (2017); Geyer et al. (2017).

In secure aggregation, encryption is used to prevent the server from accessing specific information. Under this condition, the server has only the right to access the average information collected from thousands or millions of users. On the other hand, DP adds some random data noise to the distinct summaries to obfuscate the results. However, FL is vulnerable to multiple attacks discussed in sect. 3, here is a general attack scenario demonstrated in Fig. 3. According to Fig. 3, an adversary has a local malicious model and sends malicious learning updates to the global model and vice versa. The adversary can be an evil client or server who injects malicious nodes into the model. In this case, the adversary plays the role of an insider adversary and each client downloads the global model that is transferred from the central malicious server. After sufficient training and multiple iterations between the FL clients and FL server typically named as a communication round, the model is updated on each connected device. In recent studies Kairouz et al. (2019); Bhowmick et al. (2018); Ma et al. (2020), there is evidence that the communication of model updates exposes sensitive data of associated clients and promotes security issues.

2.2 Types of federated learning

The distributional properties of the data are classified into three types, namely Vertical Federated Learning (VFL), Horizontal Federated Learning (HFL), and Federated Transfer

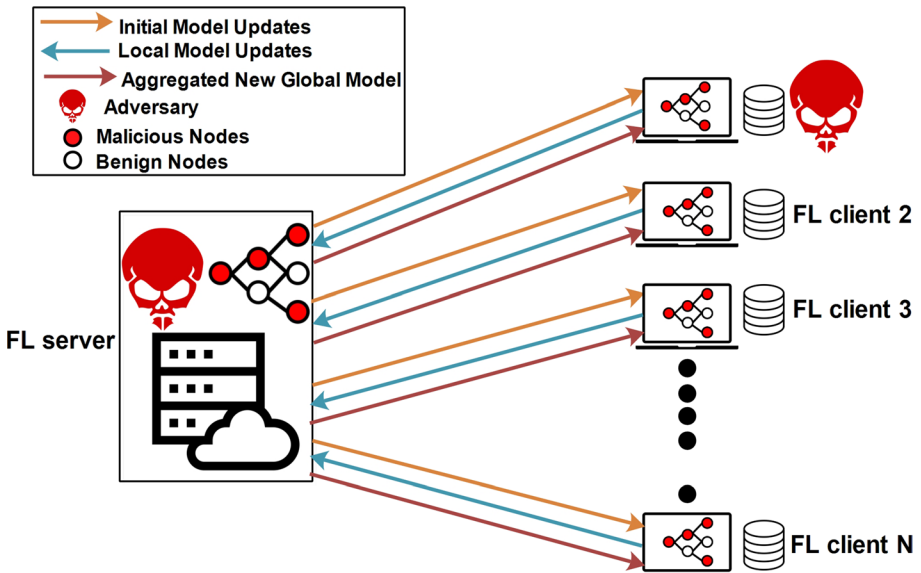


Fig. 3 Vulnerable federated learning system

Learning (FTL) Yang et al. (2019). Generally, the preparation for launching the federated learning environment starts with a categorization of data with respect to features as illustrated in Table 1. As different parties are involved in the federated learning environment and they can choose desired data partition type settings. Suppose, matrix D_i and D_j shows that the data belongs to each data holder i and j , respectively. In a matrix, each row and column represents the samples and features, respectively. Furthermore, X is used to indicate the feature space, Y shows the label ID and I is for sample space.

Vertical or feature-based or heterogeneous federated learning contains datasets with diverse kinds of features but have identical sample space. For instance, two different organizations such as banks and an e-commerce have different features in terms of expenditure behavior, credit card rating, browsing, and purchasing history of users, respectively. In this case, if both companies need predicted ML models on the basis of user information for product purchases. The vertical FL aggregates these various kinds of features, calculates the training loss and gradients by considering the privacy-preserving mechanism to construct the model. In previous studies Hardy et al. (2017); Nock et al. (2018), a privacy-preserving logistic regression model was introduced based on the VFL structure. Furthermore, in the work of Cheng et al. (2019) authors designed the SecureBoost in the environment of a vertically divided dataset. Nevertheless, discussed methods are applied in ML logistic regression model, so VFL needs much progress to apply in other ML approaches.

Besides, the horizontal or sample-based, or homogeneous FL, have identical features but differ in terms of sample or data. One of the famous example of HFL, that it is applied by Google to manage the updates in Android phones. In the work of McMahan et al. (2016), the authors build the secure client-server framework based on HFL settings that allow to build a global FL with the collaboration of client devices and server sites. Furthermore, the approach named with deep gradient compression Lin et al. (2017) was presented to decrease the communication cost in large-scale distributed training environments. Both HFL and VFL have used similar security protocols such as homomorphic encryption (HE) and secure multiparty computation (SMC) McMahan et al. (2017); Araki et al. (2016). In HFL, the central server frequently belongs to honest but curious and exposes the gradients of every client that causes information leakage. For this purpose, SMC is applied Bonawitz et al. (2017) and results proved that privacy is considerably enhanced by encrypting the model gradients. Likewise, Wei et al. (2020) suggested the VFL approach and applied homomorphic encryption for risk management. In accumulation to VFL and HFL, one more type of federated learning is FTL which has two different datasets, samples and features. In the work of the authors Liu et al. (2020b), FTL is applied to achieve a flexible and effective model without compromising the privacy of clients. Moreover, the real-time example of FTL is not much different from VFL as explained in Chen et al. (2020), a global personalized model is provided on wearable IoT devices of a specific user. In literature Nadiger et al. (2019); Lim et al. (2020), some studies belong to Federated Reinforcement Learning (FedRL), that implements the federated transfer learning from the trained security model.

Apart from, different types of data availability are also considered in FL such as cross-silo and cross-device Kairouz et al. (2019). The cross-silo FL is applied where clients are on a small scale, whereas the cross-device FL approach works with a big number of clients that have common interests from the global model. For instance, a great number of clients including mobile applications and IoT are best fits in cross-device FL. However, security in IoT devices are also at risk and increasing day by day as defined in Radanliev et al. (2021). The authors Zhang et al. (2020) developed the framework with Federated AI Technology Enabler (FATE) FATE (2021) to explain the cross-silo with

Table 1 Federated Learning Types related to Data Partition

Types	Feature space	Sample space	Definition	Security protocols	Characteristics
VFL	Different	Same	$X_i \neq X_j, Y_i \neq Y_j, I_i = I_j \forall D_i, D_j, i \neq j,$	Semi-honest Third Party (STP), Secure Multiparty Computation (SMC), and Homomorphic	Encryption and Privacy
HFL	Same	Different	$X_i = X_j, Y_i = Y_j, I_i \neq I_j \forall D_i, D_j, i \neq j,$	STP, SMC, and Homomorphic	Security and Independence
FTL	Different	Different	$X_i \neq X_j, Y_i \neq Y_j, I_i \neq I_j \forall D_i, D_j, i \neq j,$	Private Set Intersection (PSI)	Avoids from accuracy loss and Encryption

homomorphic encryption. Furthermore, cross-silo FL applications are implemented in countless areas including electronic health records FeatureCloud (2021), smart manufacturing Musketeer (2020), and finance risk prediction FedAI (2020), etc.

3 Taxonomy of attack surfaces for federated learning

The attack surface of federated learning is an overview of vulnerabilities in FL settings that are exploited by adversaries. In general, attacks are launched by compromising the security of the central servers or local devices or participants in the federated learning workflow. Moreover, once an attacker gains access to the federated learning ecosystem, the training parameters, aggregated model updates, and learning outcomes can be altered. FL viable attacks can be described in terms of their source and nature, etc. In Fig. 4, a taxonomy of FL attacks is visualized and their different types are explained in the below subsections.

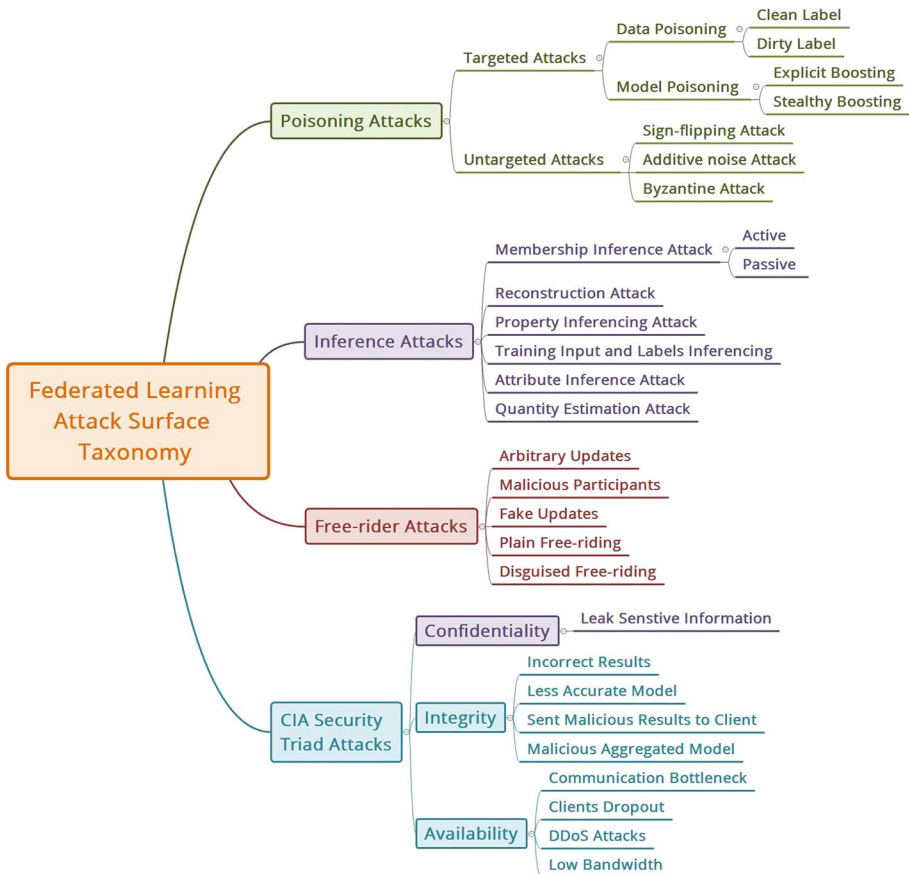


Fig. 4 Taxonomy of federated learning attacks

3.1 Poisoning attacks

The objective of a poisoning attack is to converge a benign model into a wrong model and it has a high probability of occurrence in the FL environment Kairouz et al. (2019); Bhagoji et al. (2019). Each participant in FL can equally access the training data, in this case, malicious data can be added by an adversary or malicious clients to the global FL model as visualized in Fig. 5. Each participant sends the updates to the central server then gets a fully trained FL model. So, the poisoned updates affect the training datasets or the local model, thus indirectly poisoning the global model and reducing the model accuracy. In federated learning settings, poisoning attacks are divided into two categories such as targeted and untargeted attacks Huang et al. (2011).

The targeted attacks Minghong et al. (2020); Baruch et al. (2019); El Mhamdi et al. (2018) have an aim to poison the specific labelled data whereas untargeted attacks Bagdasaryan et al. (2020); Bhagoji et al. (2019) can randomly affect the model accuracy by disregarding the specific test input. In general, targeted attacks are more difficult to launch as an adversary must achieve a particular goal. Furthermore, targeted attacks are classified as data and model poisoning attacks.

Data poisoning attacks in the FL settings ensure to manipulate the dataset at the client end rather than directly poisoning the central model. Basically, adversaries make alerts to use the client datasets in the process of model training through the clean label and dirty label attacks. In the work of Fung et al. (2018), authors demonstrate that the FL environment is vulnerable to label flipping attacks in the context of dirty labels. Further, the authors show the attack success rate and then introduce the defensive mechanism. Consequently, the proposed attack mitigation method efficiently counters the data poisoning attack. Likewise, another researcher Tolpegin et al. (2020), investigated the data poisoning attacks against a federated learning environment and showed their negative impact on the global model in terms of accuracy and recall. The experiment is conducted with a small plus large number of malicious clients to achieve the targeted poisoning ratio in the later iteration rounds. Finally, an approach is suggested to identify

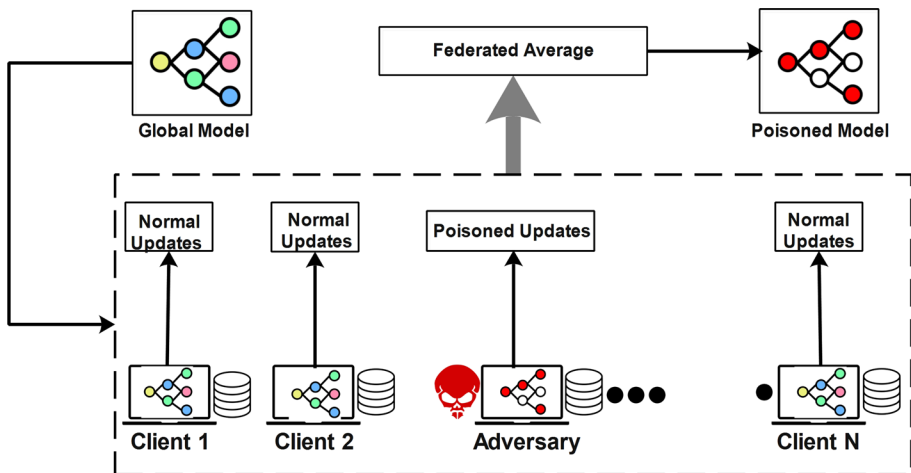


Fig. 5 Poisoning attack in a federated learning environment

the malicious clients in FL to evade poisoning attacks. As a result, the FL aggregator successfully isolates malicious and benign participants.

Besides, in model poisoning attacks an adversary directly manipulates the model rather than inserting malicious data into the training datasets. In the literature Bhagoji et al. (2019), researchers explored that model poisoning attacks are more harmful in comparison with data poisoning attacks. Moreover, model poisoning attacks through explicit and stealth boosting attacks are explained. The explicit boosting attack works based on trivial attacks with local malicious updates. By keeping the selected target in mind, adversaries train the updates. In a stealth-boosting attack, the attacker adds more loss terms to maintain some acceptable accuracy of the model and make sure that the malicious updates should be close to the benign updates. Zhou et al. (2021), presented the deep model poisoning attack by injecting malicious neurons into the global model and proved that it is more robust and persistent in FL scenarios. Accordingly, experiments have depicted that a novel model poisoning attack achieved the highest attack success rate, and is related to stealth which can easily evade prevailing defense methods Blanchard et al. (2017); Sun et al. (2019).

The untargeted attacks such as a sign-flipping attack do not alter the magnitude of model weights. For this purpose, malicious participants flip the sign of local model updates and fail the defensive mechanisms. In addition to the type of untargeted attack, the second type is additive noise attack, in which adversaries add the Gaussian noise to local model updates that ultimately degrading the model performance. However, adding some noise is used to safeguard data privacy. Note that, adding excessive noise can compromise the model accuracy Li et al. (2019). Finally, yet importantly, in the types of untargeted attacks, there is also a byzantine attack in which the client deviates from its normal behavior to malicious or abnormal behavior in the FL environment Li et al. (2019). Two reasons are the main causes of abnormal client behavior 1) an adversary can disguise itself and act like a normal client and 2) software or hardware defects may occur on the client side. So, it is necessary to detect the byzantine attacks to prevent allocating the rewards or incentives to the malicious concealed clients Kang et al. (2019). In Table 2, existing poisoning attacks studies with their defensive strategies and future work are listed.

3.2 Inference attacks

The inference attacks aim to exchange gradients that cause information leakage in federated learning frameworks Hitaj et al. (2017); Phong et al. (2018); Zhu and Han (2020). Inference attacks have a great severity of privacy leakage and can be launched through malicious centralized servers or clients in the FL workflow. In Fig. 6, inference attack is illustrated as an adversary inferring the information through gradient leakage. At each round n , an adversary downloads the current model, calculates gradient difference, and sends its local updates to the central server to update the global model. Furthermore, the adversary saves the model parameters X_n and computes the difference between successive model parameters as shown in Eq. (2), which is equal to aggregated updates of k participants.

$$\Delta X_n = X_n - X_{n-1} = \sum_k \Delta X_n^k \quad (2)$$

Accordingly, the aggregate updates of all participants except the adversary are shown in Eq. (3), where *adv* stands for adversary. Conclusively, aggregated or updated gradients cause an exploitation of data privacy.

Table 2 Summary of Existing Studies Related to Poisoning Attacks

Objective	Attack type	Defensive strategy	Datasets	Measured outcome	Results	Future work	Reference
The detection of poisoning attacks and elimination of the poisoning updates from adversaries.	Poisoning attacks	Poisoning defense generative adversarial network (PDGAN)	MNIST Lecun et al. (1998) and Fashion-MNIST Xiao et al. (2017)	Mean Overall Accuracy, and Mean Poisoning Accuracy	Before mitigating the poisoning attack the overall accuracy 80.33%. After deploying the PDGAN, overall accuracy is increased by 91.85%.	Investigate the poisoning defense with device, class, and client level DP for federated learning.	Zhao et al. (2020)
The vulnerability to poisoning attack, label flipping attacks with complex deep neural network models.	Data Poisoning	A defensive mechanism to separate malicious participants from benign	CIFAR-10 Krizhevsky et al. (2009) and Fashion-MNIST Xiao et al. (2017)	Model Accuracy, and Source Recall	Attack works effectively with small number of malicious users. Targeted attacks have a great negative impact on the subset of classes.	Effectiveness of the defensive mechanism against potential attacks such as backdoor, model poisoning, and untargeted poisoning attacks.	Tolpegin et al. (2020)
Launched the model poisoning attack in federated learning settings.	model poisoning attack		Fashion-MNIST Xiao et al. (2017)	Classification accuracy	Results proved that FL in its basic form is very vulnerable to model poisoning attack.	Robust protection mechanisms are need in future against Stealthy model poisoning.	Bhagoji et al. (2019)

Table 2 (continued)

Objective	Attack type	Defensive strategy	Datasets	Measured outcome	Results	Future work	Reference
Proposed a new optimized model poisoning attack with persistence and stealth features.	model poisoning attack		MNIST Lecun et al. (1998) and CIFAR-10 Krizhevsky et al. (2009)	Accuracy with a single shot and multi-shot attack	An attack is enough stealthy and escape from two existing defense methods.	A malicious client used the Hessian matrix to launch the attack, which takes a lot of time, during this time client can be removed from the system.	Zhou et al. (2021)
AttestedFL framework designed to identify the local model poisoning attacks through the behavior of the model and removes the unreliable nodes.	Local model poisoning attacks	attestedFL-1, attestedFL-2, and attestedFL-3		Accuracy	AttestedFL is evaluated under various settings such as continuous attacks and attackers colluding. AttestedFL increased by the accuracy of model 12% to 58%.	Optimization, computation overhead will be measured. Explore the relationship between single-shot attacks, time to generate the attack against attestedFL.	Mallah et al. (2021)

Table 2 (continued)

Objective	Attack type	Defensive strategy	Datasets	Measured outcome	Results	Future work	Reference
Backdoor attack is analyzed, defense framework “FLGuard and manipulation of the global model that causes to leak information.	Backdoor attack	FLGuard	Reddit Google BigQuery (2017), CIFAR-10 Krizhevsky et al. (2009), MNIST Lecun et al. (1998) and TinyImageNet He et al. (2016)	Backdoor Accuracy, Main Task Accuracy, Model Accuracy	FLGuard efficiently mitigates the backdoor attacks and block inference attacks to ensure the privacy of user-level training data.		Nguyen et al. (2021)
Launched the distributed backdoor attack (DBA) in federated learning settings.	Distributed backdoor attacks (DBA)		LOAN Kanwedy (2019), MNIST Lecun et al. (1998), CIFAR and Tiny-imagenet He et al. (2016)	Attack Success Rate (ASR) and Model Accuracy	DBA is a novel and powerful attack against federated learning. Threat assessment tools are suggested for evaluating the adversarial robustness of FL.	To design a protection framework against distributed backdoor attacks in federated learning settings.	Xie et al. (2019)

Table 2 (continued)

Objective	Attack type	Defensive strategy	Datasets	Measured outcome	Results	Future work	Reference
Introduced a backdoor attack with defensive method.	Backdoor attack	Norm thresholding of updates and DP	EMNIST Cohen et al. (2017)	Accuracy of attack and defense	Attack success rate depends on a large number of adversaries and adding noise helps to mitigate the attack.	Advance cyber defences are required without losing the model accuracy.	Sun et al. (2019)
Timely detection of abnormal behavior to minimize the adversarial impact. A Pre-trained autoencoder model is deployed to detect malicious model updates from clients.	Abnormal behavior	Thresholding and Credit Score schemes for anomaly detection	FEMNIST (Federated Extended MNIST) Caldas et al. (2018)	Accuracy of defense	By adopting the autoencoder method, result Shows effective performance in detection of malicious model updates from FL system.		Li et al. (2019)

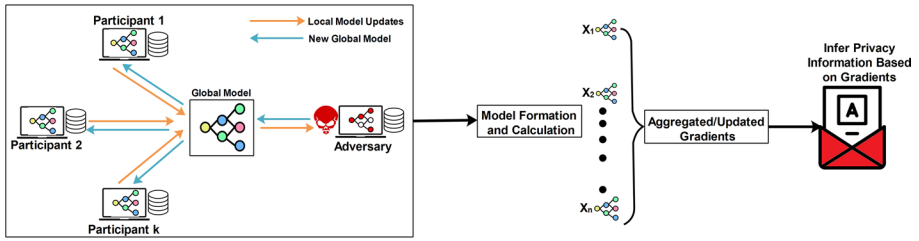


Fig. 6 Inference attacks in the federated learning environment

$$Y = \Delta X_n - \Delta X_n^{adv} \tag{3}$$

In Table 3, existing studies related to inference attacks such as membership inference attack, reconstruction attack, and feature inference attack, etc. are discussed with their type, objective, datasets, and future work.

3.2.1 Membership inference attack

The membership inference attack (MIA) has an objective to infer the hidden features of training data in contradiction of the learning model. MIA has an ability to decide whether the presented samples belong to the training data or not Shokri et al. (2017); Chen et al. (2020). For instance, assume that W_n denotes that words are included in the updates ΔX_n . During training, the adversary collects the vocabulary set such as $[V_1, \dots, V_n]$. The text record r , with words W_r , an adversary can test if $W_r \subseteq V_i$, for some i in the vocabulary set. If r is the targeted dataset, in this case, W_r will contain at least one vocabulary from the set. Hence, the adversary will use this to decide either r is a member or not.

In literature, firstly Melis et al. (2019) presented the membership inference attack for the FL process by using a batch classifier to infer properties during the learning phase. Likewise, Nasr et al. (2019) proposed a novel approach to privacy-preserving FL systems using a white-box membership inference attack. Moreover, researchers explored the vulnerability of the SGD algorithm that is lying in passive and active inference attacks. In the work of Hitaj et al. (2017), the authors exploit the shared model through mislabeled samples into the learning phase by implementing the Deep Convolutional Generative Adversarial Network (DCGAN). The presented attack shows its effectiveness in obfuscated parameters and proves that DP is an unsuccessful approach under the proposed attack.

3.2.2 Reconstruction attack

In a reconstruction attack, the adversary has an aim to reconstruct the samples from learning model parameters Fredrikson et al. (2015). The authors in Yang et al. (2019) proposed the reconstruction attack in the settings of black box and worked on reversing the original model throughout the training process. In the study of He et al. (2019), researchers presented the model inversion attack under the white box and black box structure to recover inference data. In the case of inference task distribution, any malicious client can perfectly recover random inputs, even if other clients' data or calculations are not within its access rights.

Table 3 Summary of Existing Studies Related to Inference Attacks

Objective	Attack type	Model	Datasets	Measured outcome	Results	Future work	References
Formulate the feature inference attack in a vertical FL environment.	Feature inference attack	Generative Regression network (GRN)	Bank marketing (Moro et al. (2014)), Credit card (Yeh and Lien (2009)), Drive diagnosis (Dua and Graff (2017)), and News popularity (Fernandes et al. (2015))	Mean square error (MSE) and Correct branching rate (CBR)	Extensive experiments on both kinds of datasets such as real-world and synthetic are conducted. The attack shows effectiveness in terms of MSE and CBR.	It is required to design the defensive framework to mitigate the targeted attacks.	Luo et al. (2020)
The protection strategy against GAN-based feature inference attacks like Anti-GAN.	GAN base feature inference attacks	Wasserstein WGAN	MNIST (Lecun et al. (1998)), Fashion-MNIST (Xiao et al. (2017)) and CIFAR-10 (Krizhevsky et al. (2009))	Real-Image Accuracy from Fake Images (RIAFI) and mix ratio (MR)	Distortion of the original distribution of training data leads to prevent attackers from generating different images. Anti-GAN causes slightly loss to the testing accuracy of the global classifier.	Test Anti-GAN approach for various datasets and evaluate its performance against protection techniques such as sharing fewer gradients and dropout.	Luo and Zhu (2020)

Table 3 (continued)

Objective	Attack type	Model	Datasets	Measured outcome	Results	Future work	References
Examine the membership inference attack in a sequential FL environment.	Membership Inference Attack	Neural Network	Purchases dataset kaggle (2013)	Accuracy, Precision, Recall, and F1	In training, the model's data membership inference attack has better accuracy. Attack accuracy is increased with the higher number of epochs.	To train the obtained model effectively as well as sufficiently secure it from membership attack.	Anastasia Pustozrova and Rudolf Mayer (2020)
Launch the GAN-based attack and introduce the notion of deception in the case of obfuscated parameters.	GAN attack (Inferring Class Representatives)	DCGAN	MNIST Lecun et al. (1998) and AT&T dataset of faces Samaria and Harter (1994)	GAN attack with enabled/disabled DP	The adversary has good performance in reconstruction of images. Adding noise to the parameters is not protected from GAN attack.	Possible differential privacy or homomorphic encryption might be causing to evade the attack.	Hitaj et al. (2017)
Investigate the inference attack with active and passive attacks, leakage in unintended features such as properties or subset of properties.	Membership inference and property inference		FaceScrub Moro et al. (2014), FourSquare Yang et al. (2015), Yelp-health and Yelp-auth Yelp (2020)	AUC scores and Precision	Leakage of unintended features in FL to dominant inference attacks. Membership inference attack shows a 91% precision rate.	The robust protection approach against inference attacks and more measured outcomes should use to show the effectiveness of the attack.	Melis et al. (2019)

Table 3 (continued)

Objective	Attack type	Model	Datasets	Measured outcome	Results	Future work	References
The mGAN-AI attack for federated learning settings to reconstruct private data of a specific client.	mGAN-AI (reconstruction attack)	GAN	MNIST Lecun et al. (1998) and AT&T Samaria and Harter (1994)	Inception Score, Accuracy	mGAN-AI can reconstruct samples that look like the victim's training samples. Outperforming than the attacking algorithms such as MI and GAN-based attack.	Attack generated on the subset of training samples is recommended with its protection scheme.	Wang et al. (2019)
A user-level membership inference attack is initiated and encourage the academia to work on its prevention.	membership inference attack	Generative Adversarial Networks (GANs)	MNIST Lecun et al. (1998) and CIFAR-10 Krizhevsky et al. (2009)	True Positive (TP) and False Negative (FN)	Attack success shows highest accuracy with MNIST and CIFAR-10 datasets such as 99.45% and 93.71% respectively.	Build approaches that make it difficult for the adversary to distinguish the ownership of data labels.	Chen et al. (2020)

Table 3 (continued)

Objective	Attack type	Model	Datasets	Measured outcome	Results	Future work	References
Presented the novel white-box membership inference attack and designed algorithm to exploit the privacy vulnerabilities of SGD.	membership inference attack		CIFAR100 Krizhevsky et al. (2009)	Attack accuracy, True/False positive and Prediction uncertainty	Adversarial clients in the FL setting successfully launched active membership inference attacks against other clients. Furthermore, the presented attack obtained an accuracy of 74.3%.	Theoretical bounds investigation of privacy leakage in the white-box setting remains as future topic.	Nasr et al. (2019)
Designed three types of passive attacks which aim to infer the labels, steal the quantity information of a single training round or whole training round with data of all participants.	Class sniffing, Quantity Inference, and Whole determination		MNIST Lecun et al. (1998), CIFAR-10 Krizhevsky et al. (2009) Fer2013 Goodfellow et al. (2013) and HAM10000 Tschandl et al. (2018)	True Positive (TP) and False Negative (FN)	Attacks successfully worked in FL setting with secure aggregation protocols and differential privacy.		Wang et al. (2019)

3.2.3 Property inferencing attack

The property inference attack is an attempt to extract properties from training data Truex et al. (2019). The adversary tries to recognize data patterns and then expose them, maybe the owner of the model producer does not want to reveal them. Melis et al. (2019), presented that how an attacker infers the properties of learning data with the help of some auxiliary data. Furthermore, in the study of Zhang et al. (2020), an attacker infers the sensitive attributes of other parties' datasets in collaborative learning settings.

3.2.4 Training input and label inferencing

Recently the authors Zhu and Han (2020) explained the deep leakage gradient (DLG) that exploits the training input and labels in a very small number of iterations. Similarly, another attack improved deep leakage through gradients (iDLG) Zhao et al. (2020) are presented which mines the labels from mutual gradients by manipulating the connection between labels of associated gradients. Moreover, in the study of Wang et al. (2019) inferencing label were shown in a single iteration round with the aim of stealing the quantity information.

3.2.5 Attribute inference attack

In an attribute inference attack, the adversary has an intention to de-anonymize the record holder and disclose sensitive information from clients' private data Kairouz et al. (2019). The sensitive information may include gender or location etc. In the work of Luo et al. (2020), researchers explained privacy leakage in vertical federated learning at the model prediction stage. For evaluation, the proposed attack is implemented on different datasets and the highest attack success rate were demonstrated for the required protection approach in VFL settings.

3.2.6 Quantity estimation attack

The attackers obtained information from the trained model with inadequate power and also get a quantity of each label that leads to a quantity estimation attack. Wang et al. (2019), discussed the inference attack in the context of composition proportion that causes to leaks of different labels in the federated learning process. The adversary has an aim to slink a quantity information from the desired participants after that hits the quantity proportion of all clients at various training phases. Furthermore, the authors launched the attack in passive settings and were unable to detect it by intrusion detection methods.

3.3 Free-rider attacks

In general terms, a free-rider is any individual who gains advantage from resources, services of collective nature, and public goods but does not pay at all. In literature, free-rider attacks are broadly discussed in the context of a peer-to-peer environment Tseng and Chen (2011); Feldman et al. (2006). However, in this paper, the free-rider attacks are explored in FL settings. In federated learning systems, each client has to contribute and in this regard, clients get incentives or rewards. In Fig. 7, it demonstrates that some clients act as free-riders, send fake updates to the global model, and receive an incentive for free. The free-rider

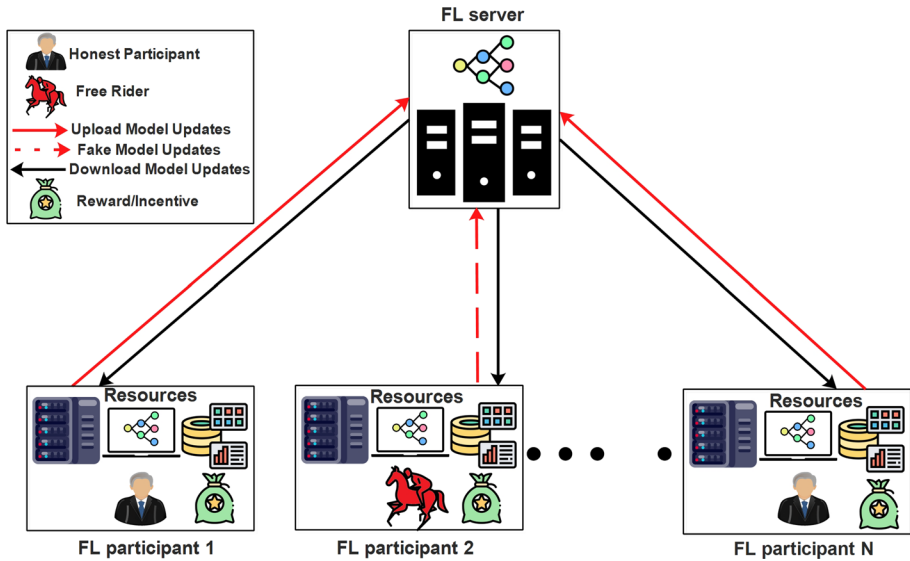


Fig. 7 Free-riding attacks in the federated learning environment

attacks got rewards in two ways 1) save local resources by not sharing them and 2) not willing to share local data in the training process of global model but consume the computing power.

In the study of Lin et al. (2019), researchers investigated the free-rider attacks and suggested a robust defensive approach as Standard deviation-Deep Autoencoding Gaussian Mixture Model (STD-DAGMM), which has an ability to identify free-riders and thwarts them from receiving aggregated model updates along with pecuniary rewards. Free-riders use the arbitrary generated values instead of gradient updates and the authors show that free-rider attack can easily bypass the autoencoder approach. A new type of attack is launched named as delta weights attack to generate the gradient updates and that can escape from the Deep Autoencoding Gaussian Mixture Model (DAGMM) method Zong et al. (2018). However, STD-DAGMM potentially mitigates the attack and detects them with certainty. Moreover, Fraboni et al. (2021) developed a theoretical framework for free-rider attacks in federated learning workflow based on the model averaging. The authors proved that returning the global model in each iteration leads to a successful plain free-riding attack. In addition, disguised free-riding attacks are explored, relying on stochastic perturbations. Besides, in future work, it is necessary to investigate the optimal disguised free-riding attacks and cyber defences.

3.4 CIA security triad attacks

CIA triad ensures the confidentiality, integrity, and availability to deploy as an ideal security model. Federated learning-based ecosystems face attacks in CIA security attributes. In Fig. 8, attacks are demonstrated with the CIA security triad in the FL environment.

Attackers compromise the confidentiality of data and leak sensitive information Bommasani et al. (2021). The attack on confidentiality in the FL environment relates to the

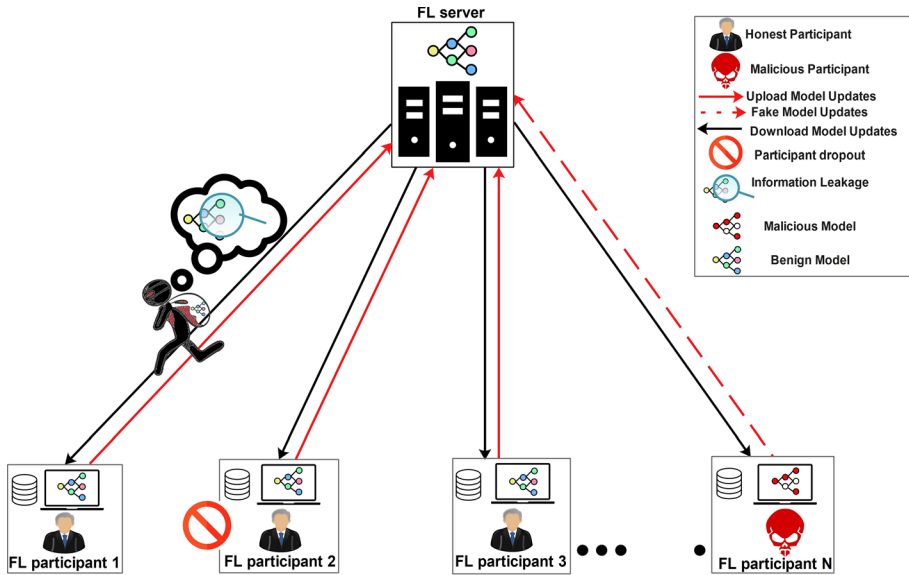


Fig. 8 CIA security triad attacks in the federated learning environment

inference attacks. Generally, FL information leakage is done through exchanging the gradients during the training process of the model. The gradients are created based on the participants' confidential data. Furthermore, model updates and inspection of data cause to expose multiple features of the data Zhao et al. (2020). Apart from that, an attack on integrity leads to dishonest action such as sending incorrect updates and modifying the model by adding malicious nodes. Both FL server and participants can manipulate the model results and send less accurate results to others. Similarly, availability is affected through communication bottlenecks and the highest dropout ratio of clients during the model training process. With constant up-gradation in hardware capabilities, computation overhead has been reduced. Authors Radanliev and De Roure (2021), reviewed the AI algorithms which are operated on low memory devices and proposed an approach by integrating the IoT devices and datasets to maximize the efficiency. However, communication overhead in FL, can be a bottleneck in the training process of large deep learning models Smith et al. (2017); Stich (2018). For instance, the ResNet-50 model comprises 23.5 million parameters that faces an overhead of 94MB and the weight is encoded with 4 bytes. A standard FL framework allows the clients to submit model updates to the FL server in each training round. So, hundreds of communications rounds are required to fully train the model in FL that poses a hurdle. In this context, compression is implemented to overcome the communication cost but it causes a decrease in model accuracy. Moreover, participants unexpectedly dropout from the FL system during the model training and resulting in further updates unavailable. The dropout of clients occurs due to network issues or the client explicitly leaves the training process. Consequently, the dropout of clients has produced fruitless results and causes fairness issues during model synchronization in FL Nishio and Yonetani (2019); Huang et al. (2020). Similarly, a compromised FL server by a Distributed Denial of Service (DDoS) attack makes updates inaccessible for all clients. Besides, CIA triad security attributes, the authentication and authorization of clients are also compromised. Attackers create multiple fake identities such as sybil attacks making it hard to

detect genuine participants, who control various malicious participants at the same time Fung et al. (2020). Additionally, sybil clones mount the poisoning attack in FL settings and the attack effectiveness is increased by using sybils. In presence of only two malicious sybils and ten honest participants, attack success shows a 96.2% ratio Fung et al. (2018).

4 State-of-the-art: securing federated learning attack surface

In this section, we elucidate the various defensive strategies proposed in the literature for securing the FL environment against different attacks. Defensive approaches are found in the existing work are grouped by their underlying protection types.

Figure 9, demonstrates the state-of-the-art cyber defense techniques in the perspective of security and privacy in the FL environment. Furthermore, federated learning tools and models are discussed, which are mostly used in current studies. Despite that, very limited research has been found on frameworks for FL attack protection. The security and privacy approaches are defined in order to overcome the poisoning and inference attacks in FL, respectively. In the case of CIA security triad approaches see subsection 4.5 the privacy-preserving, verification, and integrity-related terms are considered from literature to provide a safe FL environment. Finally, FL built-in tools such as TensorFlow Federated (TFF) Developers (2021), PySyft OpenMined (2021), and PaddleFL Wang (2019), etc. are used to provide security measures against FL attacks. However, these tools have some shortcomings such as poor documentation, partial support of other libraries, and are unable to provide alternative privacy solutions.

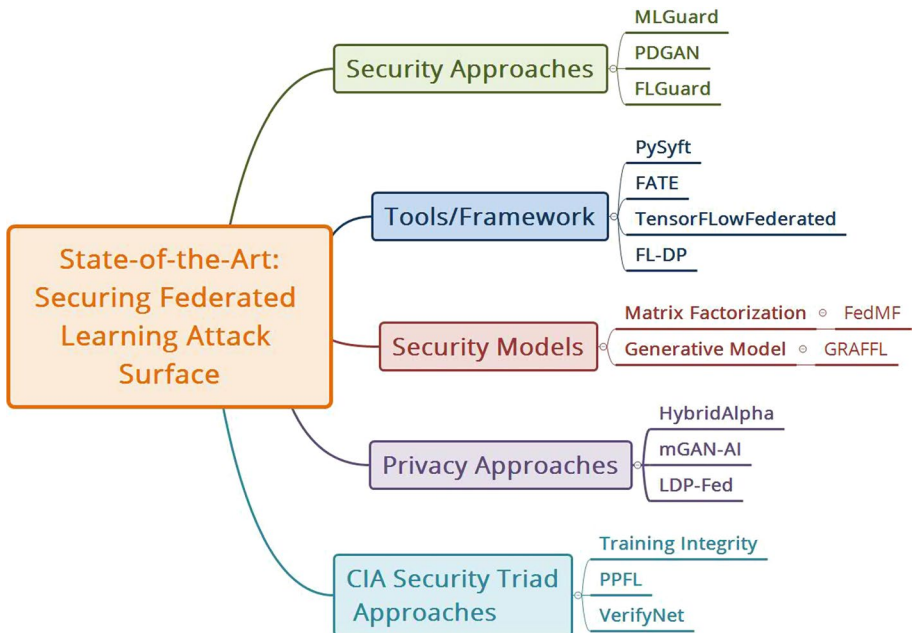


Fig. 9 State-of-the-art: securing federated learning attack surface

4.1 Defensive approaches against security attacks

Insufficient approaches have been improved to secure the federated learning system over poisoning attacks. In Fig. 10, attacks with defense techniques are portrayed. Furthermore, in Table 4, an overview of defensive approaches are illustrated with the root causes of attacks and their impact. Generally, poisoning attacks fall under the category of security attacks in the federated learning ecosystem.

A Poisoning Defense Generative Adversarial Network (PDGAN) Zhao et al. (2020), a unique approach, is presented to defend poisoning attacks. PDGAN successfully reconstructs the training samples from model updates and investigates the accuracy of each client model by employing the generated data. Consequently, whichever participant's accuracy falls below a predefined threshold is marked as a malicious client and is removed from the training process. Likewise, another approach FLGuard Nguyen et al. (2021) has been proposed to protect against backdoor attacks. FLGuard maintains the accuracy of the aggregated model and its effectiveness has been extensively calculated on various datasets including image classification, IoT intrusion detection, and word prediction. In this case, experimental results indicate that FLGuard has no effect on the accuracy of the aggregated model and provides full protection against backdoor attacks. For the detection of data poisoning attacks, alternative defensive strategies are explored in Tolpegin et al. (2020); Khazbak et al. (2020) and the authors aim to persevere the clients' privacy and mitigate the success ratio of attacks. Furthermore, another approach named attestedFL Mallah et al. (2021) is introduced to defend against model poisoning attacks, consisting of three aspects such as 1) attestedFL-1, which ensures the convergence of local model with the global model, 2) attestedFL-2, which observes the angular distance of local model updates during node training, and 3) attestedFL-3, which eliminates the local model updates where performance is not augmented. For the detection of untargeted poisoning attacks an auto-encoder-based anomaly detection and an aggregation algorithm method are explained Li et al. (2019); Fu et al. (2019). Other approaches So et al. (2020); Xu and Lyu (2020), such as Byzantine-Resilient Secure Aggregation (BREAS) and Robust and Fair Federated Learning (RFFL) are initiated as resilience approaches for byzantine users whose objective is to exploit models and datasets. For both approaches, the results indicate the acceptable accuracy for byzantine users.

4.2 Federated learning tools against privacy attacks

In literature, there are few federated learning tools that work with differential privacy techniques to secure data. The famous open-source FL tools include the tensor flow federated (TFF) Developers (2021), federated AI technology enabler (FATE) FATE (2021), PaddleFL Wang (2019), PySyft OpenMined (2021), and federated learning and differential privacy (FL and DP) Sherpa.ai (2021) are reviewed in Table 5. The tools are elaborated with their supported operating systems (OS), features, and shortcoming. PaddleFL provides the industrial framework with a high-level interface and differential privacy mechanism. With differential privacy, some noise is inserted into model updates to make the attack unsuccessful. Initially, DP provides a defence against privacy attacks but now it also protects against data poisoning attacks Ma et al. (2019). Injecting the noise into model parameters prevents information leakage from gradients. However, the insertion of noise in model parameters causes to pollute the accuracy of the model. Apart from the DP mechanism,

Table 4 Defensive Approaches against Poisoning Attacks with Root Attack and its Impact

Defensive approaches	Attacks	Root cause of attack	Impact of attack	References
MLGuard	Data poisoning	Label flipping	High	Khazbak et al. (2020)
Attested-FL and PDGAN	Model Poisoning	Explicit and stealth boosting	High	Mallah et al. (2021), Zhao et al. (2020)
Autoencoder-Based Anomaly Detection	Sign flipping	Flipping the labels	Medium	Li et al. (2019), Fu et al. (2019)
Aggregation algorithm with residual reweighting	Additive noise	Adding excessive noise (Gaussian noise)	High	Li et al. (2019)
Autoencoder-Based Anomaly Detection	Byzantine	Abnormal client behavior	Medium	So et al. (2020), Xu and Lyu (2020)
Byzantine-resilient secure aggregation (BREAS)	Backdoor attacks	Inject malicious task	High	Nguyen et al. (2021), Fu et al. (2019)
Robust and Fair Federated Learning System (RFFL)				
FLGuard				
Aggregation algorithm with residual reweighting				

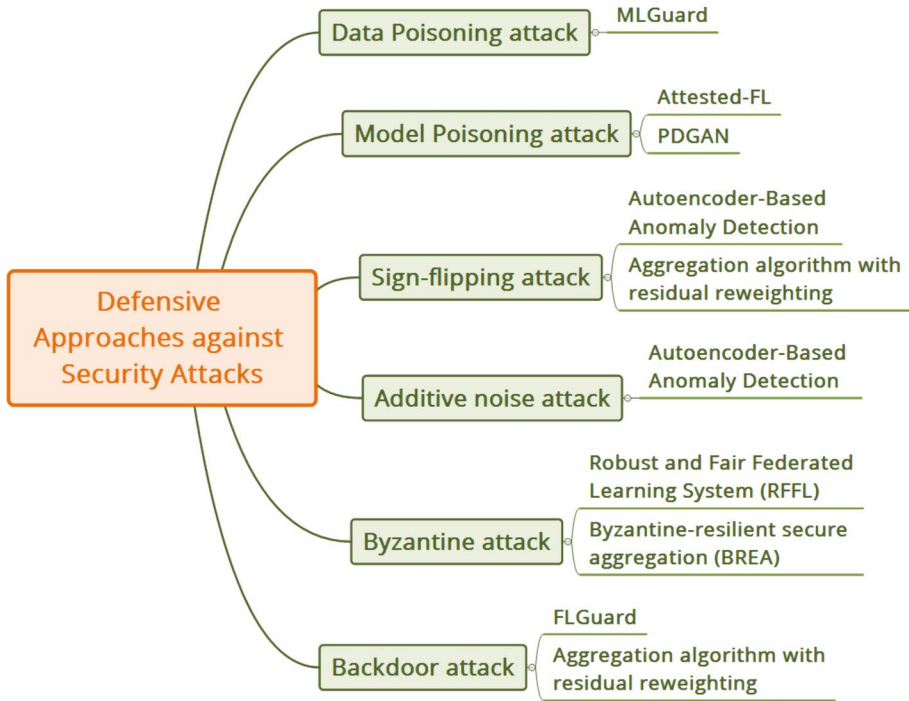


Fig. 10 Defensive approaches against federated learning security attacks

PaddleFL has poor documentation to guide the community. Comparatively, FL and DP have rich documentation and tutorials. Furthermore, it is used to simulate targeted federated attacks. Through attack simulation tools the effectiveness of the security system has been tested. Another federated framework FATE, works well with deep learning but has not applied any DP algorithm. Obviously, it is necessary to consider the TFF, developed by Google, has both high and low-level interfaces for federated learning. Furthermore, it has built-in training datasets but has the longest training time. PySyft from OpenMined has an exception of supporting various operating systems (OS) such as Mac, Linux, Windows, iOS, and Android. Additionally, it is compatible with Pytorch and TensorFlow. Last but not least, FL and DP provide solutions for clustering tasks, implement with AI mechanism, and preserve data privacy.

4.3 Federated learning models against privacy attacks

In previous studies, a few Federated Learning Models (FLM) were discussed, which overcame the issue of privacy leakage. The concept of matrix factorization (MF) was studied under machine learning Koren et al. (2009), but now it required urgent research to investigate the MF from the perspective of federated learning privacy. In existing research, MF privacy-preserving is divided into two types such as obfuscation and encryption-based techniques Kim et al. (2016); Berlioz et al. (2015). In the work of Chai et al. (2020), the authors presented federated matrix factorization (FedMF) and overcame the shortcomings in terms of no accuracy lost, because data obfuscation and

Table 5 Federated Learning Tools/frameworks Implemented in Existing Studied

Developed by	Tools	Operating system	Best features	Shortcomings	References
Google Inc.	TFF	Mac and Linux	Easily integrate TF models Built-in training datasets Provides both low and a highlevel interface for FL	Not compatible with the TensorFlow 2.x Unable to provide DP mechanism	Developers (2021)
Webank's AI department	FATE	Mac and Linux	Provide many FL algorithms Partially able to add new properties	Not applied any DP algorithm Poorly documented High-level interface dependent on the command line	FATE (2021)
Baidu	PaddleFL	Mac, Linux, and Windows	Provide high-level FL interface with DP in terms of SGD PaddleFL implements the SMC	Little documentation Difficult to develop alternative privacy-preserving approaches	Wang (2019)
OpenMined	PySyft	Mac, Linux, Windows, iOS, and Android	Compatible with TensorFlow and PyTorch Introduced the multiple Python notebooks as a learning curve	Does not support low-level FL Not applied model aggregation operators	OpenMined (2021)
Sherpa.AI	FL and DP	Linux and Windows	Rich documentation and tutorials are available Fully able to extend properties and support DP It has federated attack simulation	Partially support other libraries	Sherpa.ai (2021)

crypto services were not implemented. Each user calculates the gradients locally and uploads them to the server in place of raw data. In this way, the FedMF framework ensures model learns by uploading the gradients' information individually from each user at the server. Though it seems that uploading the gradients to the FL server does not leak the information, but it still does. To deal with this situation, the security of FedMF is increased by applying homomorphic encryption. Consequently, the efficiency of FedMF is acceptable in the case of a small dataset, however, in the case of large dataset systems it increases the overhead. Besides, the generative model in federated learning is explained in Hahn and Lee (2020). Researches presented the Gradient-Free FL framework known as GRAFFL to train the model in horizontal distributed settings instead of locally. In GRAFFL, models are locally trained but they have a pre-agreed structure and share the latest gradients with the central server. Furthermore, it ensures that gradients and model structure are not revealed to the local clients. Hence, the global model is trained at the central server (CS) as well as secure the model structure. In addition, Sufficient Auto-Encoder (SuffiAE) is used to obtain a summary of statistics, instead of deploying raw data as it leads to computational inefficiency.

4.4 Defensive approaches against privacy attacks

Federated learning offers substantial improvement for privacy preservation, but, its surface remains vulnerable to privacy attacks as discussed in sect. 3. In the literature, few approaches are available to prevent FL privacy attacks as revealed in Table 6. Xu et al. (2019), introduced the HybridAlpha that implements the DP and SMC protocol to mitigate the threat of malicious aggregators and clients to inferring personal information. The authors applied the approach both theoretically and experimentally in order to guide the cryptosystems to be adequate for FL. For experimental evaluation, the MNIST dataset was used and showed efficient training time, less communication cost, and guaranteed privacy protection. Similarly, Li et al. (2020c), presented the general procedure to protect multimedia privacy in FL during an iteration of a global model. The encryption scheme was introduced and the privacy was improved by removing hidden attributes in FL. Moreover, extensive experiments, which were deployed on various attacks capable of stealing multimedia privacy in FL, were conducted to validate the proposed solution. Finally, discernible results are provided for the eradication of residual multimedia features. Additionally, it has no computational overhead nor accuracy loss. Another model LDP-Fed Truex et al. (2020), is proposed to effectively train the complex models while protecting against inference attacks at the end of each client. On a short note, LDP-fed has an ability to deal with complex models, locally defined privacy guarantees, and protection from inference attacks in the FL environment. In the work of Wang et al. (2019), researchers launched the defensive mechanism against the GAN-based reconstruction attack. The effectiveness and superiority of the mechanism were shown by conducting exhaustive experiments and fruitfully recovering the samples from a particular user. Naseri et al. (2020), introduced the local and central differential privacy (LDP/CDP) for FL to mitigate the membership and property inference attacks. In comparison with LDP, the CDP gives better accuracy in the detection of backdoor attacks. For the future, it is essential to propose more defensive approaches against inference attacks.

Table 6 Defensive Approaches against Privacy/Inference Attacks with Attack Goal and Impact

Defensive approaches	Description	Attacks	Attack goal	Impact of attack	References
HybridAlpha	Deployed SMC protocol with functional encryption.	Inference attacks	Evasion or Exploratory attacks	High	Li et al. (2020c)
InvisibleFL	Implemented the Non-Informative Transformation(NIT) and Just Learn Over Chiper-text (JLoC).	Inference attacks	Evasion or Exploratory attacks	High	Xu et al. (2019)
LDP-Fed	Utilized the Local Differential Privacy(LDP).	Inference attacks	Evasion or Exploratory attacks	High	Truex et al. (2020)
mGAN-AI	Presented the multi-task GANAuxiliary Identification (mGAN-AI) to protect the client level privacy.	Reconstruction attack	Samples recreation	Medium	Wang et al. (2019)
Central/Local Differential Privacy (CDP/LDP)	Works with CDP and LDP to show how it can effectively provide privacy against inference attacks.	Inferring membership and Property Inference	Shadow models	High	Naseri et al. (2020)

4.5 Defensive approaches against CIA security attacks

Defense methods for CIA security attacks help to offer robustness and decrease the possibility of risk in the FL system. A Trusted Execution Environment (TEE) Subramanyan et al. (2017) was introduced as a high-level confidential system for deploying the verified codes. With TEE, digital trust is implemented by securing linked devices in the FL system. The TEE has a competency to create an isolated and cryptographic structure with authorizing end-to-end security against the insertion of incorrect training results. Furthermore, TEE guarantees the confidentiality, integrity, privacy, authentication, and authorization of data. Authors in Chen et al. (2020), applied TEE for providing integrity and privacy in the FL ecosystem. To maintain the correct results of the model, a training integrity protocol is executed with TEE in FL where servers and clients are bound collaboratively. It can detect adversaries who violate the availability of trained models during the communication rounds. Hence, the training integrity scheme is helpful where confidentiality and availability are the concerns in federated learning settings. Guowen et al. (2020), proposed the VerifyNet which aims to verify the correct results received from the FL server. Additionally, a double masking protocol is used to assure the confidentiality of data throughout the FL learning process. Likewise, a Privacy-Preserving Federated Learning (PPFL) Mo et al. (2021), is proposed to protect the FL ecosystem against privacy attacks. In PPFL, TEE is used for local model training on clients and secure aggregation of the model on the server. TEE confirms that gradient updates are obfuscated to adversaries and prevents the hacking of the central FL server, clients, and gradient leakage.

5 Challenges and future directions

Federated learning needs to consider the protection of privacy and security attacks, as well as the detection of dishonest participants who collects the incentives or rewards freely. A set of critical federated learning challenges and potential future directions that need to explore in upcoming researches are defined in this section. Based on our investigation, we have provided the list of questions as below:

5.1 Challenges

1. *High Latency Rate, Low Bandwidth, and Communication Bottlenecks:* The FL involves the notion of increased latency and low bandwidth rates. In conventional cases, it needs low latency for swift learning from the backpropagation method. This task is easily achieved in ML, but with millions of devices used to train the algorithm. It slows down the learning process and creates a high latency rate. Furthermore, bandwidth is a technical issue in FL as most of the settings are accompanied by wifi or 4G. The bandwidth of wifi or 4G is insufficient for the FL environment that induces the high latency and makes the algorithm training process relaxed. The bandwidth of the device has not improved compared to the increased computing power of the device, which creates a bottleneck in communication. It is recommended that 5G and B5G technologies are used in the FL environment and that communication costs are addressed by considering model compression and quantization methods.

2. *Contribution Measurements and Attack Incentives:* In sect. 3, we have discussed the free-rider attack who gained incentives by participating in sending the fake local model updates. In the perspective of incentives, adversaries freely gained the profit and instantaneously decrease the computational resources that have been deployed in the model training process. From an incentive perspective, the adversary is free to gain profit and immediately reduces the computational resources deployed in the model training process. More research is needed on free-rider attacks in terms of contribution measurements and rewards. In further research, other incentive strategies to detect spurious model updates should be explored, which is a more interesting topic. Some methods such as assigning the specific scores to honest and malicious participants after detection may be helpful in recognizing the true participants for algorithm training. Furthermore, defensive strategies are required under FL settings for free-rider attack, as more workers participated, in order to enhance the accuracy of the FL system.
3. *Federated Learning Protection Frameworks:* At this time, some FL tools are available that are implemented in a federated learning environment such as TFF, FATE, PaddleFL, PySyft, and FL and DP which are elaborated in sect. 4.2 with their features and limitations. Among these frameworks, at the moment only FL and DP support the federated attack simulation. Accordingly, to develop FL frameworks with the aim of providing maximum privacy protection features should be the future avenue of research.
4. *Privacy and Security Approaches:* In sect. 3, we have discussed several privacy and security attacks, which show that conventional FL does not guarantee data protection. The updated global model contains traces that cause the leakage of private and sensitive data. Previously, in federated learning, SMC and DP approaches are applied to protect information from adversaries but they have some cons such as hiding the particular updates from the clients and adding noise leading to reduce the model accuracy, respectively. The SMC required cryptographic techniques like Homomorphic Encryption (HE) Hardy et al. (2017), secret sharing, and additive masking Ács and Castelluccia (2011) methods. They cause large computation overhead in terms of encryption and communication cost. Hence, other techniques are required in order to resist potential attacks.
5. *System Heterogeneity and Training:* To train data on different devices is a challenging task and it is essential for federated learning to integrate data from various devices. Furthermore, FL has a slower impact to influence the convergence in comparison with ML. In this context, training smaller models with compressing techniques can increase the speed of convergence. A recent study Mime Karimireddy et al. (2020), is aimed at the development of future algorithms but it requires additional approaches to deal with this problem.

Apart from that, the federated learning environment suffers from multiple attacks as discussed in sect. 3. In the literature, there is no simple and straightforward approach available to defend against FL attacks due to privacy constraints.

5.2 Future directions

A Federated learning ecosystem needs extensive considerations to provide a secure and vulnerability-free surface to achieve optimized outcomes. In this regard, traceability of communication rounds between FL server and participated clients should be established to monitor the activity of model training. Traceability helps to detect internal vulnerabilities of throughout the FL life cycle. The traceability can be achieved through an unforgeable

ledger to make sure what data is employed by which training model. For instance, Owkin (2021), aims to connect the world with the first federated virtual laboratory empowered as privacy-preserving. Before aggregating the model, it should be filter out the model updates related to the malicious and benign ones. However, higher chances are still existed to corrupt the global aggregated model by crafting the local model updates to service providers. In this context, the root of trust between the parties is necessary to establish. Furthermore, implementation of bootstraps trust and assigning the test scores are helpful to cope with trust challenges Cao et al. (2020). Besides, the future may include a hierarchical root of trust structure that would include multiple subsets of samples at different trust levels. As trustworthiness and transparency of data are the critical issues and they are also essential building blocks to secure the system Bertino (2021). We believe that adopting trust and traceability techniques in a federated learning environment can mitigate other attacks. The procedure of bounces to the FL participants and owners also needs much attention in further studies. A structure should be designed in such a way for FL to get the incentives in a fair and sustainable manner. By the way of illustration, influence functions can be used to give incentives to participants by proposing a payment system Richardson et al. (2019). To safeguard the privacy leakage or inference attacks that reveal the sensitive data, Fully Homomorphic Encryption (FHE) techniques can be utilized, taking into account the high encryption overhead and data communication cost. In previous studies, to tackle the security issues, DP and SMC techniques are executed. The DP and SMC techniques have some negative effects as explained in the challenges section. Accordingly, enforcement of matrix transformation can be used to encrypt the minor part of transformed results. Through functional encryption, it is possible to achieve secure aggregation Wu et al. (2020). No doubt, a tremendous amount of efforts are in progress and federated learning studies are still need more and more in forthcoming researches. Each federated learning attack poses a unique property and require potential protection approaches in the future. Finally, in this way, advanced security approaches are needed to deal with confidential model updates, secure aggregation, trust, monitoring the entire FL process, and minimize computation overhead during multiple iterations, and encryption techniques to defend against suspicious attacks.

6 Conclusion

In this paper, the general federated learning process together with its vulnerable surface have been exposed. Adversaries exploit the FL environment through poisoned data, malicious model updates, leak information, gain incentives freely, and compromised the integrity of the model. In this context, the attack surface of FL is comprehensively illustrated in the taxonomy as poisoning, inference, free-rider, and CIA security triad attacks. Besides, FL viable attacks are explained with their types, the impact of attack and nature, etc. Limitations of the existing research are described in terms of federated learning attacks, as well as future work with suitable detection measures of FL attacks. To tackle privacy and security attacks state-of-the-art defensive approaches along with root attacks are defined. Furthermore, protection techniques against authentication and authorization are given under the umbrella term CIA security defense. Some FL tools support privacy protection and provide an attack simulations environment, however, deficiencies are also elaborated in this paper. In the end, challenges of FL and future directions are suggested to build

an uninterrupted and durable system that can be persistent to defend against adversarial attacks. The main findings of this paper in mastering FL are summarized as follows:

1. The federated learning workflow process along with a thorough analysis of the malicious FL environment and its classification are provided.
2. Identifies a taxonomy of privacy and security vulnerabilities and FL attack surfaces, as well as the CIA security triad attacks.
3. Furthermore, state-of-the-art defensive approaches associated with FL attacks, famous models against security and privacy attacks and, tools/frameworks with their best features and limitations are discussed.
4. Finally, we highlight the insights into the key federated learning challenges inherent to security problems and future directions for improving protection mechanisms.

We hope that this paper provides fundamental insights for the entire research community to grasp with FL attack surface and defensive measures. However, additional FL methodologies, security models, strategies, and protection frameworks should be considered in forthcoming research. Hence, it demands more exploration that provides stronger guarantees against FL attacks.

References

- Araki T, Furukawa J, Lindell Y, Nof A, Ohara K (2016) High-throughput semi-honest secure three-party computation with an honest majority. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM, <https://doi.org/10.1145/2976749.2978331>
- Ács G, Castelluccia C (2011) I have a DREAM! (Differentially privatE smArt metering). In Information Hiding, pages 118–132. Springer Berlin Heidelberg, https://doi.org/10.1007/978-3-642-24178-9_9
- Bagdasaryan E, Veit A, Hua Y, Estrin D, Shmatikov V (2020) How to backdoor federated learning. In International Conference on Artificial Intelligence and Statistics, pp 2938–2948. PMLR
- Baruch M, Baruch G, Goldberg Y (2019) A little is enough: Circumventing defenses for distributed learning. arXiv preprint [arXiv:1902.06156](https://arxiv.org/abs/1902.06156)
- Berlioz A, Friedman A, Kaafar MA, Boreli R, Berkovsky S (2015) Applying differential privacy to matrix factorization. In Proceedings of the 9th ACM Conference on Recommender Systems. ACM, <https://doi.org/10.1145/2792838.2800173>
- Bertino E (2021) Attacks on artificial intelligence [last word]. *IEEE Secur Privacy* 19(1):103–104
- Bhagoji AN, Chakraborty S, Mittal P, Calo S (2019) Analyzing federated learning through an adversarial lens. In *International Conference on Machine Learning*, pp 634–643. PMLR
- Bhowmick A, Duchi J, Freudiger J, Kapoor G, Rogers R (2018) Protection against reconstruction and its applications in private federated learning. arXiv preprint [arXiv:1812.00984](https://arxiv.org/abs/1812.00984)
- Blanchard P, Mhamdi EM, Guerraoui R, Stainer J (2017) Machine learning with adversaries: Byzantine tolerant gradient descent. In Proceedings of the 31st International Conference on Neural Information Processing Systems, pp 118–128
- Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E et al. (2021) On the opportunities and risks of foundation models. arXiv preprint [arXiv:2108.07258](https://arxiv.org/abs/2108.07258)
- Bonawitz K, Ivanov V, Kreuter B, Marcedone A, McMahan HB, Patel S, Ramage D, Segal A, Seth K (2016) Practical secure aggregation for federated learning on user-held data. arXiv preprint [arXiv:1611.04482](https://arxiv.org/abs/1611.04482)
- Bonawitz K, Ivanov V, Kreuter B, Marcedone A, McMahan HB, Patel S, Ramage D, Segal A, Seth K (2017) Practical secure aggregation for privacy-preserving machine learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. ACM, <https://doi.org/10.1145/3133956.3133982>
- CPRA (2020) California privacy rights act, <https://www.caprivacy.org/>
- Caldas S, Duddu Sai MK, Wu P, Li T, Konečný J, McMahan HB, Smith V, Talwalkar A (2018) Leaf: A benchmark for federated settings. arXiv preprint [arXiv:1812.01097](https://arxiv.org/abs/1812.01097)

- Cao X, Fang M, Liu J, Gong NZ (2020) Fltrust: Byzantine-robust federated learning via trust bootstrapping. arXiv preprint [arXiv:2012.13995](https://arxiv.org/abs/2012.13995)
- Chai D, Wang L, Chen K, Yang Q (2020) Secure federated matrix factorization. *IEEE Intelligent Systems*, <https://doi.org/10.1109/mis.2020.3014880>
- Chen Y, Luo F, Li T, Xiang T, Liu Z, Li J (2020) A training-integrity privacy-preserving federated learning scheme with trusted execution environment. *Inf Sci* 522:69–79. <https://doi.org/10.1016/j.ins.2020.02.037>
- Chen Y, Qin X, Wang J, Chaohui Yu, Gao W (2020) FedHealth: A federated transfer learning framework for wearable healthcare. *IEEE Intell Syst* 35(4):83–93. <https://doi.org/10.1109/mis.2020.2988604>
- Chen J, Zhang J, Zhao Y, Han H, Zhu K, Chen B (2020) Beyond model-level membership privacy leakage: an adversarial approach in federated learning. In 2020 29th International Conference on Computer Communications and Networks (ICCCN). IEEE, <https://doi.org/10.1109/icccn49398.2020.9209744>
- Cheng Y, Liu Y, Chen T, Yang Q (2020) Federated learning for privacy-preserving AI. *Commun ACM* 63(12):33–36. <https://doi.org/10.1145/3387107>
- Cheng K, Fan T, Jin Y, Liu Y, Chen T, Papadopoulos D, Yang Q (2019) Secureboost: A lossless federated learning framework. arXiv preprint [arXiv:1901.08755](https://arxiv.org/abs/1901.08755)
- Chik WB (2013) The singapore personal data protection act and an assessment of future trends in data privacy reform. *Comput Law Secur Rev* 29(5):554–575. <https://doi.org/10.1016/j.clsr.2013.07.010>
- Cohen G, Afshar S, Tapson J, Van Schaik A (2017) Emnist: Extending mnist to handwritten letters. In 2017 International Joint Conference on Neural Networks (IJCNN), pages 2921–2926. IEEE
- Developers TensorFlow (2021) Tensorflow. <https://doi.org/10.5281/ZENODO.4724125>
- Dua D, Graff C (2017) Machine learning repository, URL: <http://archive.ics.uci.edu/ml/index.php>
- El Mhamdi EM, Guerraoui R, Rouault SL (2018) The hidden vulnerability of distributed learning in byzantium. arXiv preprint [arXiv:1802.07927](https://arxiv.org/abs/1802.07927)
- FATE (2021) An industrial graded federated learning framework, URL: <https://fate.fedai.org/>
- Fang M, Cao J, Jia J, Gong N (2020) Local model poisoning attacks to byzantine-robust federated learning. In 29th USENIX Security Symposium (USENIX Security 20), pp 1605–1622
- FeatureCloud (2021) Transforming health care and medical research with federated learning, URL: <https://featurecloud.eu/about/our-vision/>
- FedAI (2020) Webank and swiss re signed cooperation mou, URL: <https://www.fedai.org/news/webank-and-swiss-re-signed-cooperation-mou/>
- Feldman M, Papadimitriou C, Chuang J, Stoica I (2006) Free-riding and whitewashing in peer-to-peer systems. *IEEE J Sel Areas Commun* 24(5):1010–1019. <https://doi.org/10.1109/jsac.2006.872882>
- Fernandes K, Vinagre P, Cortez P (2015) A proactive intelligent decision support system for predicting the popularity of online news. In *Progress in Artificial Intelligence*, pages 535–546. Springer International Publishing, https://doi.org/10.1007/978-3-319-23485-4_53
- Fraboni Y, Vidal R, Lorenzi M (2021) Free-rider attacks on model aggregation in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pp 1846–1854. PMLR
- Fredrikson M, Jha S, Ristenpart T (2015) Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, <https://doi.org/10.1145/2810103.2813677>
- Fu S, Xie C, Li B, Chen Q (2019) Attack-resistant federated learning with residual-based reweighting. arXiv preprint [arXiv:1912.11464](https://arxiv.org/abs/1912.11464)
- Fung C, Yoon CJM, Beschastnikh I (2020) The limitations of federated learning in sybil settings. In 23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID'2020), pp 301–316
- Fung C, Yoon CJ, Beschastnikh I (2018) Mitigating sybils in federated learning poisoning. arXiv preprint [arXiv:1808.04866](https://arxiv.org/abs/1808.04866)
- Geyer Robin C, Klein Tassilo, Nabi Moin (2017) Differentially private federated learning: A client level perspective. arXiv preprint [arXiv:1712.07557](https://arxiv.org/abs/1712.07557)
- Goodfellow IJ, Erhan D, Carrier PL, Courville A, Mirza M, Hamner B, Cukierski W, Tang Y, Thaler D, Lee DH, Zhou Y et al. (2013) Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, pp 117–124. Springer
- Google BigQuery (2017) Reddit dataset, URL: <https://www.reddit.com/r/bigquery/wiki/datasets>
- Guowen X, Li H, Liu S, Yang K, Lin X (2020) VerifyNet: Secure and verifiable federated learning. *IEEE Trans Inf Forensics Secur* 15:911–926. <https://doi.org/10.1109/tifs.2019.2929409>
- Hahn SJ, Lee J (2020) Graffl: Gradient-free federated learning of a bayesian generative model. arXiv preprint [arXiv:2008.12925](https://arxiv.org/abs/2008.12925)

- Hardy S, Henecka W, Ivey-Law H, Nock R, Patrini G, Smith G, Thorne B (2017) Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. arXiv preprint [arXiv:1711.10677](https://arxiv.org/abs/1711.10677)
- He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
- He Z, Zhang T, Lee RB (2019) Model inversion attacks against collaborative inference. In Proceedings of the 35th Annual Computer Security Applications Conference. ACM, <https://doi.org/10.1145/3359789.3359824>
- Hitaj B, Ateniese G, Perez-Cruz F (2017) Deep models under the GAN. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. ACM, <https://doi.org/10.1145/3133956.3134012>
- House W (2012) Consumer data privacy in a networked world: A framework for protecting privacy and promoting innovation in the global digital economy. White House, Washington, DC, pp 1–62
- Huang W, Li T, Wang D, Du S, Zhang J (2020) Fairness and accuracy in federated learning. arXiv preprint [arXiv:2012.10069](https://arxiv.org/abs/2012.10069)
- Huang L, Joseph AD, Nelson B, Rubinstein BIP, Tygar JD (2011) Adversarial machine learning. In Proceedings of the 4th ACM workshop on Security and artificial intelligence - AISec '11. ACM Press, <https://doi.org/10.1145/2046684.2046692>
- Jie X, Glicksberg BS, Chang S, Walker P, Bian J, Wang F (2020) Federated learning for healthcare informatics. *J Healthcare Informatics Res* 5(1):1–19. <https://doi.org/10.1007/s41666-020-00082-4>
- Kaggle (2013) Acquire valued shoppers challenge, URL: <https://www.kaggle.com/c/acquire-valued-shoppers-challenge/data>
- Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Bhagoji AN, Bonawitz K, Charles Z, Cormode G, Cummings R et al. (2019) Advances and open problems in federated learning. arXiv preprint [arXiv:1912.04977](https://arxiv.org/abs/1912.04977)
- Kang J, Xiong Z, Niyato D, Yu H, Liang YC, Kim DI (2019) Incentive design for efficient federated learning in mobile networks: A contract theory approach. In 2019 IEEE VTS Asia Pacific Wireless Communications Symposium (APWCS). IEEE, <https://doi.org/10.1109/vts-apwcs.2019.8851649>
- Kanwendy. Lending club loan data, 2019. URL: <https://www.kaggle.com/wendykan/lending-club-loan-data>
- Karimireddy SP, Jaggi M, Kale S, Mohri M, Reddi SJ, Stich SU, Suresh AT (2020) Mime: Mimicking centralized stochastic algorithms in federated learning. arXiv preprint [arXiv:2008.03606](https://arxiv.org/abs/2008.03606)
- Khazbak Y, Tan T, Cao G (2020) MLGuard: Mitigating poisoning attacks in privacy preserving distributed collaborative learning. In 2020 29th International Conference on Computer Communications and Networks (ICCCN). IEEE, <https://doi.org/10.1109/iccn49398.2020.9209670>
- Kim S, Kim J, Koo D, Kim Y, Yoon H, Shin J (2016) Efficient privacy-preserving matrix factorization via fully homomorphic encryption. In Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security. ACM, <https://doi.org/10.1145/2897845.2897875>
- Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. *Computer* 42(8):30–37. <https://doi.org/10.1109/mc.2009.263>
- Krizhevsky Alex, Hinton Geoffrey, et al. (2009) Learning multiple layers of features from tiny images
- Kuchler H (2019) Pharma groups combine to promote drug discovery with ai, URL: <https://www.ft.com/content/ef7be832-86d0-11e9-a028-86cea8523dc2>
- Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. *Proc IEEE* 86(11):2278–2324. <https://doi.org/10.1109/5.726791>
- Li H, Ota K, Dong M (2018) Learning IoT in edge: Deep learning for the internet of things with edge computing. *IEEE Network* 32(1):96–101. <https://doi.org/10.1109/mnet.2018.1700202>
- Li T, Sahu AK, Talwalkar A, Smith V (2020) IEEE Signal Process Mag. Federated learning: challenges, methods, and future directions. 37(3):50–60. <https://doi.org/10.1109/msp.2020.2975749>
- Li Z, Sharma V, Mohanty SP (2020) Preserving data privacy via federated learning: Challenges and solutions. *IEEE Consumer Electron Mag* 9(3):8–16. <https://doi.org/10.1109/mce.2019.2959108>
- Li L, Wei X, Chen T, Giannakis GB, Ling Q (2019) RSA: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets. *Proceed AAAI Conf Artif Intell* 33:1544–1551. <https://doi.org/10.1609/aaai.v33i01.33011544>
- Li Q, Zhu W, Wu C, Pan X, Yang F, Zhou Y, Zhang Y (2020) InvisibleFL: Federated learning over non-informative intermediate updates against multimedia privacy leakages. In Proceedings of the 28th ACM International Conference on Multimedia. ACM, <https://doi.org/10.1145/3394171.3413923>
- Li S, Cheng Y, Liu Y, Wang W, Chen T (2019) Abnormal client behavior detection in federated learning. arXiv preprint [arXiv:1910.09933](https://arxiv.org/abs/1910.09933)
- Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V (2018) Federated optimization in heterogeneous networks. arXiv preprint [arXiv:1812.06127](https://arxiv.org/abs/1812.06127)

- Lim HK, Kim JB, Kim CM, Hwang GY, Choi HB, Han YH (2020) Federated reinforcement learning for controlling multiple rotary inverted pendulums in edge computing environments. In 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIC). IEEE. <https://doi.org/10.1109/icaic48513.2020.9065233>
- Lin J, Du M, Liu J (2019) Free-riders in federated learning: Attacks and defenses. arXiv preprint [arXiv:1911.12560](https://arxiv.org/abs/1911.12560)
- Lin Y, Han S, Mao H, Wang Y, Dally WJ (2017) Deep gradient compression: Reducing the communication bandwidth for distributed training. arXiv preprint [arXiv:1712.01887](https://arxiv.org/abs/1712.01887)
- Liu Y, Huang A, Luo Y, Huang H, Liu Y, Chen Y, Feng L, Chen T, Han Yu, Yang Q (2020) FedVision: An online visual object detection platform powered by federated learning. Proceed AAAI Conf Artif Intell 34(08):13172–13179. <https://doi.org/10.1609/aaai.v34i08.7021>
- Liu Y, Kang Y, Xing C, Chen T, Yang Q (2020) A secure federated transfer learning framework. IEEE Intell Syst 35(4):70–82. <https://doi.org/10.1109/mis.2020.2988525>
- Long G, Tan Y, Jiang J, Zhang C (2020) Federated learning for open banking. In *Lecture Notes in Computer Science*, pages 240–254. Springer International Publishing, https://doi.org/10.1007/978-3-030-63076-8_17
- Luo X, Wu Y, Xiao X, Ooi BC (2020) Feature inference attack on model predictions in vertical federated learning. arXiv preprint [arXiv:2010.10152](https://arxiv.org/abs/2010.10152)
- Luo X, Zhu X (2020) Exploiting defenses against gan-based feature inference attacks in federated learning. arXiv preprint [arXiv:2004.12571](https://arxiv.org/abs/2004.12571)
- Lyu L, Yu H, Ma X, Sun L, Zhao J, Yang Q, Yu PS (2020) Threats to federated learning. In *Lecture Notes in Computer Science*, pages 3–16. Springer International Publishing, https://doi.org/10.1007/978-3-030-63076-8_1
- Ma C, Li J, Ding M, Yang HH, Shu F, Quek TQS, Vincent Poor H (2020) On safeguarding privacy and security in the framework of federated learning. IEEE Network 34(4):242–248. <https://doi.org/10.1109/mnet.001.1900506>
- Ma Y, Zhu X, Hsu J (2019) Data poisoning against differentially-private learners: Attacks and defenses. arXiv preprint [arXiv:1903.09860](https://arxiv.org/abs/1903.09860)
- Mallah RA, Lopez D, Farooq B (2021) Untargeted poisoning attack detection in federated learning via behavior attestation. arXiv preprint [arXiv:2101.10904](https://arxiv.org/abs/2101.10904)
- McMahan HB, Ramage D, Talwar K, Zhang L (2017) Learning differentially private recurrent language models. arXiv preprint [arXiv:1710.06963](https://arxiv.org/abs/1710.06963)
- McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA (2017) Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR
- McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA (2016) Federated learning of deep networks using model averaging. arXiv preprint [arXiv:1602.05629](https://arxiv.org/abs/1602.05629)
- Melis L, Song C, De Cristofaro E, Shmatikov V (2019) Exploiting unintended feature leakage in collaborative learning. In 2019 IEEE Symposium on Security and Privacy (SP). IEEE, <https://doi.org/10.1109/sp.2019.00029>
- Mo F, Haddadi H, Katevas K, Marin E, Perino D, Kourtellis N (2021) Ppfl: Privacy-preserving federated learning with trusted execution environments. arXiv preprint [arXiv:2104.14380](https://arxiv.org/abs/2104.14380)
- Moro S, Cortez P, Rita P (2014) A data-driven approach to predict the success of bank telemarketing. *Decis Support Syst* 62:22–31. <https://doi.org/10.1016/j.dss.2014.03.001>
- Musketeer. Smart manufacturing and health care, 2020. URL: <https://musketeer.eu/project/>
- Nadiger C, Kumar A, Abdelhak S (2019) Federated reinforcement learning for fast personalization. In 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE). IEEE, <https://doi.org/10.1109/aike.2019.00031>
- Naseri M, Hayes J, Emiliano DC (2020) Toward robustness and privacy in federated learning: Experimenting with local and central differential privacy. arXiv preprint [arXiv:2009.03561](https://arxiv.org/abs/2009.03561)
- Nasr M, Shokri R, Houmansadr A (2019) 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, <https://doi.org/10.1109/sp.2019.00065>
- Nguyen TD, Rieger P, Yalame H, Mollering H, Fereidooni H, Marchal S, Miettinen M, Mirhoseini A, Sadeghi AR, Schneider T et al. (2021) FGuard: Secure and private federated learning. arXiv preprint [arXiv:2101.02281](https://arxiv.org/abs/2101.02281)
- Nilsson A, Smith S, Gustavsson E, Jirstrand M (2018) A performance evaluation of federated learning algorithms. In *Proceedings of the Second Workshop on Distributed Infrastructures for Deep Learning*. ACM, <https://doi.org/10.1145/3286490.3286559>

- Nishio T, Yonetani R (2019) Client selection for federated learning with heterogeneous resources in mobile edge. In ICC 2019 - 2019 IEEE International Conference on Communications (ICC). IEEE, <https://doi.org/10.1109/icc.2019.8761315>
- Nock R, Hardy S, Henecka W, Ivey-Law H, Patrini G, Smith G, Thorne B (2018) Entity resolution and federated learning get a federated resolution. arXiv preprint [arXiv:1803.04035](https://arxiv.org/abs/1803.04035)
- OpenMined (2021) Let's solve privacy, URL: <https://www.openmined.org/>
- Owkin. Federated learning, 2021. URL: <https://owkin.com/federated-learning/>
- O'Driscoll A (2021) 30+ data breach statistics and facts, <https://www.comparitech.com/blog/vpn-privacy/data-breach-statistics-facts/>
- Paul V, von dem Axel B (2017) The EU General data protection regulation (GDPR). Springer International Publishing, Berlin. <https://doi.org/10.1007/978-3-319-57959-7>
- Phong LT, Aono Y, Hayashi T, Wang L, Moriai S (2018) Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Trans Inf Forensics Secur* 13(5):1333–1345. <https://doi.org/10.1109/tifs.2017.2787987>
- Pustozero A, Mayer R (2020) Information leaks in federated learning. In Proceedings of the Network and Distributed System Security Symposium
- Radanliev P, De Roure D (2021) Review of algorithms for artificial intelligence on low memory devices. *IEEE Access* 9:109986–109993
- Radanliev P, De Roure D, Burnap P, Santos O (2021) Epistemological equation for analysing uncontrollable states in complex systems: Quantifying cyber risks from the internet of things. *The Review of Socio-network Strategies*, pp 1–31
- Richardson A, Filos-Ratsikas A, Faltings B (2019) Rewarding high-quality data via influence functions. arXiv preprint [arXiv:1908.11598](https://arxiv.org/abs/1908.11598)
- Samarakoon S, Bennis M, Saad W, Debbah M (2020) Distributed federated learning for ultra-reliable low-latency vehicular communications. *IEEE Trans Commun* 68(2):1146–1159. <https://doi.org/10.1109/tcomm.2019.2956472>
- Samaria FS, Harter AC (1994) Parameterisation of a stochastic model for human face identification. In Proceedings of 1994 IEEE Workshop on Applications of Computer Vision. IEEE Comput Soc Press <https://doi.org/10.1109/acv.1994.341300>
- Satariano A (2019) Google is fined 57 million under europe's data privacy law URL: <https://www.nytimes.com/2019/01/21/technology/google-europe-gdpr-fine.html>
- Sherpa.ai. (2021) We research and build artificial intelligence technology and services, URL: <https://sherpa.ai/>
- Shokri R, Stronati M, Song C, Shmatikov V (2017) Membership inference attacks against machine learning models. In 2017 IEEE Symposium on Security and Privacy (SP). IEEE
- Smith SL, Kindermans PJ, Ying C, Le QV (2017) Don't decay the learning rate, increase the batch size. arXiv preprint [arXiv:1711.00489](https://arxiv.org/abs/1711.00489)
- So J, Guler B, Avestimehr AS (2020) Byzantine-resilient secure federated learning. *IEEE J Sel Areas Commun*. <https://doi.org/10.1109/jsac.2020.3041404>
- Song M, Wang Z, Zhang Z, Song Y, Wang Q, Ren J, Qi H (2019) Beyond inferring class representatives: User-level privacy leakage from federated learning. In IEEE INFOCOM 2019 - IEEE Conference on Computer Communications. IEEE, <https://doi.org/10.1109/infocom.2019.8737416>
- Stich SU (2018) Local sgd converges fast and communicates little. arXiv preprint [arXiv:1805.09767](https://arxiv.org/abs/1805.09767)
- Subramanyan P, Sinha R, Lebedev I, Devadas S, Seshia SA (2017) A formal foundation for secure remote execution of enclaves. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. ACM, <https://doi.org/10.1145/3133956.3134098>
- Sun Z, Kairouz P, Suresh AT, McMahan HB (2019) Can you really backdoor federated learning? arXiv preprint [arXiv:1911.07963](https://arxiv.org/abs/1911.07963)
- Tan K, Bremner D, Le Kernec J, Imran M (2020) Federated machine learning in vehicular networks: A summary of recent applications. In 2020 International Conference on UK-China Emerging Technologies (UCET). IEEE, <https://doi.org/10.1109/ucet51115.2020.9205482>
- Tolpegin V, Truex S, Gursoy ME, Liu L (2020) Data poisoning attacks against federated learning systems. In *Computer Security – ESORICS 2020*, pages 480–501. Springer International Publishing. https://doi.org/10.1007/978-3-030-58951-6_24
- Truex S, Liu L, Chow K-H, Gursoy ME, Wei W (2020) LDP-fed. In Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking. ACM, <https://doi.org/10.1145/3378679.3394533>
- Truex S, Liu L, Gursoy ME, Yu L, Wei W (2019) Demystifying membership inference attacks in machine learning as a service. *IEEE Transactions on Services Computing*, pages 1–1. <https://doi.org/10.1109/tsc.2019.2897554>

- Tschandl P, Rosendahl C, Kittler H (2018) The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientif Data* 5(1):1–9
- Tseng Y-M, Chen F-G (2011) A free-rider aware reputation system for peer-to-peer file-sharing networks. *Expert Syst Appl* 38(3):2432–2440. <https://doi.org/10.1016/j.eswa.2010.08.032>
- Wang H (2019) Baidu paddlepaddle releases 21 new capabilities to accelerate industry-grade model development, URL: <http://research.baidu.com/Blog/index-view?id=126>
- Wang H, Yurochkin M, Sun Y, Papailiopoulos D, Khazaeni Y (2020) Federated learning with matched averaging. arXiv preprint [arXiv:2002.06440](https://arxiv.org/abs/2002.06440)
- Wang L, Xu S, Wang X, Zhu Q (2019) Eavesdrop the composition proportion of training labels in federated learning. arXiv preprint [arXiv:1910.06044](https://arxiv.org/abs/1910.06044)
- Wei O, Zeng J, Guo Z, Yan W, Liu D, Fuentes S (2020) A homomorphic-encryption-based vertical federated learning scheme for risk management. *Comput Sci Inf Syst* 17(3):819–834. <https://doi.org/10.2298/osis190923022o>
- Wu D, Pan M, Xu Z, Zhang Y, Han Z (2020) Towards efficient secure aggregation for model update in federated learning. In *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*. IEEE, <https://doi.org/10.1109/globecom42002.2020.9347960>
- Xiao H, Rasul K, Vollgraf R (2017) Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint [arXiv:1708.07747](https://arxiv.org/abs/1708.07747)
- Xie C, Huang K, Chen PY, Li B (2019) Dba: Distributed backdoor attacks against federated learning. In *International Conference on Learning Representations*
- Xu X, Lyu L (2020) Towards building a robust and fair federated learning system. arXiv preprint [arXiv:2011.10464](https://arxiv.org/abs/2011.10464)
- Xu R, Baracaldo N, Zhou Y, Anwar A, Ludwig H (2019) HybridAlpha. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security - AISec'19*. ACM Press, <https://doi.org/10.1145/3338501.3357371>
- Yang Q, Liu Y, Chen T, Tong Y (2019) Federated machine learning. *ACM Trans Intell Syst Technol* 10(2):1–19. <https://doi.org/10.1145/3298981>
- Yang D, Zhang D, Chen L, Qu B (2015) NationTelescope: Monitoring and visualizing large-scale collective behavior in LBSNs. *J Netw Comput Appl* 55:170–180. <https://doi.org/10.1016/j.jnca.2015.05.010>
- Yang Z, Zhang J, Chang EC (2019) Adversarial neural network inversion via auxiliary knowledge alignment. arXiv preprint [arXiv:1902.08552](https://arxiv.org/abs/1902.08552)
- Yeh I-C, Lien C (2009) The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Syst Appl* 36(2):2473–2480. <https://doi.org/10.1016/j.eswa.2007.12.020>
- Yelp. Yelp open dataset, 2020. URL: <https://www.yelp.com/dataset>
- Zhang C, Li S, Xia J, Wang W, Yan F, Liu Y (2020) Batchcrypt: Efficient homomorphic encryption for cross-silo federated learning. In *2020 USENIX Annual Technical Conference (USENIXATC 20)*, pp 493–506
- Zhang W, Tople S, Ohrimenko O (2020) Dataset-level attribute leakage in collaborative learning. arXiv preprint [arXiv:2006.07267](https://arxiv.org/abs/2006.07267)
- Zhao Y, Chen J, Zhang J, Wu D, Teng J, Yu S (2020) PDGAN: A novel poisoning defense method in federated learning using generative adversarial network. In *Algorithms and Architectures for Parallel Processing*, pages 595–609. Springer International Publishing, https://doi.org/10.1007/978-3-030-38991-8_39
- Zhao B, Mopuri KR, Bilien H (2020) idlg: Improved deep leakage from gradients. arXiv preprint [arXiv:2001.02610](https://arxiv.org/abs/2001.02610)
- Zheng Z, Zhou Y, Sun Y, Wang Z, Liu B, Li K (2021) Federated learning in smart cities: A comprehensive survey. arXiv e-prints, pages arXiv–2102
- Zhou X, Zhang X, Yiming W, Zheng N (2021) Deep model poisoning attack on federated learning. *Future Internet* 13(3):73. <https://doi.org/10.3390/fi13030073>
- Zhu L, Han S (2020) Deep leakage from gradients. In *Lecture Notes in Computer Science*, pages 17–31. Springer International Publishing, https://doi.org/10.1007/978-3-030-63076-8_2
- Zong B, Song Q, Min MR, Cheng W, Lumezanu C, Cho D, Chen H (2018) Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations*