

Trusted AI in Multi-agent Systems: An Overview of Privacy and Security for Distributed Learning

Chuan Ma, *Member, IEEE*, Jun Li, *Senior Member, IEEE*, Kang Wei, *Member, IEEE*,
Bo Liu, *Senior Member, IEEE*, Ming Ding, *Senior Member, IEEE*, Long Yuan,
Zhu Han, *Fellow, IEEE*, and H. Vincent Poor, *Life Fellow, IEEE*

Abstract—Motivated by the advancing computational capacity of distributed end-user equipment (UE), as well as the increasing concerns about sharing private data, there has been considerable recent interest in machine learning (ML) and artificial intelligence (AI) that can be processed on distributed UEs. Specifically, in this paradigm, parts of a ML process are outsourced to multiple distributed UEs. Then the processed information is aggregated on a certain level at a central server, which turns a centralized ML process into a distributed one, and brings about significant benefits. However, this new distributed ML paradigm raises new risks of privacy and security issues. In this paper, we provide a survey on the emerging security and privacy risks of distributed ML from a unique perspective of information exchange levels, which are defined according to the key steps of a ML process, i.e.: i) the level of pre-processed data, ii) the level of learning models, iii) the level of extracted knowledge and, iv) the level of intermediate results. We explore and analyze the potential of threats for each information exchange level based on an overview of the current state-of-the-art attack mechanisms, and then discuss the possible defense methods against such threats. Finally, we complete the survey by providing an outlook on the challenges and possible directions for future research in this critical area.

This work was supported in part by the National Key RD Program of China 2022YFF0712100, in part by the National Natural Science Foundation of China under Grant 62002170 and 61902184, in part by the Fundamental Research Funds for the Central Universities with No. 30921013104, in part by the Science and Technology on Information Systems Engineering Laboratory WZC20205250411 in part by the Future Network Grant of Provincial Education Board in Jiangsu, in part by the Youth Foundation Project (No. K2023PD0AA01) of Zhejiang Lab, in part by the Research Initiation Project of Zhejiang Lab, and in part by the U.S National Science Foundation under Grants CNS-2128448 and ECCS-2335876. (Corresponding authors: Jun Li and Kang Wei)

Chuan Ma is with Zhejiang Laboratory, Hangzhou 311121. He is also with Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China. (Email: chuan.ma@zhejianglab.edu.cn)

Jun Li and Kang Wei are with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210096, China. Kang Wei is also with the Department of Computing, Hong Kong Polytechnic University, Hong Kong 999077, China (Email: {jun.li, kang.wei}@njust.edu.cn).

Bo Liu is with the University of Technology Sydney, NSW 2007, Australia (Email: bo.liu@uts.edu.au).

Ming Ding is with Data61, CSIRO, NSW 2015, Australia (Email: Ming.Ding@data61.csiro.au).

Long Yuan is with the School of Computer Science, Nanjing University of Science and Technology, Nanjing 210096, China (Email: longyuan@njust.edu.cn).

Zhu Han is with the Department of Electrical and Computer Engineering in the University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul, South Korea (Email: zhan2@uh.edu).

H. Vincent Poor is with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544 USA (Email: poor@princeton.edu).

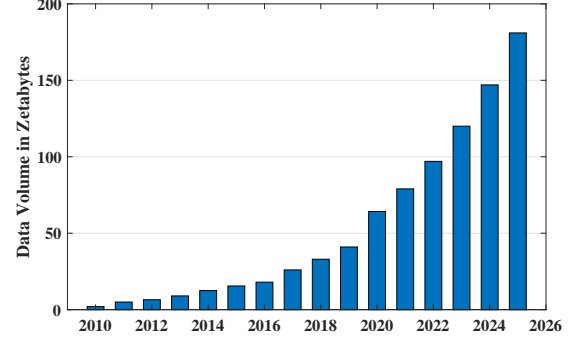


Fig. 1. Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025.

Index Terms—Trusted AI, Multi-agent Systems, Distributed Machine Learning, Federated Learning, Privacy, Security

I. INTRODUCTION

An explosive growth in data availability arising from proliferating Internet of Things (IoT) and 5G/6G technologies, combined with the availability of increasing computational resources through cloud and data servers, promote the applications of machine learning (ML) in many domains (e.g., finance, health-care, industry, and smart city). ML technologies, e.g., deep learning, have revolutionized the ways that information is extracted with ground-breaking successes in various areas [1]. Meanwhile, owing to the advent of IoT, the number of intelligent applications with edge computing, such as smart manufacturing, intelligent transportation, and intelligent logistics, is growing dramatically.

As such, conventional centralized deep learning can no longer efficiently process the dramatically increased amount of data from the massive number of IoT or edge devices. For example, as shown in Fig. 1, the expected volume of data will be 181 zetabytes in 2025¹. In addition, the long runtime of training models steers solution designers towards using distributed systems for an increase of parallelization and the total amount of wireless bandwidth, as the training data required for sophisticated applications can easily be on the order of terabytes [2]. Examples include transaction processing for larger enterprises on data that is stored in

¹<https://www.statista.com/statistics/871513/worldwide-data-created/>

different locations [3] or astronomical data that is too large to move and collect [4].

To address this challenge, distributed learning frameworks have emerged. A typical distributed learning fashion involves the cooperation of multiple clients and servers, which thus involves a decentralization and aggregation process along with the ML process [5]. With the increasing capability of edge devices, distributed clients are able to execute simple ML tasks. For example, federated learning (FL) [6]–[8] enables the decoupling of data provisioning by distributed clients and aggregating ML models at a centralized server. In certain ML tasks, the model sometimes can be so large that it cannot be trained in a reasonable amount of time, and cannot run completely on a single machine. Therefore, large-scale distributed ML is proposed in [9] where datasets in each client will be re-analyzed and pre-trained locally, and the knowledge is aggregated by a central server. In addition, aggregating learning results [10] by the server is another part of distributed ML technology.

To complete a ML task successfully, we need to preserve the integrity and security of the system, along with the privacy of participating clients. As the manufacturers can potentially fail to implement a robust security system in distributed devices, experts on security have warned of potential risks of large numbers of unsecured devices connecting to the Internet [11]. Security and privacy are very significant issues for distributed ML, which introduce a new level of emergent concerns for participants. This is because these devices collect not only personal and sensitive information, e.g., names and telephone numbers but also monitor daily activities. Due to the regular stream of news stories about privacy leakage through major data breaches, users are wary of using personal data in public or private ML tasks with good reasons [12].

There are some related surveys on security and privacy issues in distributed ML. For example, the challenges and opportunities of distributed learning over conventional (centralized) ML were discussed in [13], [14], which elaborated on limited privacy and security issues. In [15], [16], the authors focused on the adversarial models related to private information leakage and corresponding defensive mechanisms in ML, and the work [17] investigated privacy issues in distributed ML. Moreover, differential privacy (DP) based protection methods were introduced in [18]. In addition, to protect the privacy of the IoT data, the work [19] surveyed the ML-based method to address the privacy issues including scalability, inter-operability, and limitations on resources, such as computation and energy. The works [20]–[22] addressed security and privacy issues in FL, together with related solutions. The summary of the related surveys on security and privacy issues in ML is listed in Table I.

Different from the above-mentioned surveys, in this work,

- We first give a clear and fresh definition of distributed learning, and develop the distributed learning framework in four levels in terms of sharing different information, namely sharing data, sharing model, sharing knowledge, and sharing results.
- We then provide an extensive overview of the current state-of-the-art related to the attacks and defensive mechanisms on the privacy and security issues for each level. Real examples are also listed for each level.

Real examples are also listed for each level.

- In addition, learned lessons from each aspect are described, which can indeed help readers to avoid potential mistakes.
- Several research challenges and future directions are further discussed, which can provide insights into the design of advanced learning paradigms.

II. BACKGROUND OF DISTRIBUTED ML AND THE PAPER STRUCTURE

In Section II, we first describe the detailed process that how a machine learning task is executed, and then transit the centralized learning to distributed paradigms, and develop a decentralized learning framework. In addition, we provide descriptions of several widely-studied distributed learning frameworks.

A. Background of Machine Learning

Generally speaking, the core idea of ML algorithms can be summarized as training the machine to learn rules or patterns underlying some phenomenon using data and then making decisions or inferences based on new data using the learned rules or patterns. Many ML algorithms fall into the category of pattern recognition (PR), including face recognition, voice recognition, character recognition, and so on [25]. Since humans cannot easily program machines to follow all detailed rules and judgments, ML can be used to help machines learn hidden and even implied rules by themselves. This process is described simply as follows.

Suppose we are going to train a machine to classify whether a fruit is an apple or a banana (a classification task). We first collect some samples that can be labeled and learned by the machine (dataset). So some apples and bananas from this dataset along with their features, including shape, color, weight, size, and so on are recorded. Now, a labeled fruit (apple or banana) with a set of ground-truth features together builds up a sample, and these labeled samples constitute the training dataset. The goal of this ML task is to make the machine learn features from the training dataset and output good predictions given new samples without labels (test dataset). This learning process can be expressed as fitting a function that takes the features as inputs and outputs a value that is as close as possible to the true label. Fig. 2 illustrates the procedure of ML with four main steps listed as follows:

- **Data collection.** The quantity and quality of the collected data dictate how accurate the model is, and the dataset can be divided into training, validation, and test dataset [26].
- **Model training.** For different ML tasks, an appropriate model should be chosen wisely first. Then, the training dataset with the right labels is fed as inputs to the model to start training.
- **Knowledge extraction.** During training, features of the input samples are extracted by some metrics or combinations of metrics (e.g. linear or nonlinear combinations),

TABLE I
EXISTING SURVEYS ON PRIVATE AND SECURE MACHINE LEARNING

Related Survey	Topic	Key contributions
[20]	Privacy preserving in federated learning for IoT data	This work mainly focused on the survey on the use of federated learning for private data analysis in IoT, i.e., highly skewed non-IID data with high temporal variability, to address privacy concerns, bandwidth limitations and power/compute limitations.
[19]	Machine learning-based solutions to protect privacy in IoT	This work surveyed the works that leverage machine learning as a strategy to address the privacy issues of IoT including scalability, inter-operability, and resource limitation such as computation and energy.
[23]	Data Security Issues in Deep Learning	This survey investigated the potential threats of deep learning with respect to black and white box attacks and presented related countermeasures on offense and defense.
[18]	Differentially private machine learning	This survey investigated the existing differentially private machine learning technologies and categorized them as the Laplace/Gaussian/exponential mechanism and the output/objective perturbation mechanism
[13], [14]	Machine learning in distributed systems	These articles provided an overview by outlining the challenges and opportunities of distributed machine learning over conventional (centralized) machine learning, and discussing available techniques.
[20]–[22]	Attacks and defensive strategies on federated deep learning	These works investigated existing vulnerabilities of FL and subsequently provided a literature study of defensive strategies and algorithms for FL aimed to overcome these attacks.
[15], [16], [24]	Privacy in machine learning	These surveys focused on machine learning and algorithms related to private information leakage and corresponding defensive mechanisms.
[17]	Privacy in distributed machine learning	This work focused on the privacy leakage issues in distributed learning and studied benefits, limitations, and trade-offs for defensive algorithms.
Our paper	Privacy and Security in distributed learning	Our work is different from the above survey articles in the following aspects: 1. our work first develops a distributed framework into four levels; 2. the state-of-the-art on the private and secure issues in each level are investigated and summarized; 3. the characteristics of the adversary at each level are further discussed.

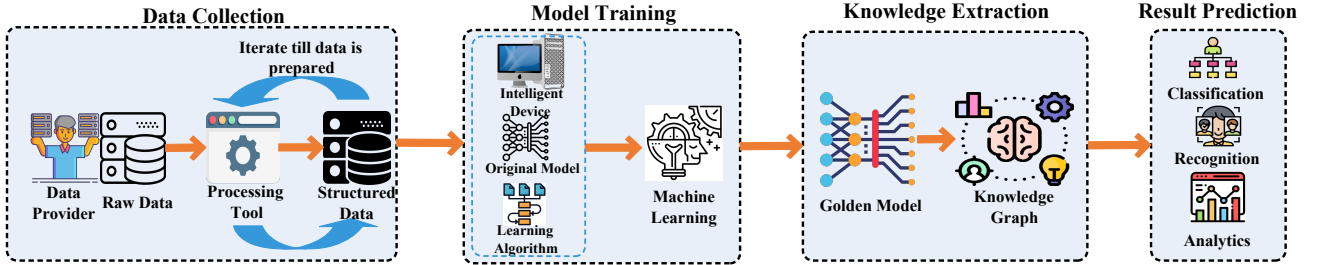


Fig. 2. The process of machine learning in four key steps: data collection, model training, knowledge extraction, and result prediction.

and this knowledge helps the model updates its weights in structures.

- **Result prediction.** The test dataset which has been withheld from the model is used and outputs the prediction results, such as labels, values, vectors (e.g., generative time series), and matrices (e.g., generative images).

B. Background of Distributed Machine Learning

Distributed ML systems and algorithms have been extensively studied in recent years to scale up ML in the presence of big data. Existing work focuses either on the theoretical convergence speed of proposed algorithms or on the practical system aspects to reduce the overall model training time [27]. Bulk synchronous parallel algorithms (BSP) [28] are among the first distributed ML algorithms. Due to the hash constraints on the computation and communication procedures, these schemes share a convergence speed that is similar to traditional synchronous and centralized gradient-like algorithms. The Stale synchronous parallel (SSP) algorithm [29] is a more practical alternative that abandons strict iteration barriers, and

allows the workers to be off synchrony up to a certain bounded delay. The convergence results have been developed for both gradient descent and stochastic gradient descent (SGD) [29]–[31] as well as proximal gradient methods [32] under different assumptions of loss functions. In fact, SSP has become central to various types of currently distributed parameter server architectures [33]–[36]. Depending on how the workload is partitioned [27], distributed ML systems can be categorized into four levels:

- **Level 0: sharing data.** After collecting and pre-processing data locally, each UE will upload its private/anonymized data to a central server, and then the server will use this aggregated data to complete the learning task.
- **Level 1: sharing model.** Different from uploading data directly, each UE can train a local ML model using its own data and **share** the trained model with the server. Then the server will aggregate the collected model and re-transmit the global model to UEs for the next round of learning.

- **Level 2: sharing knowledge.** Different from sharing ML models, the extracted knowledge from training local data, such as the relationship between different attributes, is further shared.
- **Level 3: sharing result.** The task training is completely processed locally, and each UE only shares ML results/outputs to the central server.

The detailed framework of the four-level distributed ML is illustrated in Fig. 3, which is composed of a local and global plane. In the local plane, different information, i.e., data or models, are processed and generated in local devices, and then transmitted to a centralized server for aggregation. Four levels of the proposed distributed learning framework are described in detail, i.e., sharing data, sharing models, sharing knowledge, and sharing results, which are exemplified by representative ML techniques.

C. Existing Distributed Learning Frameworks

In this subsection, we will introduce some popular distributed learning models in the literature, which includes federated learning, split learning, SGD-based collaborative learning, and multi-agent reinforcement learning.

1) *Federated Learning*: FL is a collaborative ML technique [37]–[39] developed by Google, which allows decoupling of data provision at UEs, and machine learning model aggregation, such as network parameters of deep learning, at a centralized server. A structure of FL is plotted in Fig. 4. The purpose of FL is to cooperatively learn a global model without directly sharing data. In particular, FL has distinct privacy advantages compared to data center training on a dataset. At a server, holding even an anonymized dataset can put client privacy at risk via linkage to other datasets. In contrast, the information transmitted for FL consists of minimal updates to improve a particular ML model. The updates can be ephemeral, and will not contain more information than the raw training data (by the data processing inequality). Further, the source of the updates is not needed by the aggregation algorithm, and so updates can be transmitted without identifying metadata over a mixed network such as Tor [40] or via a trusted third party. General categories are distributed horizontal FL, where clients have different sample spaces with the same feature space, and share models during aggregation, distributed vertical FL with the same sample space and different feature spaces, sharing models or knowledge to the central server, and distributed transfer learning with various sample and feature spaces when uploading model or knowledge in aggregation [41].

However, although the data is not explicitly shared in the original format, it is still possible for adversaries to reconstruct the raw data approximately, especially when the architecture and parameters are not completely protected. In addition, FL can expose intermediate results such as parameter updates from an optimization algorithm like SGD, and the transmission of these gradients may actually leak private information when exposed together with a data structure such as image pixels. Furthermore, the well-designed attacks such as inference attack (stealing membership information) [42]–[44], and poisoning attack (polluting the quality of datasets or parameter models) [45] may induce further security issues.

2) *Split Learning*: Split learning, as a type of distributed deep learning [17], [47]–[49], has another name of split neural network (SplitNN). Similar to FL, split learning is effective when data uploading is not available because of privacy and legal restrictions. In the SplitNN, each participant first trains a NN until a predefined layer, called the cut layer, and then transmits the output of the cut layer to the server. Upon receiving the outputs, a central server will continue training the rest layers. Then, the loss function value is calculated and back-propagated to the participant. When receiving the feedback, the participant continues the back-propagation until the network finishes training. In Fig. 5, we show a combination of FL and split learning, where the logits are shared and aggregated at a centralized server.

The computational and communication costs on the client side are reduced in split learning because part of the network is processed locally. In addition, instead of transmitting the raw data, the activation function of the cut layer is uploaded to the server, which has a relatively smaller size. Some experimental results show that split learning has higher performances and fewer costs than FL over figure classification tasks, i.e., CIFAR-100 datasets, using Resnet-50 architectures for hundreds of clients-based setups [47]. However, it needs further explanations on how split learning works and makes decisions, which is linked to the trust of distributed networks, especially in the health area [47].

3) *Large Batch Synchronous SGD (LBS-SGD)*: The difference between the large batch synchronous SGD-based collaborative learning and FL lies in that the updates in LBS-SGD are processed on each batch of training data, and multiple epochs of local training are required before uploading in FL. In LBS-SGD, model parallelism and data parallelism are two common ways to support updating, such as distributed large mini-batch SGD [50], distributed synchronous SGD with backups [17], [51], and selective SGD [52]. In [52], each participant is asked to choose a part of the models to update at each epoch and share them asynchronously with others. The work [50] considered synchronous SGDs by dividing local epochs into mini-batches over multiple clients and model aggregations. While the aggregated updates were performed synchronously in [50] that the aggregator will wait for all clients, the straggler may slow down the learning, and a synchronous optimization with backup participants has been provided in [51].

4) *Multi-Agent Reinforcement Learning*: Reinforcement learning (RL) is trial-and-error learning by interacting directly with the environment, training according to the feedback, and finally achieving the designed goal. Specifically, RL defines a decision maker as an agent and the interaction with the environment, where three essential elements: the state, action, and reward, are used to describe the interaction. For each interaction, the client arrives at a certain state and processes a corresponding action, and then obtains feedback that is used to alter the current state to the next state. However, a single RL framework has no capability to address complex real-world problems, and thus, a multi-agent reinforcement learning system (MARL) has attracted increasing attention. Within a MARL, agents will cooperate with each other and observe the complex environment in a more comprehensive

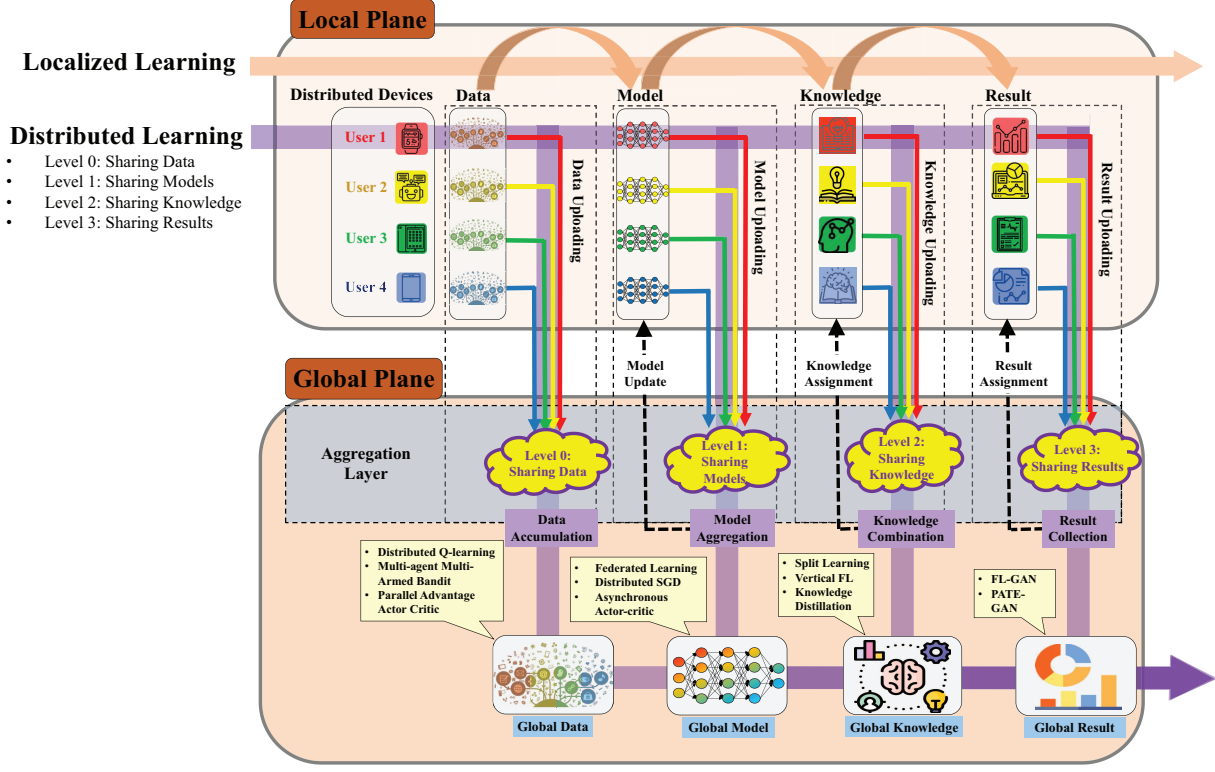


Fig. 3. The framework of distributed learning, which is composed of a local and global plane. In the local plane, different information, i.e., data or models, are processed and generated in local devices, and then transmitted to a centralized server for aggregation. Four levels of the proposed distributed learning framework are described in detail, i.e., sharing data, sharing models, sharing knowledge, and sharing results, which are exemplified by representative ML techniques.

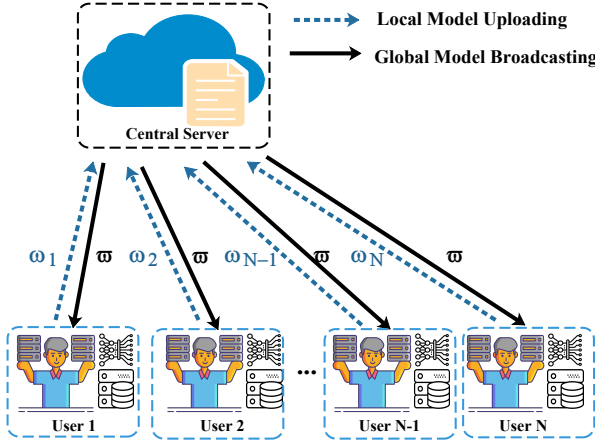


Fig. 4. The structure of federated learning, where users train a ML model using their local data and share the models to a centralized server.

way. For example, as shown in Fig. 6, a three-agent reinforcement learning system, where actions and rewards are shared between different users, is provided. By absorbing the learning experiences from the user-self and other participants, a faster convergence rate with better performance is always achieved. However, compared to the single-agent setting, controlling multiple agents poses several additional challenges, such as

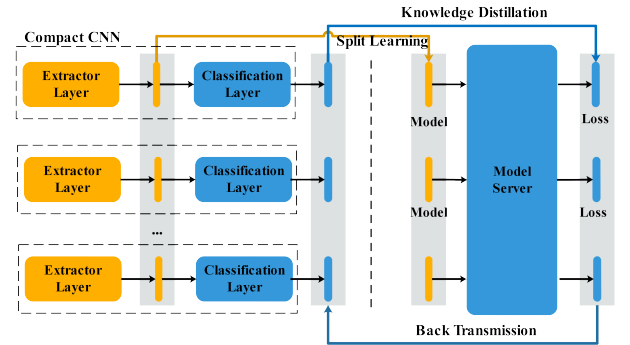


Fig. 5. A reformulation of FL with assisted by the split learning and knowledge distillation [46].

the heterogeneity of participants, the design of achieved goals, and the more serious malicious client problem. Although a number of methods have been proposed to address these challenges, e.g., approximate actor-critic [53] and lenient-DQN, limitations like nonseasonal communication among agents and privacy leakage prevent the rapid development of MARL and existing methods cannot be extended to large-scale multi-agent scenarios.

Following the discussed background of distributed ML, we present the structure of this survey work in Fig. 7. The rest of the paper is structured as follows. In Section III, privacy

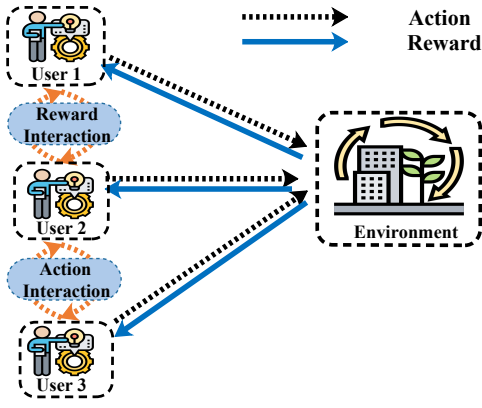


Fig. 6. A framework of multi-agent reinforcement learning, where multiple users communicate and interact to change information, and also process actions to obtain feedback from the environment.

and security issues are discussed and several robust protection methods are provided in Section IV. Then, in Section V, we survey the attacks and defenses in various paradigms in distributed ML. Several research challenges and future directions are shown in Section VI. Finally, conclusions are drawn in Section VII. In addition, a list of important abbreviations is provided in Table II.

III. PRIVACY AND SECURITY RISKS IN DISTRIBUTED ML

In Section III, we will introduce the potential risks of privacy and security, which are measured by factors including threat models, adversarial models, and attack methods.

A. Threat Models

1) *Malicious/Curious Participant*: Participants in distributed ML can be malicious or curious. For example, a car insurance company with limited user attributes might want to improve its risk evaluation model by incorporating more attributes from other businesses, e.g., a bank, a taxation office, etc. The role of the other participants is simply to provide additional feature information without directly disclosing their data to other participants, and in return, obtain financial and/or reputation rewards. However, the competitors may be disguised as collaborators, and then damage the training process or steal the ML model.

2) *External Attackers*: In terms of exchanged information, eavesdropping, modification or deletion can occur during communication in distributed ML as well. We can notice that the exchanged information usually contains the updated direction and extracted features from private data, and thus it is crucial to ensure its correctness, especially for the client-server framework. An external attacker may control the final output by modifying or deleting the exchanged information in the communication. In addition, via eavesdropping on the extracted features from private data, an external attacker can further infer sensitive information [54].

B. Adversarial Models

In this subsection, we will discuss adversarial goals related to leaking information from the training data or destroying models during learning.

1) Access:

- **White-Box**: The adversary is assumed to acknowledge certain information about the training data or the learning model, e.g., model parameters, network structures, or part of/the whole training dataset.
- **Black-Box**: The adversary does not have any knowledge about the ML model, but the adversary can further explore the model by injecting some designed inputs and observing related outputs [55].

2) *Training v.s. inference*: The second factor is the place where the attack happens:

- **Training Stage**: The adversary attempts to learn the model by accessing a part or all of the training data, and creating a substitute model, i.e., a shadow model.
- **Inference Stage**: The adversary observes the outputs from the learning and sums up the model characteristics [54].

3) *Passive vs. Active*: A third factor is to distinguish between passive and active attacks.

- **Passive attack**: The adversary can passively observe and obtain the updates but change nothing during the training process.
- **Active attack**: The adversary actively performs and adjusts the learning operation. For example, the adversary can upload unreasonable parameters to degrade the aggregate model in FL [56].

C. Attack Methods

In this subsection, several attack methods are investigated as follows.

1) *Poisoning Attack*: The goal of a poisoning attack is to degrade the model quality, which misleads the learning to an incorrect direction by carefully crafting poisoning samples during training, also called adversarial examples [57]. In the black-box attack, the attacker can only inject a relatively small amount of crafted/poisoned data into the training model, where the amount and the undiscovered capability of these poisoning data are two basic metrics to estimate the attacking performance. For example, the authors in [58] have first investigated poisoning attacks against linear regression models and proposed a fast optimization algorithm with limited crafting samples to perturb outputs. Further, Suciu *et al.* have investigated the minimum information required by the attacker to achieve various attacking goals [59]. In the white-box attack, the adversaries have full knowledge of the training model and can take advantage of it to reconstruct a powerful poisoning attack. For example, Yuan *et al.* in [60] have proposed a white-box attack with perfect knowledge under different goals. Although the mentioned method might be unrealistic in practical settings, it can achieve almost five times than the black-box attack in success rate.

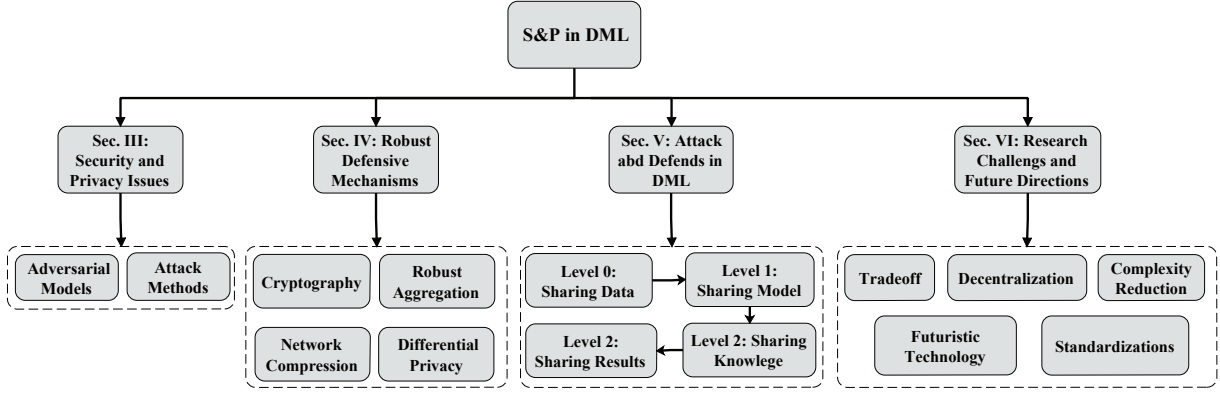


Fig. 7. The structure of the survey with key compositions.

TABLE II
LIST OF IMPORTANT ABBREVIATIONS.

Abbr.	Definition	Abbr.	Definition	Abbr.	Definition
ML	Machine Learning	DL	Deep Learning	RL	Reinforcement Learning
DQN	Deep Q-Learning	AC	Actor-Critic	A3C	Asynchronous Advantage Actor-Critic
TRPO	Trust Region Policy Optimization	PG	Policy Gradient	PPO	Proximal Policy Optimization
DP	Differential Privacy	HE	Homomorphic Encryption	SMC	Secure Multiparty Computation
SGD	Stochastic Gradient Descent	FL	Federated Learning	NN	Neural Network

2) *Evasion Attack*: An evasion attack often happens in the prediction process, which aims to mislead the outputs. In detail, the evasion attack is to change real data from one category to a determined or random one and destroy the integrity of the original dataset. From a black-box attack angle, the adversary only knows the type of the training dataset and observes the outputs. Based on this assumption, the authors in [61] have realized it in the speech recognition system. The generated adversarial samples achieve a 91.67% successful rate on moving one data from one category to another. While in the white-box attack, the adversary is able to acknowledge more useful information, such as the network structure and the type of training samples, rather than the predictive interface. For example, Kevin Eykholt *et al.* in [62] has shown the weakness for DNNs when random noises are added to the inputs and an advanced robust physical perturbations-based method has been proposed.

3) *Model Inversion Attack*: The model inversion attack proposed in [63] works in a black-box fashion, and the adversary only knows the input and can observe the corresponding outputs, which is used to detect correlations between uncertain inputs and respective outputs. A follow-up work has presented a combination with a black-and-white box attack [43]. The proposed attack aims to predict the highest probability of one input for a given label, which the adversary is able to reconstruct the input for a known label, i.e., a figure from a specific class. However, the proposed model inversion attack only works in linear models for most cases, and a major weakness is that the complexity grows exponentially with the input size since it relies on searching all linear combinations by brute force.

4) *Membership Inference Attack*: The membership inference attack (MIA) is mainly focused on [privacy attacks](#). A

previous attack targeting distributed recommend systems [64] intended to infer which input will lead to a change in the output by observing temporal patterns from the learning model. In [54], Shokri *et al.* have investigated the differences between the models to infer whether an input exists in the training dataset for the supervised model. In particular, a shadow model analogs as a similar structure to the targeted model in a black-box fashion. Following [54], Song *et al.* in [65] attempted to record the training data with black-box access. Then, the authors in [66] have exploited the knowledge of learning models to hide the Markov model and attack support vector machine in classification tasks. Also, related works [44], [67], [68] presented inference attacks against distributed deep learning [37], [52]. In particular, Aono *et al.* [67] aimed to attack the privacy-preserving learning framework proposed in [52], and revealed that partial data samples can be revealed by an honest-but-curious server. However, the operation that the single-point batch size limits its effectiveness. Also, a white-box attack against [52] has been proposed in [44], which used generative adversarial networks (GAN) to produce similar samples with a targeted training dataset, however, the proposed algorithm lost effectiveness in the black-box access. Finally, Truex *et al.* in [69] has shown that the MIA is usually data-driven, and Melis *et al.* in [68] have demonstrated the way that a malicious participant infers sensitive properties in distributed learning. Other MIAs focused on genomic research studies [70], [71], in which the attack is designed to infer the presence of specific information of individuals within an aggregated genomic dataset [71], locations [72], and noisy statistics in general [73].

5) *Model and Functionality Stealing*:

- **Model Extraction**. The aim of model extraction is first proposed in [74], in which they proposed to infer the

parameters from a trained classifier with a black-box fashion; however, it only works when the adversary has access to the predictions, i.e., the probabilities for each class in a classification task. In follow-up works, other researchers went a step further to perform hyper-parameter stealing [75], which are external configurations. These values cannot be estimated by data samples, architecture extraction [76] that infers the deep model structures as well as the updating tools (e.g., SGD or alternating direction method of multipliers (ADMM)), etc.

- **Functionality Extraction.** The concept of functionality extraction is, rather than stealing the model, to create knock-off models. Orekondy *et al.* [77] have processed this attack only based on design inputs and relative outputs to observe correlation from ML as a service (MLaaS) queries. In particular, the adversary uses the input-output pairs, e.g., image-prediction pairs in a figure classification task, to train a knock-off model, and compares it with one of the victims for the same task. In addition, the authors in [55] have trained a shadow model to replace a DNN which directly uses inputs generated by the attacker and labeled by the attacking DNN.

D. Section Summary

To sum up, the attack target can be regarded as a clue to distinguish the privacy and security risks from the adversary aspect. A common aim for the privacy attack is to infer a membership of participants without degrading the learning performance, i.e., membership inference attack, and model and functionality stealing, while malicious clients usually aim to destroy the integrity of the learning system, i.e., model poisoning, evasion, and inversion attack.

IV. ROBUST DEFENSIVE MECHANISMS

In Section IV, we will present an overview of several robust defensive mechanisms that include cryptography, robust aggregation, network compression, and differential privacy to reduce information leakage and address security issues.

A. Cryptography

Cryptography is a vital part of distributed ML as it has the ability to support confidential secure computing scenarios. There are a vast of research algorithms and prototypes in literature, which allow participants to obtain learning outputs without uploading their raw data to the server. For instance, in the supervised ML task, secure multi-party computation (SMPC) and homomorphic encryption (HE) based privacy-enhancing tools have been proposed to enable secure computing. Typical examples are, neural networks [78]–[80], matrix factorization [81], linear regressions [82], decision trees [83], and linear classifiers [84], [85].

Specifically, SMPC allows two or more participants to jointly complete a ML task over the shared data without revealing it to others. Popular SMC prototypes are usually developed for two parties, such as [80], [82], [86], [87] designed for distributed ML tasks. For more than two parties,

algorithms based on three-party communication have been provided in [88]–[91], which all rely on the majority of semi-honest or honest participants. For example, Bonawitz *et al.* in [78] has proposed a mixture of several communicating schemes to enable secure computing of participants in FL by blurring the aggregation from the server.

Regard to HE, it mainly uses the encryption and decryption protocol to transform the original message by certain mathematical operations, and there are three common forms for HE: 1) Partially Homomorphic Encryption (PHE) supports one type of mathematical operation; 2) Somewhat Homomorphic Encryption (SWHE) that uses a number of mathematical operations for limited use-cases; 3) Fully Homomorphic Encryption (FHE) supports unlimited numbers of mathematical operations with no other limits [92]. For example, Phong *et al.* in [67] have developed a novel homomorphic scheme based on additive operations for FL with no performance degradation [67]. Other distributed learning strategies, such as [93], [94] used HE to encrypt data and the central server can train a learning model based on the encrypted one. However, the drawbacks of HE are obvious. First, it is usually hard or even impractical to implement HE since this will generate a huge computation overhead [87], [95], [96]. Second, with the increasing number of homomorphic operations, the size of the encrypted models grows exponentially, especially in the SWHE [95], which usually largely surpasses the original model. Third, extra communications between the client and server are required to facilitate key-sharing protocols, which will increase communication costs.

B. Robust Aggregation

The robust aggregation protection methods are used designed for distributed ML that a server needs to aggregate something from clients. To prevent malicious clients, or a group of collusive malicious clients, such as the Byzantine attack in FL [97], the authors in [98] have proposed Krum, a robust aggregation scheme. By minimizing the sum of squared Euclidean distances over the aggregated models, Krum can effectively recognize and remove these outliers. Several follow-ups [99]–[101] aimed to recognize malicious clients. In addition, Chang *et al.* [102] have developed a knowledge-sharing-based algorithm to preserve privacy. The proposed Cronus algorithm relies on a public dataset that is available to all clients. Instead of transmitting parameters, clients will upload the predicted results from this public dataset, and a mean estimation algorithm [103] was used to aggregate these high dimensional label samples. Although Cronus has been proven to defend against basic model poisoning attacks with an acceptable performance loss, sharing labels will lead to privacy leakage to a certain extent.

C. Network Compression

The main purpose of compressing the network is to reduce information transmission, which saves communication resources and accelerates learning. As well, it can also reduce the information exposed to the adversary. Typical methods include quantization [104]–[106], network sparsification [107], [108],

knowledge distillation [109], [110], network pruning [111], [112] and Sketch [113]–[115]. Specifically, an initial work [52] provided the ideal to transmit a subset of all gradients in distributed SGD, and based on it, the authors in [116] have proposed a novel gradient subset scheme that uploads sparse and chosen gradients can improve the prediction accuracy in the non-independent and identically distributed (non-IID) settings. However, as the gradients keep their own form, recent works [42], [117] shown that such methods cannot prevent a specific adversary from inferring available information from these frameworks [42], [117].

Another approach is using lossy compression techniques to decrease the transmitted bits, and it may facilitate certain forms of information security. The authors in [118] quantized the updates using the low-precision quantizer proposed in [104] and provided a smooth tradeoff between compression rate and the convergence performance in convex and non-convex settings. In [119], a count Sketch method with momentum and error accumulation was provided for FL while achieving a high compression rate with good convergence. On the basis of it, the authors in [115] have proved such a quantization method can provide a certain differential privacy guarantee. Moreover, a Sketch-based method was proposed in [114], which sorts gradient values into buckets and encodes them with bucket indexes. In addition, a stochastic-sign-based gradient compressor was used and analyzed to enable communication efficiency [120], and an auto-encoder compressor was proposed in [121] in which the auto-encoder is trained based on dummy-gradients, and the server will release the coded part to clients while keeping the decoder part secretive.

Different from the above methods, a technique called dropout can also be used to defend [122], although it is usually used to prevent overfitting problems in training [123]. By applying dropout, there will be no deterministic outputs (e.g., the updating gradients) on the same training dataset, which can reduce the exploitable attack fact [42].

D. Differential Privacy

Differential privacy (DP) is a standard definition for privacy estimation [124]. A query mechanism is first defined as a property to a dataset, DP-based analytical methods are then extended for ML models on private training data, such as SVM [125], linear regression [126], and deep learning [52], [127]. On neural networks, differentially private stochastic gradient descent [127] is the most famous method that adds random noises on the updating gradients to achieve DP guarantee.

DP sets up a game where the adversary is trying to determine whether a training model has an input D or D' which are adjacent datasets and only differ in one sample. If the adversary can distinguish which dataset (D or D') is used to train by observing the outputs, we can say this training model leaks private information. A formal definition of (ϵ, δ) -DP is expressed as follows:

Definition 1. $((\epsilon, \delta) - DP)$. A randomized mechanism $f : \mathcal{D} \mapsto \mathcal{R}$ offers (ϵ, δ) -DP if for any adjacent input $d, d' \in \mathcal{D}$ and $S \subset \mathcal{R}$,

$$\Pr[f(d) \in S] \leq e^\epsilon \Pr[f(d') \in S] + \delta, \quad (1)$$

where $f(d)$ denotes a random function of d .

To estimate the accumulated privacy budget in multiple learning iterations, the composition theory in [124] shown the effectiveness, and other variants of DP [128], [129] use slightly different formulations with (1), and can achieve a tighter privacy delimitation. Recently, the authors in [130] have derived a lower bound of DP from the adversary perspective, and the Monte Carlo-based method is the first trial of obtaining the privacy level empirically. In addition, the concept of local DP was proposed firstly in [131], [132], and enjoys its popularity gradually.

E. Section Summary

To sum up, general defensive schemes, such as cryptography, robust aggregation, and network compression, can provide thorough protection on security and preserve privacy, where the application of DP is particularly for privacy issues.

V. ATTACKS AND DEFENCES IN VARIOUS LEVELS OF DISTRIBUTED LEARNING

In Section V, we will provide a detailed discussion on the state-of-the-art of attacks and defenses in each level of distributed ML.

A. Level 0: Sharing Data

Data collection plays an important role in various data-governed distributed ML algorithms. However, original data usually contain sensitive information such as medical records, salaries, and locations, and thus a straightforward release of data is not appropriate. Correspondingly, research on protecting the privacy of individuals and the confidentiality of data with an acceptable performance loss has received increasing attention from many fields such as computer science, statistics, economics, and social science.

1) *Threat Models*: Although existing works have proposed a mount of mechanisms to hide identifiers of the raw data, it is also possible for attackers to steal privacy by analyzing hidden features [133]. Moreover, deep neural networks have been proven vulnerable to adversarial examples, which poses security concerns due to the potentially severe consequences [134]. This means that if some adversaries successfully make adversarial examples participate in system training, the training performance will be unacceptable.

2) *Taxonomy of Attacks*: Attacks on data publishing models can be mainly categorized as adversarial examples and feature identification based on their goals. As shown in Table III, we summarize possible attacks as follows.

- **Adversarial examples (data poisoning)**. The work in [134] integrated the momentum term into the iterative process for attacks and generated more transferable adversarial examples by stabilizing update directions and escaping from poor local maxima during the generating iterations. The research on this area is faced with an “arms race” between attacks and defenses, i.e., a defense method proposed to prevent the existing attacks will be soon evaded by new attacks.

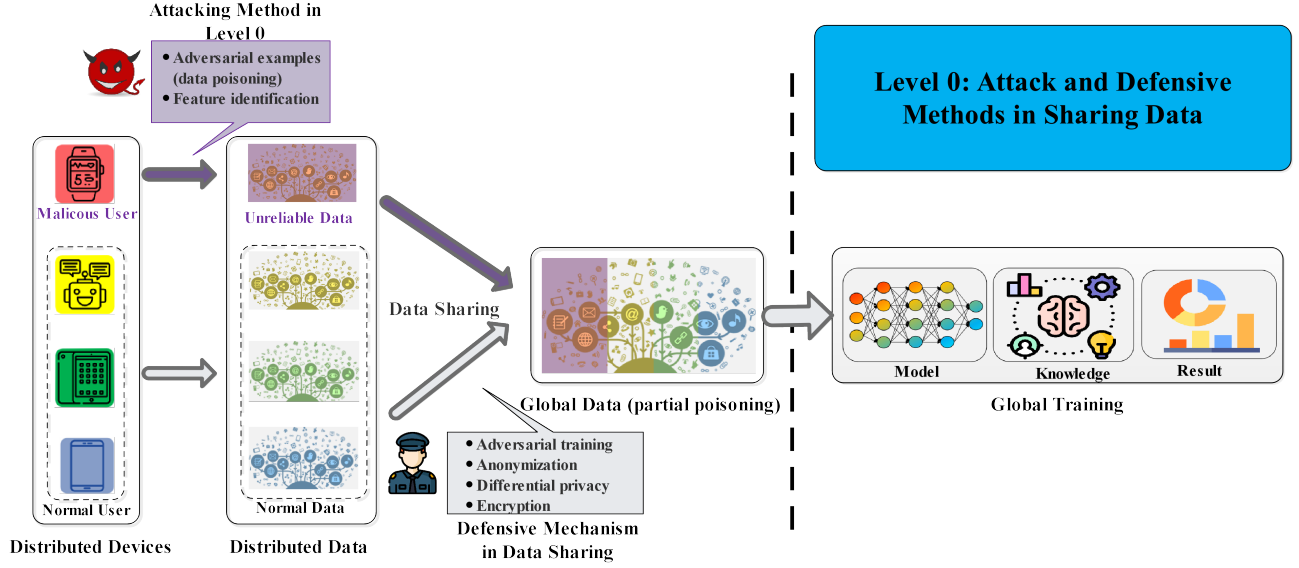


Fig. 8. A breakout figure from Fig. 3: an illustration of privacy and security issues in Level 0 distributed learning with sharing data.

- **Feature identification.** Although many works have proposed efficient methods to process original data in order to preserve sensitive information. Many feature identification attacks are emerging to expose hidden information. As one of the feature identification attacks, structure-based de-anonymization attacks to graph data have been proposed, which aims to de-anonymize the private users in terms of their uniquely distinguishable structural characteristics [135].

3) *Taxonomy of Defences:* Many defensive mechanisms have been designed against aforementioned attacks as shown in Table IV, and we will discuss various defenses as follows.

- **Adversarial training.** Adversarial training is among the most effective techniques to improve model robustness by augmenting training data with adversarial examples. The work in [140] has proposed an adversarial distributional training (ADT) framework, which is formulated as a mini-max optimization problem and improves the model robustness obviously. In this framework, the inner maximization aims to learn an adversarial distribution to characterize the potential adversarial examples around a natural one under an entropic regularizer, and the outer minimization aims to train robust models by minimizing the expected loss over the worst-case adversarial distributions.
- **Anonymization.** An anonymization operation comes in several flavors: generalization, suppression, anatomization, permutation, and perturbation [141], [142]. These techniques aim to remove or hide identifying characteristics from raw data while guaranteeing the data utility. An information-theoretic approach has been formulated and proposed a new multi-objective loss function for training deep auto-encoders [143], which helps to minimize user-identity information as well as data distortion to preserve the application-specific utility. The work in [144] has proposed the conditional identity anonymization gener-

ative adversarial networks (CIA-GAN) model, which can remove the identifying characteristics of faces and bodies while producing high-quality images and videos that can be used for various computer vision tasks, such as detection or tracking. Unlike previous methods, CIA-GAN has full control over the de-identification (anonymization) procedure, ensuring both anonymization as well as diversity. In summary, the choice of anonymization operations has an implication for the search space of anonymous tables and data distortion. The full-domain generalization has the smallest search space with the largest distortion, and the local recording scheme has the largest search space but the least distortion.

- **Dummy.** Existing research methods to protect data privacy mainly focus on the protection of the user's identities through anonymity. User attributes can be classified into identity information, quasi-identifier, and sensitive information. Given an anonymity table, if the attributes in the table have not been properly treated, an adversary may deduce the relationship between the user's identity and sensitive information according to the user's quasi-identifier, such as age and gender. A popular approach for data anonymity is k -anonymity, and any record in a k -anonymized dataset has a maximum probability $1/k$ of being re-identified [145]–[147]. The privacy model l -diversity and t -closeness in [148] further refines the concept of diversity and requires that the distribution of the sensitive values of each equivalent class should be as close as to the overall distribution of the dataset. The common rules for these algorithms are basically to produce dummy records to hide the real ones. In addition, the dummy-based methods also work for location privacy protection. Dummy data along with the true one will be sent to the server from users, which may hide the client's contribution during training [149]. Because the collection is processed on the server, the system performance can

TABLE III
TAXONOMY OF ATTACKS IN LEVEL-0 DISTRIBUTED ML WITH SHARING DATA.

Issue	Ref.	Attacker's knowledge	Learning Model	Effectiveness
Adversarial examples	[134]	White-box, black-box	Inception v2, Inception v3, Inception v4, Resnet v2-152	Attack a white-box model with a near 100% success rate and more than 50% for black-box models
	[136]	White-box, black-box	DQN, A3C, TRPO	Physically interfering with the observations of the victim
	[137]	Black-box	AC	Directly attack actions to achieve the designated purposes
	[138]	Black-box	AC	Taking actions to induce natural observations (environment dynamic) that are adversarial to the victim
Feature identification	[139]	A little bit about an individual subscriber	-	Identify the Netflix records of known users, uncovering users' preferences and other sensitive information

still be guaranteed. As an efficient method to generate realistic datasets, GANs provide an alternative to balance user privacy and training performance. The work in [150] has proposed a novel data augmentation technique based on the combination of real and synthetic heartbeats using GAN to improve the classification of electrocardiogram (ECG) heartbeats of 15 different classes from the MIT-BIH arrhythmia dataset².

- **DP.** As a promising solution, a mechanism is said to be differentially private [124] if the computation result of a dataset is robust to any change of the individual sample. Several differentially private machine learning algorithms [151] have been developed in the community, where a trusted data curator is introduced to gather data from individual owners and honestly runs the private algorithms. Compared to DP, Local DP (LDP) [131], [132] eliminates the need for a trusted data curator and is more suitable for distributed ML. Rappor [152], which applies LDP by Google, is designed to collect the perturbed data samples from multiple data owners. Besides simple counting, a follow-up paper [153] shows that Rappor can also compute other types of statistics such as joint-distribution estimation and association testing. Besides Rappor, an alternative way that achieves DP is to add random noise on the sample value before publishing [131], [154]. To process this method, a numerical sample is always normalized and a categorical one is transformed to the same range by one-hot coding. In addition, the authors in [155] adopted the DP algorithm to handle the privacy concern in a communication problem that each distributed client needs to transmit data to one aggregated center to learn a model. The work [156] has proposed a distributed edge computing which for image classification, where each edge will upload its raw data after coding to latent data to protect privacy.
- **Encryption.** The work in [157] has instantiated a scalable privacy-preserving distributed learning (SPINDLE), an operational distributed system that supports the privacy-preserving training and evaluation of generalized linear

models on distributed datasets. Moreover, it relies on a multiparty HE scheme to execute high-depth computations on encrypted data without significant overhead. The work in [158] has proposed a distributed algorithm for distributed data, where privacy is achieved by the data locality property of the Apache Hadoop architecture and only a limited number of cryptographic operations are required.

- **Others.** The work in [159] has aimed to develop secure, resilient, and distributed ML algorithms under adversarial environments. This work has established a game-theoretic framework to capture the conflicting interests between the adversary and a set of distributed data processing units. The Nash equilibrium of the game has allowed for predicting the outcome of learning algorithms in adversarial environments and enhancing the resilience of the ML through dynamic distributed learning algorithms.

4) Real Examples for Level-0 Distributed ML:

- **RAPPOR.** Randomized aggregatable privacy-preserving ordinal response provides a privacy-preserving way to learn software statistics to better safeguard users security, find bugs, and improve the overall user experience. Building on the concept of randomized response, RAPPOR enables learning statistics about the behavior of users software while guaranteeing client privacy [152]. The guarantees of differential privacy, which are widely accepted as the strongest form of privacy, have almost never been used in practice despite intense academic research. RAPPOR introduces a practical method to achieve those guarantees. In detail, the core of RAPPOR is a randomized response mechanism [161] for a user to answer a yes/no query to the record aggregator. A classic example is to collect statistics about a sensitive group, in which the aggregator asks each individual: "Are you a doctor?" To answer this question, each individual tosses a coin, gives the true answer if it is a head, and a random yes/or answer otherwise. This randomized approach provides plausible deniability to the individuals. Meanwhile, it is shown to satisfy ϵ -LDP, and the strength of privacy protection (i.e., ϵ) can be controlled by using a biased coin. Based on the

²<https://www.physionet.org/content/mitdb/1.0.0/>

TABLE IV
TAXONOMY OF DEFENSES IN LEVEL-0 DISTRIBUTED ML WITH SHARING DATA.

Method	Ref.	Use case	Key idea	Effectiveness
Adversarial training	[140]	Against adversarial examples	Formulating a minimax optimization problem, Parameterizing the adversarial distributions	Improving model security and robustness
Anonymization	[142]	Removing unique identifiers of spatiotemporal trajectory datasets	Clustering the trajectories using a variation k -means algorithm	Enhancing the k -anonymity metric of privacy
	[143]	Motion data	A multi-objective loss function involving an information-theoretic approach	Concealing user's private identity
	[144]	Image and video	Conditional generative adversarial networks	Removing the identifying characteristics of faces and bodies for privacy
Dummy	[145]–[148]	Tabular dataset	Generating fake samples to hide real one	Realizing k -anonymity or similar metrics for privacy
	[150]	Balance MIT-BIH arrhythmia dataset	Generative adversarial networks (GANs)	Generating high quality dummy samples for privacy
DP	[131], [132], [152], [153]	Localized or tabular dataset	Using random response to perturb the value of local data	Achieving LDP for privacy
	[155]	PAC-learning from distributed data	General upper and lower bounds for quantities such as the teaching-dimension	Achieving DP without incurring any additional communication penalty for privacy
	[156]	Communication bandwidth limitation and security concerns of data upload	Training autoencoder, Transmitting latent vectors	Reducing the communications overhead and protecting the data of the end users
Encryption	[160]	Enforcement of access policies, Support of policies updates	Defining their own access policies over user attributes and enforce the policies on the distributed data	Securely manage the data distributed
	[157]	Complete ML workflow by enabling the execution of a cooperative GD	Multiparty homomorphic encryption	Preserving data and model confidentiality with up to $N - 1$ colluding parties
	[158]	Distributed training data, a large volume of the shared data portion.	Data locality property of Apache Hadoop architecture, a limited number of cryptographic operations	Achieving privacy-preservation with an affordable computation overhead
Others	[159]	A learner with a distributed set of nodes	Establishing a game-theoretic framework to capture the conflicting interests between the adversary and data processing units	Obtaining the network topology with a strong relation to the resiliency

collected randomized answers, the aggregator estimates the percentage of users whose true answer is “yes” (resp. “no”). RAPPOR allows the software to send reports that are effectively indistinguishable and are free of any unique identifiers. RAPPOR is currently an available implementation in Chrome, which learns statistics about how unwanted software is hijacking users settings.

- DP in the IOS system. Apple has adopted and further developed local DP to enable Apple to learn about the user community while avoiding learning about individuals [162]. DP perturbs the information shared with random noise before it ever leaves the users device such that Apple can never reproduce the raw data. The power of additive noise that has been added can be reduced without exposing raw data from users by averaging out over large numbers of data points, and meaningful information emerges. DP is utilized as the first step of a system for data analysis that consists of robust privacy protections at every stage. The system is optional and developed

to provide transparency to users. Device identifiers are removed from the data, and it is transmitted to Apple over an encrypted channel. The Apple analysis system ingests the differentially private contributions, dropping IP addresses and other metadata. The final stage is aggregation, where the private records are processed to compute the relevant statistics and the aggregate statistics are then shared with relevant Apple teams. Since both the ingestion and aggregation stages are performed in a restricted access environment, the raw data is not broadly accessible to the public.

5) *Brief Summary:* The guarantee of privacy and security in terms of data sharing models relies on the pre-processing of the raw data, such as perturbation, dummy, anonymization, and encryption. As shown in Fig. 8, data pre-processing happens at the first stage of a ML task, and thus, these pre-processing techniques are usually harmful to the utility of systems or involved extra computations. Therefore, it is more practical to select a proper mechanism to hide sensitive information from

shared data while alleviating the negative influences on the system's utility.

B. Level 1: Sharing Model

In model sharing systems, all distributed nodes need to share their training models with the central server or other participants. Via the interaction between independent data training and local model aggregation, model sharing systems can capture a required learning model over data that resides at the associated nodes.

1) *Threat Models*: Although data is not required to upload in model sharing systems, private information can still be divulged by analyzing uploaded model parameters, e.g., weights trained in deep neural networks. Moreover, adversarial participants may degrade or even destroy the training systems by uploading unreliable models. Attacks can be carried out by the following three aspects.

- **Insiders vs. outsiders.** Insider attacks include those launched by the server and the participants in the model sharing systems. Outsider attacks include those launched by the eavesdroppers in the wireless transmission environment between participants and the server, and by users of the final model when it is deployed as a service. Insider attacks are generally stronger than outsider attacks, as it strictly enhances the capability of the adversary.
- **Semi-honest vs. malicious.** Under the semi-honest setting, adversaries are considered passive or honest but curious. They try to learn the private states of other participants without deviating from the model sharing protocol. The passive adversaries are assumed to only observe the aggregated or averaged gradient, but not the training data or gradient from other honest participants. Under the malicious setting, an active, or malicious adversary tries to learn the private states of honest participants and deviates arbitrarily from the model sharing protocol by modifying, re-playing, or removing messages. This strong adversary model allows the adversary to conduct particularly devastating attacks.
- **Poisoning vs. inference.** Attacks at the poisoning phase attempt to learn, influence, or corrupt the model sharing itself. During the poisoning phase, the attacker can run data poisoning attacks to compromise the integrity of training dataset collection, or launch model poisoning attacks to compromise the integrity of the learning process. The attacker can also launch a range of inference attacks on an individual participant's update or on the aggregation of updates from all participants.

2) *Taxonomy of Attacks*: Attacks to model sharing models can be categorized as poisoning attacks, inference attacks, and model inversion based on their various goals as shown in Table V. We also summarize them as follows.

- **Poisoning attack.** Compromised clients by attackers always have opportunities to poison the global model in model sharing systems, in which local models are continuously updated by clients throughout their deployments. Moreover, the existence of compromised clients may induce further security issues such as bugs in

pre-processing pipelines, noisy training labels, as well as explicit attacks that target training and deployment pipelines [188]. In order to destroy ML models, poisoning attackers may control part of clients and manipulate their outputs sent to the server. For example, the compromised clients can upload noisy and reversed models to the server at each communication round [178], [189], which has the advantage of low complexity to mount attacks. Other attackers may manipulate the outputs of the compromised clients carefully to achieve the evasion of defenses and downgrade the performance of ML models. Furthermore, the authors in [165], [190] have formulated the local model poisoning attack as optimization problems, and then apply this attack against recent Byzantine-robust FL methods. In this way, attackers can improve the success rate of attacks, dominate the cluster and change the judgment boundary of the global model, or make the global model deviate from the right direction. Besides, attackers may hope to craft the ML model to minimize this specific objective function instead of destroying it. Via using multiple local triggers and model-dependent triggers (i.e., generated based on local models of attackers), the collusive attackers can conduct backdoor attacks successfully [191]. Bagdasaryan *et al.* in [163] have developed and evaluated a generic constrain-and-scale technique that incorporates the evasion of defenses into the attacker's loss function during training. The work in [164] has explored the threat of model poisoning attacks on FL initiated by a single, non-colluding malicious client where the adversarial objective is to cause the model to misclassify a set of chosen inputs with high confidence.

- **Inference attack.** The work in [190] has presented a new attack paradigm, in which a malicious opponent may interfere with or backdoor the process of distributed learning by applying limited changes to the uploaded parameters. The work in [163] has proposed a new model-replacement method that demonstrated its efficacy on poisoning models of standard FL tasks. Inferring privacy information about clients for attackers is also possibly achievable in ML models. A generic attacking framework mGAN-AI that incorporates a multi-task GAN has been proposed in [192], which conducted novel discrimination on client identity, achieving attack to clients' privacy, i.e., discriminating a participating party's feature values, such as category, reality, and client identity.
- **Model inversion.** By casting the model inversion task as an optimization problem, which finds the input that maximizes the returned confidence, the work in [43] has recovered recognizable images of people's faces given only their names and accesses to the ML model. In order to identify the presence of an individual's data, an attack model trained by the shadow training technique has been designed, and can successfully distinguish the target model's outputs on members versus non-members of its training dataset [52].

Specifically, in distributed reinforcement learning (DRL)

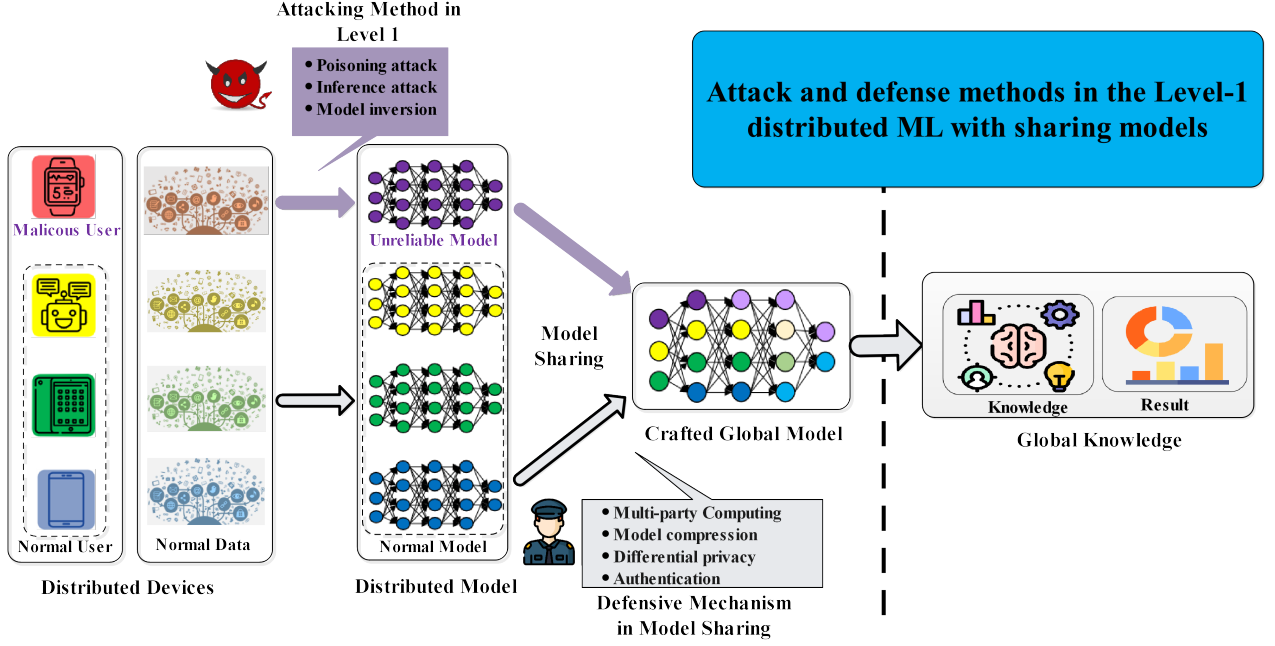


Fig. 9. A breakout figure from Fig. 3: an illustration of privacy and security issues in Level 1 distributed learning with sharing model.

TABLE V
TAXONOMY OF ATTACKS IN LEVEL-1 DISTRIBUTED ML WITH SHARING MODELS.

Issue	Ref.	Attacker's knowledge	Learning Model	Effectiveness
Model poisoning	[163]	Black-box	LSTM, ResNet	Manipulating the RL to achieve the designated purposes
	[164]	Black-box	CNN	Manipulating the RL to achieve the designated purposes
	[165]	White-box, Black-box	LR, CNN	Destroying the system performance
Inference attacks (Snooping attack)	[42]	Black-box	CNN	Inferring certain sensitive characteristics of clients, such as locations and gender, etc.
	[166]	Black-box access to the trained policy, access to the state space, the action space, the initial state distribution and the reward function	DQN, PG, PPO	Inferring certain sensitive characteristics of the training environment transition dynamics, such as dynamics coefficients, environment transition dynamics
	[167]	Black-box	DQN, A2C	Consistently predicting RL agents' future actions with high accuracy
Model inversion	[44]	Black-box	CNN	Reconstructing raw training data
	[168]	Black-box	CNN	Reconstructing the actual training samples without affecting the standard training

systems, there has been literature available on security vulnerabilities. We provide many characteristics of an adversary's capabilities and goals that can be studied as follows. First, we divide attacks based on what components in an MDP the attacker chooses to attack: the agent's observations, actions, and environment (transition) dynamics. Then, we discuss the practical scenarios where attacks happen on these components.

- **Observations.** Existing work on attacking DRL systems with adversarial perturbations focuses on perturbing an agent's observations, i.e., states and rewards, that are communicated between the agent and the environment. This is the most appealing place to start, with seminal results already suggesting that recognition systems are

vulnerable to adversarial examples [136], [193]–[201]. Sandy *et al.* [136] have first shown that adversarial attacks are also effective when targeting neural network policies in RL adversarial examples. Based on this technique, part of the works enhance adversarial examples to attack DRL. To improve the attack efficiency, the strategically-timed attack [193], consuming a small subset of time steps in an episode, has been explored. Via stamping a small percentage of inputs of the policy network with the Trojan trigger and manipulating the associated rewards, the work in [197] has proposed the TrojDRL attack, which can deteriorate drastically the policy network in both targeted and untargeted settings. Another fancy idea

TABLE VI
TAXONOMY OF DEFENSES IN LEVEL-1 DISTRIBUTED ML WITH SHARING MODELS.

Method	Ref.	Description	Key Challenges	Effectiveness
DP	[127], [169]–[174]	Introducing a level of uncertainty into the released model sufficient to mask the contribution of any individual user	Finding a balance between the training performance and privacy level	Low complexity in preserving privacy
Model compression	[115], [175]	Encoding local models before transferring them to the server	Measuring the effect on the privacy and reduce the negative effect on the training performance	Low complexity and high communication efficiency
HE	[117], [176]	Mathematical operations applied on an encrypted message result in the same mathematical operation being applied to the original message	Increasing computation complexity and transmission bits	Strongly effective in security
Secure MPC	[78]	Allowing two or more participants to jointly compute functions over their collective data without disclosing any sensitive information	Lack of a common protocol for various tasks	A lower complexity than HE and a higher security than DP
Statistical analysis	[177], [178]	Detecting and filtering the outliers based on the statistical information, e.g., Euclidean distance and principle component	Destroying the training performance especially in the non-i.i.d. setting	Low complexity to detect outliers
Pretest on Auxiliary Datasets	[179], [180]	Calculating the accuracy score for all local models and reducing the effect of low-quality ones	Performance governed by the quality of auxiliary datasets	Directly detecting malicious users with sensitive datasets
Authentication	[181]	Using trust composition for determining the trust and reputation values for unknown agents	Relying on the trust transfer and vulnerable to the collusion	Low complexity in security
	[182]–[184]	Combining blockchain technology and reaching an agreement by a group of agents	Vulnerable to the 51% attack	Guaranteeing fairness in integrity
Authorization	[185]–[187]	Constructing capability-based access and different agent privilege levels	Formulating corresponding authorization standards for differential privilege levels	Guaranteeing the quality of participants

for a reward-poisoning attack is to design an adaptive disturbing strategy [198], where the infinity norm constraint is adjusted on the DRL agent’s learning process at different time steps. For the theoretical analysis, two standard victims with adversarial observations, i.e., tabular certainty equivalence learner in reinforcement learning and linear quadratic regulator in control have been analyzed in a convex optimization problem on which global optimality and the attack feasibility and attack cost have been provided [196]. In addition, the effectiveness of a universal adversarial attack against DRL interpretations (i.e., UADRLI) has been verified by the theoretical analysis [199], from which the attacker can add the crafted universal perturbation uniformly to the environment states in a maximum number of steps to incur minimal damage. In order to stealthily attack the DRL agents, the work in [200] has injected adversarial samples in a minimal set of critical moments while causing the most severe damage to the agent. Another work in [201] has formulated an optimization framework in a stealthy manner to find an optimal attack for different measures of attack cost and solved it with an offline and online setting.

- **Actions.** Attacks applied on the action space usually

aim to minimize the expected return or lure the agent to a designated state, e.g., the action outputs can be modified by installing some hardware virus in the actuator executing process. This can be realistic in certain robotic control tasks where the control center sends some control signals to the actuator. A vulnerability in the implementation, i.e., the vulnerability in the blue-tooth signal transmission, may allow an attacker to modify that signal [202]. A training policy network to learn the attack has been developed, which treats the environment and the original policy together as a new environment, and views attacks as actions [137]. However, existing works only concentrate on the white-box scenario, i.e., knowing the victim’s learning process and observations, which is not practical and inaccessible to attackers.

- **Environment Dynamics.** The environment (transition) dynamics can be defined as a probability mapping from state-action pairs to states, which is governed by the environmental conditions. For attacks applied on the environment dynamics, an attacker may infer environment dynamics [166] or perturb a DRL system’s environment dynamics to make an agent fail in a specific way [137], [138], [201], [203]. In the autonomous driving case, the

attacker can change the material surface characteristic of the road such that the policy trained in one environment will fail in the perturbed environment. In a robot control task, the attacker can change the robot's mass distribution so that the robot may lose balance when executing its original policy because it has not been trained in that case.

Then, we categorize these attacks based on what knowledge the attacker needs. Broadly, this breaks attacks down into the already recognized white-box attacks, where the attacker has full knowledge of the DRL system, and black-box attacks, where the attacker has less or no knowledge.

- **White-Box.** If the adversary attacks the DRL system with the capability of accessing the architecture, weight parameters of the policy and Q networks, and querying the network, we can call it a white-box attack. Clearly, the attacker can formulate an optimization framework for the white-box setting and derive the optimal adversarial perturbation [136], [199]. Moreover, via the theoretical analysis of the attack feasibility and cost, the adversary can further decrease the efficiency and stealth of the learning [137], [196]. However, this setting is inaccessible for the adversary in most practical scenarios.
- **Black-Box.** In general, the trained RL models are kept private to avoid easy attacks by certain secure access control mechanisms. Therefore, the attacker cannot fully acknowledge the weight parameters of the policy network and Q networks, and may or may not have access to query the policy network. In this case, the attacker can train a surrogate policy to imitate the victim policy, and then use a white-box method on the surrogate policy to generate a perturbation and applies that perturbation to the victim policy [137]. The finite difference (FD) method [204] in attacking classification models can be utilized to estimate the gradient on the input observations, and then perform gradient descent to generate perturbations on the input observations [137]. In this black-box setting, the adversary becomes difficult to perturb a DRL system and needs to estimate the victim's information with large computation costs, such as policies and observations.

Based on the adversary's objective, adversarial attacks are divided into two types: poisoning attacks and snooping attacks.

- **Poisoning Attack.** In particular, for poisoning attacks, there are at least two dimensions to potential attacks against learning systems as untargeted attacks [136] and targeted (induction) attacks [194]. In untargeted attacks, attackers focus on the integrity and availability of the DRL system, i.e., minimizing the expected return (cumulative rewards). Specifically, the work [136] has shown existing adversarial example crafting techniques can be used to significantly degrade the test-time performance of trained policies. However, in terms of defensive mechanisms, the attacker may control time steps [200] or solve an optimization framework in a stealthy manner [199]. Another attack of this category aims at maliciously luring an agent to a designated state more than decreasing the cumulative rewards [194]. Via combining a generative

model and a planning algorithm, the generative model predicts the future states and the planning algorithm generates a preferred sequence of actions for luring the agent [193]. Similar to untargeted attacks, by solving an optimization framework in a stealthy manner [201], the attacker can easily succeed in teaching any target policy.

- **Snooping Attack.** Different from poisoning attacks, the attacker only aims to eavesdrop on environment dynamics, the action and reward signals being exchanged between the agent and the environment. If the adversary can train a surrogate DRL model that closely resembles the target agent [166], [167], the desired information can be estimated by this model. Furthermore, the adversary only needs to train a proxy model to maximize reward, and adversarial examples crafted to fool the proxy will also fool the agent [205]. We can note that the snooping attacks can still launch devastating attacks against the target agent by training proxy models on related tasks, and leveraging the transfer-ability of adversarial examples.

3) *Taxonomy of Defences:* Defensive mechanisms found in multiple works of literature are grouped by their underlying defensive strategies as shown in Table VI. We will discuss various defenses in model sharing frameworks as follows.

- **DP.** DP tackles the privacy leakage about the single data change in a dataset when some information from the dataset is publicly available and is widely used due to its strong theoretical guarantees [206]. Common DP mechanisms will add an independent random noise component to access data, i.e., the shared models in this level, to provide privacy. DP preserving distributed learning systems have been studied from various paradigms, such as distributed principal component analysis (PCA) [169], distributed ADMM [170], distributed SGD [127], FL [171], [172] and multi-agent reinforcement learning [173], [174]. In order to provide fine-tuned control over the trade-off between the estimation accuracy and privacy preservation, a distributed privacy-preserving sparse PCA (DPS-PCA) algorithm that generates a min-max optimal sparse PCA estimator under DP constraints has been proposed in [169]. Similarly, for distributed ADMM, distributed SGD, FL, and multi-agent reinforcement learning systems, all related works focus on improving the utility-privacy trade-off via two aspects as follows: a) analyzing the learning performance with a DP constraint and then optimizing system parameters; b) enhancing the DP mechanism by obtaining tighter estimates of the overall privacy loss.
- **Model compression.** Model compression techniques for distributed SGD and FL systems, e.g., sketches, can achieve provable privacy benefits [115], [175]. Therefore, a novel sketch-based framework (DiffSketch) for distributed learning has been proposed, improving absolute test accuracy while offering a certain privacy guarantee and communication compression. Moreover, the work in [175] has presented a family of vector quantization schemes, termed Vector-Quantized Stochastic Gradient Descent (VQSGD), provides an asymptotic reduction in

the communication cost and automatic privacy guarantees.

- **Encryption.** Encryption, e.g., HE [117] and MPC [78], is also adopted to protect user data privacy through parameter exchange under the well-designed mechanism during ML. A novel deep learning system [117], bridging asynchronous SGD and cryptography, has been proposed to protect gradients over the honest-but-curious cloud server, using additively homomorphic encryption, where all gradients are encrypted and stored on the cloud server. To verify whether the cloud server is operating correctly, VerifyNet [176] has been proposed to guarantee the confidentiality of users' local gradients via a double-masking protocol in FL, where the cloud server is required to provide proof of the correctness of its aggregated results to each user.
- **MPC.** The work in [78] has outlined an approach to advancing privacy-preserving ML by leveraging MPC to compute sums of model parameter updates from individual users' devices in a secure manner. The problem of computing a multiparty sum where no party reveals its updates to the aggregator is referred to as secure aggregation. Via encoding local models into multiple secret shares in the first round, and then splitting each share into a public share and a private share, the work in [207] can provide stronger protections for the security and privacy of the training data. MPC integrates the encryption technology and interactive protocols, aiming to make the receiver keep away from sensitive information and obtain the necessary messages [208]–[211].
- **Statistical analysis.** The work in [177] has proposed a robust aggregation rule, called adaptive federated averaging, that detects and discards bad or malicious local model updates based on a hidden Markov model. To tackle adversarial attacks in the FL aggregation process, the work in [178] presented a novel aggregation algorithm with the residual-based re-weighting method, in which the weights for the average of all local models are estimated robustly. Via controlling the global model smoothness based on clipping and smoothing on model parameters, a sample-wise robustness certification FL framework has been proposed, which can train certifiably robust FL models against backdoors [212]. Most of the defenses for FL aim to explore the latent model exception, such as similarities between malicious and benign clients, and then lessen the influence of these exceptional models [213]–[216].
- **Pretest on auxiliary datasets** For detecting poisoned updates in collaborative learning [179], the results of client-side cross-validation were applied for adjusting the weights of the updates when performing aggregation, where each update is evaluated over other clients' local data. The work in [179] considered the existence of unreliable participants and used the auxiliary validation data to compute a utility score for each participant to reduce the impact of these participants. The work in [180] has proposed a novel poisoning defense method in FL, in which the participant whose accuracy is lower than a predefined threshold will be identified as an attacker,

and the corresponding model parameters will be removed from the training procedure in this iteration.

- **Authentication and access control.** The key question in considering security in a MARL consists of increasing the confidence that all parties involved in the system (agents, platforms, and users) will behave correctly, and this can be achieved through the authentication of these parties. The identification of the parties can make up a system and possibly establish an agent-trust relationship. Thus, how to design efficient identity certification mechanisms to uniquely authenticate known and trusted users and agents in the system has drawn heated attention. A domain-independent and reusable MARL infrastructure has been developed in [217], in which the system uses a certification authority (CA) and ensures full cooperation of secured agents and already existing (unsecured) agents. The work in [181] has introduced a method called trust composition, which combines several trust values from different agents. We can note that the trust composition can play a critical role in determining the trust and reputation values of unknown agents since it is impractical for an agent to get complete knowledge about other agents. A work called PTF (Personalized Trust Framework) has been proposed to establish a trust/reputation model for each application with personalized requirements [218]. Naturally, the idea of using blockchain technology to solve security problems in multi-robot systems was discussed in [182]. The work in [182] stated that combining peer-to-peer networks with cryptographic algorithms allows reaching an agreement by a group of agents (with the following recording this agreement in a verifiable manner) without the need for a controlling authority. Thus, blockchain-based innovations can provide a breakthrough in MARL applications. The work in [183] has developed an approach to using decentralized programs based on smart contracts to create secure swarm coordination mechanisms, as well as for identifying and eliminating Byzantine swarm members through collective decision-making. The work in [184] has proposed an approach combining blockchain technology and explainability supporting the decision-making process of MARL, in which blockchain technology offers a decentralized authentication mechanism capable of ensuring trust and reputation management.
- **Authorization and trust model.** Combined with authentication, authorization is used to restrict the actions that an agent can perform in a system, and control the access to resources by these agents. Sensitive information about principals is transferred online even across the Internet and is stored in local and remote machines. Without appropriate protection mechanisms, a potential attacker can easily obtain information about principals without their consent. In the context of authorization mechanisms, the algorithm proposed in [185] is designed to solve the problem of systems that are constantly changing. The main goal is to build a flexible and adaptive security policy management capable to configure itself and reflect the actual needs of the system. According to the authors,

a system is not safe if a security model is developed but never managed afterward. Security of the proposed system in [186] has been further explored in the form of authorization and encryption of the data by introducing an authorization layer between the user and the system that will be responsible for providing access to the legitimate users of the system only. The work in [187] has ensured agent authorization and platform security with capability-based access and different agent privilege levels, in which the agent behavior is modeled with an activity transition graph (ATG) and implemented entirely in JavaScript with a restricted and encapsulated access to the platform API (AgentJS).

4) Real Examples for Level-1 Distributed ML:

- **Electronic Medical Records (EMR)** [219]. The use of information and network technologies in the healthcare field inevitably produces EMR, which is a necessary trend for the modernization of medical records in hospitals. The initial adoption of EMR in clinical practice has vastly improved the efficiency and quality of health care provided by hospitals. Empowered by algorithm technologies and data reconstruction, BaseBit [219] has constructed a robust and comprehensive knowledge base system and has a series of intelligent models with excellent abilities of expression. In various applications centered around electronic medical records, the proposed models effectively improve the abilities such as automatic medical record writing, overall quality control, cost monitoring systems for single diseases, early warning for infectious diseases, prompt for critical illnesses, clinical decision-making assistance for rare diseases, enabling hierarchical diagnosis and treatments.

5) *Brief Summary:* As shown in Fig. 9, although due to the local training process, the raw data of each participant will not be exposed to the curious server or external attackers, defensive mechanisms are also necessary because of the existing possibility of feature inference and data reconstruction from models sharing, in addition to the model poisoning paradigm. Traditional HE and DP are proven beneficial to privacy preservation but lead to low efficiency or damaged utility. Therefore, the quantitative analysis of the relationship between the sensitive feature and the published model is imperative.

C. Level 2: Sharing Knowledge

Recent configurations that rely on knowledge sharing techniques can be summarized as split learning [47], vertical FL [9], and distillation-based FL [223]. Split learning allows multiple clients to hold different modalities of vertically partitioned data and learn partial models up to a certain layer (the so-called cut layer). Then the outputs at the cut layer from all clients are then concatenated and sent to the server that trains the rest of the model. In vertical FL, participants hold the same set of samples but with disjoint features and only one participant owns the labels, which need to combine split NNs and privacy-preserving techniques [224]. Distillation-based FL [46], [223], [225] exchanges model outputs instead of

model parameters, where the communication overhead cannot scale up according to the model size and has been proven to satisfy the DP guarantee.

1) *Threat Models:* In knowledge sharing paradigms, adversarial participants or eavesdroppers still possibly exist. The adversarial participants can be categorized into two kinds: a) honest-but-curious (semi-honest) participants, who do not deviate from the defined learning protocol, but attempt to infer private training data from the legitimately received information; b) malicious participants, who may deviate from the defined learning protocol, and destroy this training task or inject Trojans to the training model.

2) *Taxonomy of Attacks:* Existing attacks on knowledge sharing paradigms can be mainly categorized as label leakage, feature inference, and data reconstruction as shown in Table VII. Then, we discuss existing attacks as follows.

- **Label leakage.** The labels in distributed learning frameworks might be highly sensitive, e.g., whether a person has a certain kind of disease. However, the bottom model structure and the gradient update mechanism of VFL or split learning can be exploited by a malicious participant to gain the power to infer the privately owned labels [226]. Worse still, by abusing the bottom model, he/she can even infer labels beyond the training dataset [227]. The work in [220] first made an attempt at a norm attack that uses the norm of the communicated gradients between the parties, and it can largely reveal the ground-truth labels from participants. The adversary (either clients or servers) can accurately retrieve the private labels by collecting the exchanged gradients and smashed data [228]. Thus, it is necessary to make gradients from samples with different labels similar.
- **Feature inference.** Through analysis, the work in [229], [230] demonstrated that, unless the feature dimension is exceedingly large, it remains feasible, both theoretically and practically, to launch a reconstruction attack with an efficient search-based algorithm that prevails over current feature protection techniques. In this paper, the authors have performed the first systematic study on relation inference attacks to reveal VFL's risk of leaking samples' relations. Specifically, the adversary is assumed to be a semi-honest participant. Then, according to the adversary's knowledge level, the work [230] formulated three kinds of attacks based on different intermediate representations and revealed VFL's risk of leaking samples' relations. Luo *et al.* [192] considered the most stringent setting that the active party (i.e., the adversary) only controls the trained vertical FL model and the model predictions, and then observed that those model predictions will leak a lot of information about features by learning the correlations between the adversary's and the attacking target's features.
- **Data reconstruction.** The work in [221] has provided the leakage analysis framework via three empirical and numerical metrics (distance correlation and dynamic time warping) indicating that the activated outputs after two and more convolutional layers can be used to reconstruct the raw data, i.e., sharing the intermediate activation

TABLE VII
TAXONOMY OF ATTACKS IN LEVEL-2 DISTRIBUTED ML WITH SHARING KNOWLEDGE.

Method	Ref.	Attacker's knowledge	Learning Model	Effectiveness
Label leakage	[220]	Black box	Split learning	Revealing the ground-truth labels from the participants
Feature inference	[192]	Black box	Vertical FL	Inferring the feature values of new samples belong to the passive parties successfully
Data reconstruction	[221]	Black box	Split learning	Activated output after two and three convolutional layers can be used to reconstruct the raw data
	[222]	Black box	Vertical FL	Stealing partial raw training data successfully

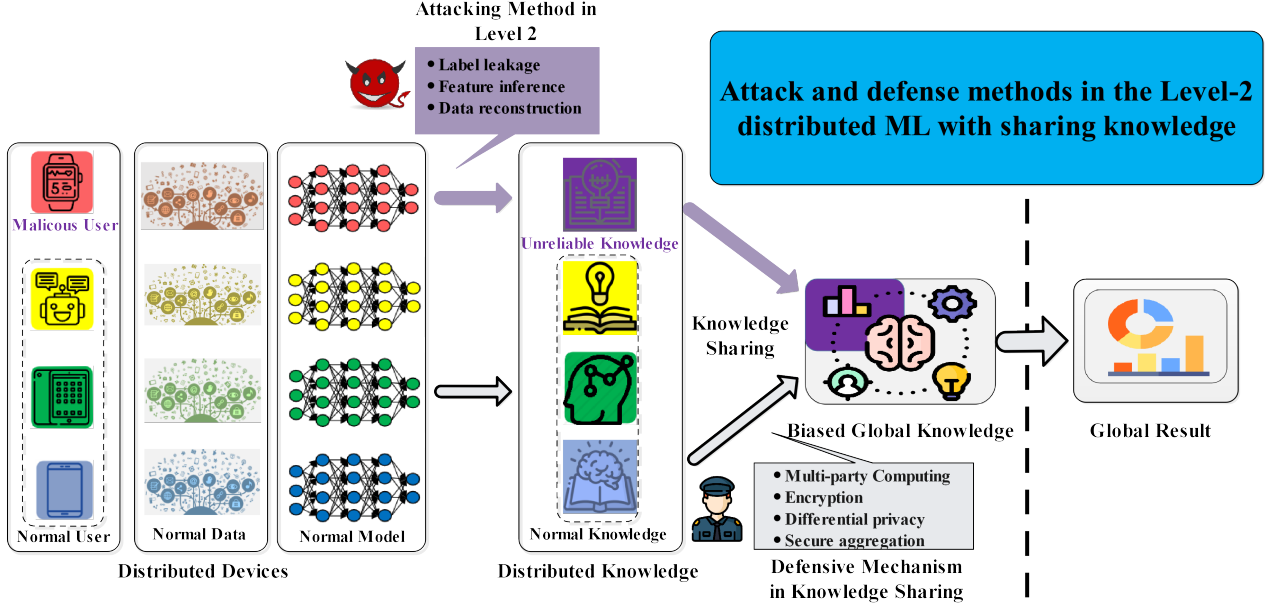


Fig. 10. A breakout figure from Fig. 3: an illustration of privacy and security issues in Level 2 distributed learning with sharing knowledge.

from these layers may result in severe privacy leakage. In vertical FL, two simple yet effective attacks, reverse multiplication attack and reverse sum attack, have been proposed to steal the raw training data of the target participant [222]. Though not completely equivalent to the raw data, these stolen partial orders can be further used to train an alternative model which is as effective as the one trained on the raw data [231].

3) *Taxonomy of Defences:* Defensive mechanisms found in multiple works of literature are grouped by their underlying defensive strategy as shown in Table VIII. Hence, we will discuss various defenses in model sharing frameworks as follows.

- **DP.** The work in [232] has proposed a privacy-preserving protocol for composing a differentially private aggregate classifier using local classifiers from different parties. In order to overcome the effects of the proposed information inference attacks [221], DP has been proven helpful in reducing privacy leakage but leading to a significant drop in model accuracy.
- **MPC.** The work in [233] has proposed a novel solution for privacy-preserving vertical decision tree training and

prediction, termed Pivot, ensuring that no intermediate information is disclosed other than necessary releases (i.e., the final tree model and the prediction output).

- **Encryption.** A novel privacy-preserving architecture has been proposed in [234], which can collaboratively train a deep learning model efficiently while preserving the privacy of each party's data via the HE technique. The work in [234] has explored a lossless privacy-preserving tree-boosting system known as SecureBoost by using the additive HE scheme.
- **Secure aggregation.** The work in [235] has proposed the vertical FederBoost which runs the gradient boosting decision tree (GBDT) training algorithm in exactly the same way as the centralized learning. Via further utilizing packetization and DP, this algorithm can protect the order of samples: participants partition the sorted samples of a feature into buckets, which only reveals the order of the buckets and add differentially private noise to each bucket.
- **Others.** The work in [236] has presented TIPRDC to learn a feature extractor that can hide the private information from the intermediate representations using an

adversarial training process while maximally retaining the original information embedded in the raw data to accomplish unknown learning tasks. In [221], adding more hidden layers to the client side was proven helpful in reducing privacy leakage, but increasing the number of layers seems ineffective with the most highly correlated channels. In order to relieve the negative impact of random perturbation preserving techniques on the learned model's predictive performance, the work in [220] has introduced an improved way to add Gaussian noise by making the expected norm of the positive and negative gradients in a mini-batch equal (un-distinguishable).

4) Real Examples for Level-2 Distributed ML:

- **FATE.** An open-source project, named FATE, provides a secure computing framework to support the federated AI ecosystem [237], led by Webank's AI Department. It can enable big data collaboration without privacy leakage by implementing multiple secure computation protocols, such as DP, HE, and so on. FATE accesses out-of-box usability and excellent operational performance with a modular modeling pipeline, explicit visual interface, and flexible scheduling system [238]. eHi Car Services, a national chain car rental brand, and WeBank jointly announced a deep strategic partnership, announcing that the two sides will carry out multi-scene and multi-dimensional innovation cooperation in car travel, member services, finance and insurance, blockchain technology, and other fields. eHi Car Services uses federal transfer learning, AI face authentication technology, payment technology, and other fin-techs to deeply integrate into the car rental service process for the purpose of optimizing and improving user experience, and combines the car rental scene with the bank's big data risk control system, so as to provide a new way of travel and life for the young and long-term rental customers.

5) *Brief Summary:* As shown in Fig. 10, split learning, vertical FL, and distillation-Based FL are the classical knowledge sharing systems, in which the knowledge can be viewed as the **partial** processing result to meet the requirement of the system learning. It is also challenging for knowledge sharing systems to hide sensitive information from the shared knowledge.

D. Level 3: Sharing Results

We define the sharing results category as follows: there is no interaction or communication during the process of training. The distributed clients only share the training results after the process ends. The history of sharing results can be traced back to ensemble ML over partitioned datasets [239], [240], where a number of base classifiers collectively determine the output for an instance based on a pre-defined aggregation strategy. Ensemble techniques were originally introduced to increase the overall performance of the final classification, but it is also straightforward to utilize them for distributed ML systems [241]. The shared results [242] in distributed learning can be either the final training models, e.g., PATE and multi-agent multi-arm bandits (MAMAB), or the prediction (output) of the models, e.g., crowd-sourcing.

1) *Threat Models:* For the result sharing models, malicious participants may exist and provide false advice or results to hinder the learning performance of other participants or the global model. In addition, curious participants can infer some confidential information from the shared results.

2) *Taxonomy of Attacks:* As stated by da Silva *et al.* [243], the existence of malicious participants is a key concern in agent advises. The work in [244] has proposed the attack model that some of these agents might become self-interested and try to maximize car owners' utility by sending out false information. Based on [244], Hayes *et al.* [245] have investigated attacks in the setting where the adversary is only permitted to access the shared results (such as the generated samples set in GAN), by retraining a local copy of the victim model. In addition, Hilprecht *et al.* [246] have proposed to count the number of generated samples that are inside an ϵ -ball of the query, based on an elaborate design of distance metric. The work in [247] has presented the first taxonomy of membership inference attacks and focused on membership inference attack against deep generative models that reveals information about the training data used for victim models. In spirit to Hilprecht *et al.* [246], this work scored each query by the reconstruction error directly, which does not introduce additional hyper-parameter while achieving superior performance. We further summarize these attacks in Table IX.

3) *Taxonomy of Defences:* In results sharing paradigms, Tab. X summarizes the use case, key idea, and effectiveness for existing attacks. Moreover, we will discuss various defenses in model sharing frameworks as follows.

- **DP.** The work in [174] has proposed a novel differentially private agent advising approach, which employs the Laplace mechanism to add noise to the rewards used by student agents to select teacher agents. By using the advising approach and the DP technique, this approach can reduce the impact of malicious agents without identifying them and naturally control communication overhead. The work in [248] adopted DP and studied the regret upper and lower bounds for MAB algorithms with a given local DP guarantee. The differentially private PATE framework has been proposed to achieve individual privacy guarantees with provable privacy bounds [250], [251].
- **MPC.** Zhao [249] has proposed to use the teacher-student framework in a more general distributed learning setting. The goal of this work is to address distributed deep learning under DP using the teacher-student paradigm. In the setting, there are a number of distributed entities and one aggregator. Each distributed entity leverages deep learning to train a teacher network on sensitive and labeled training data. The knowledge of the teacher networks is transferred to the student network at the aggregator in a privacy-preserving manner that protects the sensitive data. This transfer results from training non-sensitive and unlabeled data, which also applies secure MPC to securely combine the outputs of local ML for updating.
- **Others.** If an ensemble contains enough models, and each model is trained with disjoint subsets of the training

TABLE VIII
TAXONOMY OF DEFENCES IN LEVEL-2 DISTRIBUTED ML WITH SHARING KNOWLEDGE.

Method	Ref.	Use case	Key idea	Effectiveness
DP	[232]	Deriving aggregate information without revealing information about individual data instances	Differentially private aggregate in a multi-party setting	DP analysis on the perturbed aggregate classifier
	[221]	Against DCM and DTWM attacks in split learning	Laplace mechanism on the split layer activation	Strong DP level ($\epsilon = 1$) works but degrading the classification accuracy
MPC	[233]	Vertical decision tree training, random forest (RF), and gradient boosting decision tree (GBDT)	A hybrid framework of threshold partially HE (TPHE) and MPC	Be independent of any trusted third party against a semi-honest adversary that may compromise $m - 1$ out of m clients
Encryption	[234]	Asymmetrically split learning	Partial HE (PHE), additive noise	Achieving a lossless performance and more than 100 times speedup
	[234]	Vertical tree-boosting system	HE	Revealing no information of each participant and achieving a lossless performance
Secure aggregation	[235]	Vertical GBDT	Lightweight secure aggregation because the whole training relies on the order of the data instead of the values	Achieving the same level of the area under the ROC curve (AUC) with centralized training
Others	[236]	Privacy attributes inferring from extracted features	Adversarial training and neural network based mutual information estimator	First task-independent privacy-respecting data crowdsourcing framework
	[221]	Against DCM and DTWM attacks in split learning	Adding more hidden layers	Preventing privacy leakage with a slight reduction in performance
	[220]	Against norm-based attack	Adding Gaussian noise by making the expected norm of the positive and negative gradients in a mini-batch equal	Preventing label leakage against some extreme scenarios.

TABLE IX
TAXONOMY OF ATTACKS IN LEVEL-3 DISTRIBUTED ML WITH SHARING RESULTS.

Method	Ref.	Attacker's knowledge	Learning Model	Effectiveness
Poisoning attack	[244]	Black box	Street random waypoint (STRAW) mobility	Average speed of vehicles in the network decreases as the percentage of liars increases
Inference attack	[245]	White-box, black-box	GAN	Achieving 100% and 80% successful at membership inferring in white-box and black-box settings, respectively
	[246]	Black-box	GAN, variational autoencoders (VAEs)	Success rates superior to previous work with mild assumptions
	[247]	White-box, partial black-box, black-box	GAN	Consistently outperforms the state-of-the-art models with an increasing number of generated samples

data in a distributed manner, then “any predictions made by most of the models should not be based on any particular part of the training data” [252]. The private aggregation of teacher ensembles (PATE) is based on this idea [10]. In more detail, the ensemble is seen as a set of “teachers” for a new “student” model. The student is linked to the teachers only by their prediction capabilities, and is trained by “querying the teachers about unlabelled examples”. The prediction result is disjointed from the training data through this process. Therefore data privacy can be protected. The privacy budget for PATE is much lower than traditional DP-based ML approaches. But it may not work in many practical scenarios as it relies on an unlabelled public dataset.

4) Real Examples for Level-3 Distributed ML:

- Large-scale online taxicab platforms, such as Uber and DiDi, have revolutionized the way people travel and socialize in cities worldwide and are increasingly becoming essential components of the modern transit infrastructure [253], [254]. The reinforcement learning-based dynamic bipartite graph matching approach has been adopted to assign each worker with one or more tasks to maximize the overall revenue of the platform, where the workers are dynamic while the tasks arrive sequentially. Specifically, for each worker-task pair, the platform can obtain a reward based on value-based reinforcement learning. Then, via some solutions to bipartite graph matching, such as greedy search, the platform can

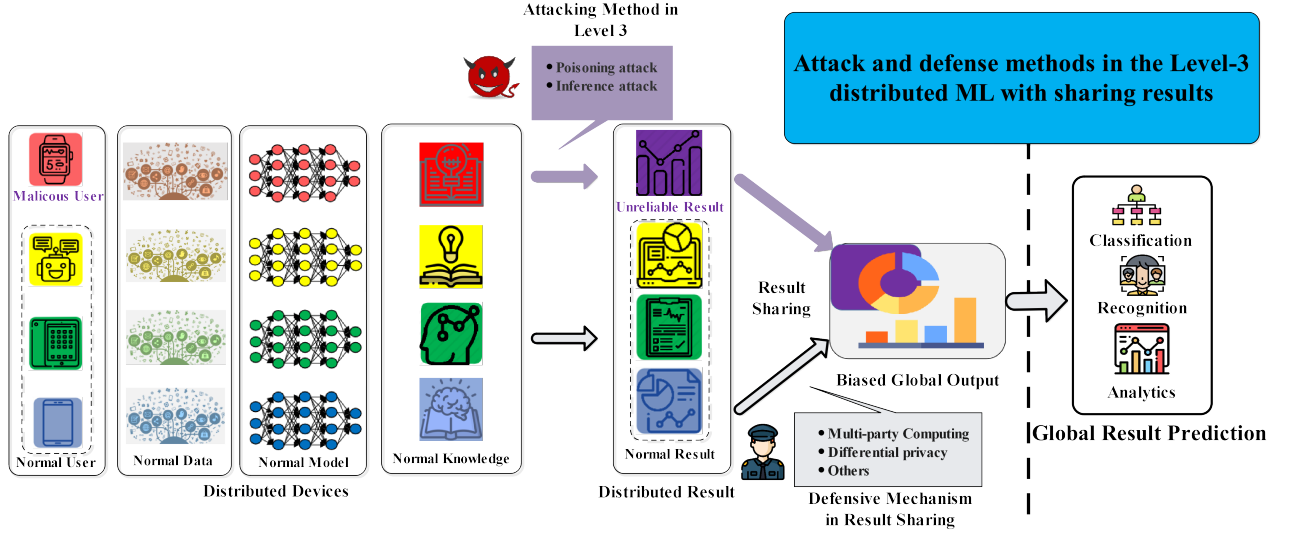


Fig. 11. A breakout figure from Fig. 3: an illustration of privacy and security issues in Level 3 distributed learning with sharing results.

TABLE X
TAXONOMY OF DEFENSES IN LEVEL-3 DISTRIBUTED ML WITH SHARING RESULTS.

Method	Ref.	Use case	Key idea	Effectiveness
DP	[174]	Malicious agent advising	Laplace mechanism	Reducing the impact of malicious agents without identifying them
	[248]	Against inference attacks from any party or eavesdropper	Laplace mechanism, Bernoulli mechanism	Providing regret upper and lower bounds for MAB with local DP
MPC	[249]	PATE	Training non-sensitive and unlabeled data, Securely combining the outputs by MPC	Guarantee data security
Others	[10]	PATE	The student is linked to the teachers only by their prediction capabilities and trained by “querying the teachers about unlabelled examples”	Achieving much lower privacy budget than traditional DP approaches

make near-optimal decisions. However, if the platform can obtain all workers information and its purpose is only aiming to maximize the overall revenue, workers may be out of control. Thus, using DP to achieve fairness may be a solution [255].

5) *Brief Summary*: As shown in Fig. 11, although the results from ML systems are various from the raw data, they are also existing risks of privacy leakage, such as the generated samples from the generator in GAN. Hence, several defensive mechanisms are utilized for preventing privacy leakage and against malicious participants.

E. Relationship among the privacy and security issues in the four levels of distributed ML

From level 0 to level 3, there is no certain law for the privacy and security level, but we may conclude that the forms of data show expose different degrees of information in the considered four levels. For example, compared to the prediction results in level 3, much more information can be extracted from the raw or original data in level 0. Regarding to the protection methods, designing a general mechanism for the four levels is an un-trivial task. For example, the DP-based mechanisms

can be well adopted in level 0 (i.e., local DP [131], [152]), level 1 (i.e., DP in deep learning [127]) and level 3 (i.e., PATE-GAN [10]), but it may lose the effectiveness in level 2 (sharing knowledge).

VI. LESSONS LEARNED

In this section, we summarize the key lessons learned from this survey, which provides an overall view of the current research on security and privacy issues in distributed learning.

A. Lessons Learned from Definitions of Security and Privacy

The public often mixes up the terminologies of “Privacy” and “Security”, which are in fact distinctively different. From the expression of privacy and security in distributed learning, we can learn lessons as follows.

1) *Difference between Security and Privacy*: The concerns of security and privacy issues are different [256]–[258]. On the one hand, security issues refer to unauthorized/malicious access, change, or denial of data or learning models. Such attacks are usually launched by adversaries with expert/full knowledge of the target system. Hence, the fundamental three

goals of security are confidentiality, integrity, and availability [165]. On the other hand, privacy issues generally refer to the unintentional disclosure of personal information. For example, from a side-by-side comparison of a vote registration dataset and an anonymous set of health-care sensor records (e.g., no individuals name and ID), an adversary may have the ability to identify particular individuals and the health conditions of these individuals leaks [68], [259], [260]. This is because attributes such as gender, birth date, and zip code are the same in both datasets.

2) *Connection between Security and Privacy*: Security and privacy go hand-in-hand. Privacy issues can further induce security issues in some scenarios. If an adversary steals the private information of individuals, substantial profit from the information can be easily obtained. For example, when the adversary extracts the health conditions of an important person, he/she can blackmail the victim person by threatening to reveal the information. We know that one can envision an environment that is secure but does not guarantee privacy. Similarly, one can imagine an environment that is private, but it does not guarantee security from outsiders. Security can be achieved without privacy, but privacy cannot be achieved without security. This is because whether the security is weak or vulnerable, it will automatically affect privacy.

B. Lessons Learned from Evaluations of Security and Privacy

The evaluations on security and privacy guide the research directions in this area. In the following, we will provide some lessons by reviewing the state-of-the-art.

1) *Bayes-based Methods*: Privacy leakage can be formalized as a Bayes optimization problem from the aspect of an adversary with different assumptions on the probability distributions of the input data and interactive messages (such as gradients and extracted features). For example, the work in [261] constructed a theoretical framework that can measure the expected risk that an adversary has in the process of reconstructing an input, given the joint probability distribution of inputs and their gradients. This framework can reveal the gradient leakage level by analyzing the Bayes optimal adversary, which minimizes this risk with a specific optimization problem involving the joint distribution. DP constitutes a strong standard for privacy guarantees for algorithms on aggregate databases [124], [127], [172]. It is defined in terms of the application-specific concept of adjacent databases and aims to hide whether one sample exists in the database. Thus, DP is defined as the detecting probability of outputs of any two adjacent databases.

2) *Experiment-based Methods*: Attack algorithms can evaluate the security and privacy levels directly. In order to evaluate the adversarial robustness of image classification tasks, large-scale experiments have been conducted and the performance of different defense methods can be evaluated [262]. In addition, we can apply adversaries to DP-SGD, which allows for evaluating the gap between the private information that an attacker leaks (a lower bound) and what the privacy analysis establishes as being the maximum leak (an upper bound) [130]. We can notice that attack methods

constantly emerge to face advanced defense methods. Thus, the experiment-based methods need to consume a lot of computation resources, such as 3,000 GPU hours with parallelized over 24 GPUs as shown in [130].

C. Lessons Learned from Attacks and Defenses

The research on attacks and defenses in distributed learning is faced with an “arms race”, i.e., a defense method proposed to prevent the existing attacks will be soon evaded by new attacks, and vice versa.

1) *Attacks in Distributed Learning*: Attack algorithms in the white-box scenario draw a lot of attention in the last few decades, but they seem to be impractical and can only be used as an upper bound. For example, model poisoning attacks in FL can be divided into three scenarios based on various levels of background knowledge, i.e., full knowledge, partial knowledge, and no knowledge. The attack performance decreases drastically as the background knowledge decreases [165], [263], [264]. In this context, practical attack algorithms with no knowledge should be studied to explore potential privacy and security risks. In addition, the organizer usually obtains more background knowledge than the rest of the participants. In order to mitigate the risk of the organizer being an adversary/eavesdropper, the decentralized framework can be adopted as a solution.

For the same attack purpose, different levels of distributed learning require different background knowledge, since the level of distributed learning determines interactive messages which usually contain the private information of participants, such as extracted features and neural network gradients of private data. Thus, various attack methods have emerged to infer private information or poison training process instead of unified attack schemes. For example, MIA in level 1 (sharing model) needs shadow datasets to train shadow models and then estimates the confidence of the training models [54]. We know that the shadow datasets and their distribution affect the attack performance obviously. However, how to obtain the shadow datasets becomes controversial, such as generative networks, stealing, and so on.

2) *Defenses in Distributed Learning*: Although distributed learning can achieve privacy-enhanced and scalable data sharing, it also presents some security and privacy risks. Four-level distributed learning frameworks show various risk levels of privacy leakage, due to the different interactive messages [265], [266]. The interactive messages usually contain the private information of participant users, such as extracted features and neural network gradients of private data. This data process can protect private data to some degree. Thus, it is of interest to study the potential privacy protection levels owing to these data process functions, and then design effective protection schemes to achieve a better trade-off between training performance and privacy.

Privacy/confidential computing for distributed learning is a high requirement compared with conventional privacy protection. However, existing privacy computing techniques usually cannot provide systematic privacy preservation, which will degrade the learning performance or training efficiency [155],

[267]. In addition, the protection effectiveness of different privacy computing techniques varies. For example, DP is seen as an effective method to prevent membership inference attacks by perturbing the impact on whether one instance exists in the training process. Thus, the sensitivity of interactive messages in distributed learning for DP should be carefully investigated when estimating the privacy budget. MPC is another widely used privacy computing technique. However, the transfer ability of MPC is limited and the MPC protocols for different paradigms of distributed learning need to be well-designed. Overall, it is crucial to combine these privacy computing techniques and design a general privacy-preserving framework for different paradigms of distributed learning [268], [269].

D. Lessons Learned from Federated Learning

Reviewing the state-of-the-art in the field, we find that FL plays an increasingly important role in facilitating training ML models for distributed data, as highlighted as follows.

1) *The Advantages of Federated Learning*: Three classic paradigms in FL, i.e., horizontal FL, vertical FL, and federated transfer learning, can be categorized as level 1, level 2 and level 3 of distributed learning, and have the capability to address most of the challenges of training ML models in distributed scenarios. FL is an efficient approach for federated data sharing among multiple clients, in which raw data are kept on the client side, which in turn protects data privacy for tensor mining. The primary purpose of FL is to train a satisfied ML model without exposing participants' data privacy. Thus, when we select or design a training framework, both participants data characteristics and privacy requirements should be considered. In addition, an increasing number of advanced paradigms have emerged to handle various challenges in FL training, such as multi-modal FL [270]–[272], federated knowledge distillation [273]–[275], quantized FL [118] and so on, which help to construct a secure and efficient federated AI ecosystem.

2) *The Disadvantages of Federated Learning*: Although FL can benefit data privacy, security and privacy risks induced by the interactive messages also exist. Particularly, FL can be combined with other privacy techniques, such as DP, MPC, HE, and so on, to improve the privacy of local updates, by integrating them into gradient descent training to enable privacy-enhancing FL. Moreover, the security of FL-based data sharing can be improved by combining it with blockchain technology [276]–[279]. In this context, the information of trained parameters can be appended into immutable blocks on a blockchain during client-server communications. Further, the vast communication cost in vertical FL should be noticed [280]–[282]. Specifically, in vertical FL, the total computation and communication cost is proportional to the training dataset size. In other words, the widely adopted batch computation method in horizontal FL cannot be applied to vertical FL. When facing a massive amount of data, e.g., billions of advertising data, communication, and local computation may be in many orders of magnitude, and the system may lose vitality due to limited resources, such as hardware capacity, bandwidth, and power.

VII. RESEARCH CHALLENGES AND FUTURE DIRECTIONS

As discussed in the above sections, distributed learning systems can alleviate security and privacy concerns by advancing defense mechanisms. In Section VII, we provide and reveal several critical research challenges for further improvement in system implementation. In addition, related possible solutions are also discussed.

A. Balance between ML performance and Security/Privacy Level

1) *Convergence analysis*: As mentioned above, DP has widely been adopted to train a distributed ML model, which will add random noise to gradients during the training process. However, a strict privacy guarantee usually requires a large noise variance injected, so the DP-based training will lead to significant performance degradation. Although existing works in [172], [283] have explored the training performance of the differentially private distributed learning systems and provided some theoretical results, these results can only bring out some intuitions and cannot enhance the learning performance directly. Therefore, an accurate estimation of convergence performance on the differentially private ML training is beneficial to find a proper balance between utility and privacy.

2) *Dynamic parameter optimization*: In addition to the accurate estimation of convergence performance, dynamic parameter optimization is also a promising direction to balance the trade-off between utility and privacy. Because of privacy protection, the training performance caused by the original parameters has been changed. Correspondingly, the conventional parameter optimization method for distributed ML also becomes inapplicable. For example, the work in [172] has developed the upper bound on the differential private FL and revealed that there exists an optimal number of communication rounds with a given privacy level. This discovery brings a new look at the communication round in FL and rethinks the choice of communication parameters. The dynamic parameter optimization for differentially private ML has also been considered, which implements a dynamic privacy budget allocator over the course of training to improve model accuracy [284]. Although existing dynamic optimization methods have already been proposed and proven to improve a number of distributed learning systems obviously, there is still a huge room for improvement.

3) *Specific/personalized protection mechanism*: The various requirements for different scenarios or different participants in distributed ML systems are also challenging, especially when the data distribution is non-independently identically distributed [285], [286]. Therefore, designing a specific/personal protection mechanism for the distributed ML system can bring out a better balance between utility and privacy. The work in [287] has considered a social network and achieved a proven DP requirement by perturbing each participant's option with a designated probability in each round. Combining sketch and DP techniques, the work in [115] has proposed a novel sketch-based framework, which compresses the transmitted messages via sketches to simultaneously achieve communication efficiency and provable privacy

TABLE XI
SUMMARY OF CHALLENGES ALONG WITH THEIR DESCRIPTIONS, AND POSSIBLE SOLUTIONS.

Challenges	Description	Solution
Balance between ML performance and Security/Privacy Level	The tradeoff between the Learning performance, such as convergence, and the privacy and security level should be well designed.	Dynamic parameter optimization Specific/personalized protection mechanism
Decentralized Paradigm	In the distributed fashion, the regulations as well as the incentives among multiple participants should be investigated.	Authentication and access control Consensus design Blockchain assisted distributed learning
Complexity Reduction	Distributed learning with a high complexity security and privacy protection is sometimes impractical. How to alleviate this complexity burden under a required protection level still needs investigation.	Lightweight encryption High-efficiency secure protocol Model compression

benefits. These designs can obtain a satisfactory trade-off between utility and privacy, because of the deep combination of original scenarios and DP techniques. Therefore, how to balance utility and privacy in the amount of distributed learning scenarios has not been fully explored.

4) *Private set intersection (PSI)*: PSI is an important step in distributed learning because of the feature or individual differences among multiple users. For example, in horizontal FL/SGD systems, we need to ensure that each record has the same features. Classical PSI protocols are third party-based PSI [288], [289], public-key-based PSI [290], [291], circuit-based PSI [292] and OT-based PSI [293]. However, there is still a research gap that using PSI in distributed learning to investigate the tradeoff between the privacy level and the learning performance.

B. Decentralized Paradigm

1) *Authentication and access control*: The key question in adding security to a decentralized diagram is to increase the confidence that all parties involved in the system (agents, platforms, and users) will behave correctly, and can be achieved by authentication. The identification of the parties can make up a system and possibly establish a trusting environment between clients. Cryptology is proven useful in a large number of authentication and access control scenarios, but it cannot address the problem of fully new participants. In addition, a trust/reputation model has been proposed to determine the participating values for unknown clients, since it is hard for an agent to obtain complete knowledge about other participants [181], [217], [218]. Consequently, how to design efficient identity certification mechanisms to uniquely authenticate known, and trusted users and agents in the system has drawn much attention.

2) *Consensus design*: Coordination and cooperative control of multi-client in distributed ML always attract lots of attention from various research communities, where a fundamental approach to achieving cooperative control is the consensus-based algorithm [294]. Traditional consensus designs are mostly based on single and finite-time domain [295], [296], where in reality, the dynamics of the system are usually complicated and non-linear. Therefore, a useful and effective consensus design with dynamic or unknown parameters is urgent in future research. For example, the time-varying resources and requirements for participating clients are key and un-trivial

factors in design. In addition, the security of consensus also raises several issues recently [297]. How to protect the integrity of the consensus from inside or outside attackers and how to prevent private information leakage from the published consensus are other interesting research directions.

3) *Blockchain assisted distributed learning*: The reasons for implementing blockchain in a distributed learning system are to increase the interaction efficiency between participants by providing more trusted information exchange, reaching a consensus in trust conditions, assessing participant productivity or detecting performance problems, identifying intruders, allocating plans and tasks, and deploying distributed solutions and joint missions [298], [299]. However, the challenges consist of assessing feasibility and finding an architectural approach for combining blockchain-based consensus algorithms with real-time distributed learning systems, while assuring incentive information exchange and compatibility with the already existent local processing protocols [258]. In addition, the incentive mechanism is also vital for the consensus design [300], [301].

4) *Fairness*: Fairness attracts increasing attention in recent years, especially in the scenario where multiple participants are evolved in one learning task [302]. A max-min fairness distributed learning system has been developed in [303], where multiple clients are matched with the bandits with the minimum regret. Furthermore, collaborative fairness in FL has been investigated in [304]. Although several works throw out the idea of fairness, there is a lack of a common definition of fairness in distributed learning. Whether attending the same rounds of training or allocating training trials according to the users' capability represents fairness is still an unclear question. In addition, the relationship between fairness with security and privacy also requires further discussion.

C. Complexity Reduction

1) *Lightweight encryption*: One of the oldest and most popular techniques used in information security is cryptography, and its use to protect valuable information is usually relying on symmetric encryption and decryption algorithms such as elliptic curve cryptography (ECC), homomorphic hash function, and secret sharing technology. A secure lightweight ECC-Based protocol, i.e., Broadcast based Secure Mobile Agent Protocol (BROSMAP) [305], has been improved to fulfill the needs of Multi-agent based IoT Systems in general and

obtained better performance than its predecessor with the same security requirements. HE assisted MPC framework [176], enabling a participant to compute functions on values while keeping the values hidden, can allow certain mathematical operations (such as aggregation) to be performed directly on ciphertexts, without prior decryption. However, cryptography algorithms usually require complicated computation protocols and may not be achieved efficiently.

2) *High-efficiency secure protocol*: Secure protocols are designed to enable computation over data distributed between different parties so that only the result of the computation is revealed to the participants, but no other private information. Secure protocols usually combine several efficient security and privacy techniques, e.g., MPC, DP, and HE, and need several interactions to exchange intermediate results. However, too many interactions may increase the information leakage risk, communication, and computing overhead. Besides, it is also challenging to explore generic secure protocols over remote parties, especially for complicated scenarios and various applications. To realize an efficient communication protocol in a trusted and secure environment, an alternative way is to increase the transmission rate using an intelligent reflecting surface (IRS) by smartly reconfiguring the wireless propagation environment, with the help of massive low-cost passive reflecting elements integrated on a planar surface and to enable cover communication [306].

3) *Model compression*: High accuracy of large neural networks is often achieved by paying the cost of hungry memory consumption and complex computational capability, which greatly impedes the deployment and development in distributed systems [307]. To efficiently accelerate the learning process, privacy preservation-based methods, such as compact model [308], [309], tensor decomposition [310], data quantization [311] and network sparsification [312], are recent key advances.

D. Distributed ML and Futuristic Technologies

1) *Robotics*: Distributed ML can enhance the ability to identify and control robotics with remote and distributed control or wireless connections to clouds. This scenario requires high precision control, which raises increasing security issues and vulnerability to transmission errors [313], [314]. How to preserve the integrity of the control system and how to prevent information leakage during data transmission needs further investigation. In addition, ethical issues related to bionic robots are hotly debated concerns [315], [316].

2) *Virtual reality (VR) and augmented reality (AR)*: ML and its distributed styles can improve the quality of generated images and videos, such as GAN and diffusion models. With the rapid development in VR and AR-based applications, private information from generated videos may lead to personal information leakage [317], [318]. Adversaries can take advantage of the fake videos to analyze the unique behaviors, personal interests, and background environments of participants [319].

3) *Distributed quantum computing*: Quantum ML operates based on quantum mechanics, taking advantage of superposition to store and process information [320], [321]. However, if

information sources are from distributed clients, information leakage and inside or outside attacks may occur during data transmission. Thus, conducting the protection on distributed ML raises several challenging problems, such as identifying attackers, ensuring the integrity and availability of transmission data, and preserving privacy.

4) *Metaverse*: Metaverse seamlessly integrates the real world with the virtual one. It allows avatars to carry out rich activities, including creation, display, entertainment, social networking, and trading. Thus, it is promising to build an exciting digital world and transform a better physical scenario by exploring the Metaverse [322], [323]. Intuitively, the breakthroughs of AI in the real world motivate people to realize the Metaverse. For example, distributed ML via integrating distributed data from Metaverse users can provide technical support for Metaverse systems to reach or exceed the level of human learning. This can significantly affect the operational efficiency and the intelligence of the Metaverse. Intelligent voice services provide technical support, such as voice recognition and communication. However, several new security and privacy challenges that can compromise the systems or divulge users' privacy raise attention in the interaction process, such as the communication between metaverse users and service providers.

5) *Digital twin*: The digital twin can fill the gap between physical systems and digital spaces. Leveraging FL to construct digital twin models of IoT devices based on their running data has been proposed in [324], [325]. The physical security of IoT devices is critical as they can be damaged, destroyed, or even stolen by attackers. Digital twin systems also have other priorities than the traditional network/system security requirements because of their interactions with the physical components. For instance, defects in a critical product may lead to death, injuries, or environmental damage. For this reason, safety could be ranked as the top security requirement. Safety can broadly be defined as the avoidance of harm or hazard to the physical environment and infrastructure that could occur from system faults [326]. Meanwhile, the possible privacy leakage from the interactions with the physical components must also be considered.

6) *Web 3.0*: Web 3.0 has attracted considerable attention due to its unique decentralized characteristics [327]. In Web 3.0, data presents a distributed storage structure, so there will be no central node for data management, significantly reducing the service cost of managing data. Web 3.0 emphasizes the protection of users personal data, and therefore, as a key technology to solve the data privacy problem, privacy computing is becoming the immediate need of Web 3.0 existence. Privacy computing technology can analyze and calculate data under the premise of protecting data privacy and security, which provides a strong guarantee for the efficient and safe circulation of data across industries and organizations.

7) *Generative design AI*: Generative design uses AI to come up with multiple design variations for products or parts. This leads to a faster generation of design options than would be developed through manual design, which leads to faster product development times and more creative choices to select from. For example, the meteoric rise of diffusion models

has been one of the most significant developments in ML in the past several years [328]. Although generative design AI can improve the qualities of several tasks, it also relies on massive data and may induce several security and privacy issues, especially for fake digital assets, like photos or videos, that are indistinguishable from real things.

E. Development of IEEE standardizations, policy, and regulations

Privacy and security are paramount considerations in the field of distributed learning, where data is shared and processed across various decentralized nodes. To ensure a robust and trustworthy environment for distributed learning systems, several IEEE standards, policies, and regulations come into play. These guidelines help establish a solid foundation for protecting user data and maintaining the integrity of the learning process.

1) IEEE Standards:

- IEEE 1363 (Standard Specifications for Public-Key Cryptography): Encryption is vital for securing data in distributed learning. IEEE 1363 provides specifications for public-key cryptography algorithms, ensuring confidentiality and integrity of communication in distributed systems.
- IEEE P2089 (Standard for Privacy Impact Assessment for Internet of Things): This standard provides a framework for assessing the privacy impact of IoT systems, which often play a crucial role in distributed learning scenarios. It guides the identification of potential privacy risks and suggests mitigation strategies.
- IEEE 3652.1-2020 (Guide for Architectural Framework and Application of Federated Machine Learning)³: It provides a blueprint for data usage and model building across organizations and devices while meeting applicable privacy, security and regulatory requirements in FL. In detail, the description and definition; the categories and the application scenarios to which each category applies; the performance evaluation; and the associated regulatory requirements of FL are defined.
- IEEE P7000 series (Model Process for Addressing Ethical Concerns During System Design): Distributed learning involves ethical considerations, and this series offers a comprehensive model process to address ethical concerns throughout system design and development. It emphasizes transparency, accountability, and user consent.

2) Policies and Regulations:

- GDPR (General Data Protection Regulation)⁴: Although not an IEEE standard, GDPR is a significant regulation that affects distributed learning. It emphasizes the protection of personal data and requires explicit user consent for data processing. Organizations handling data in distributed learning must adhere to GDPR's principles to ensure user privacy.

- HIPAA (Health Insurance Portability and Accountability Act)⁵: In healthcare-related distributed learning applications, HIPAA plays a crucial role. It sets regulations for protecting the privacy and security of patient's health information, including data used in distributed learning scenarios.
- NIST (National Institute of Standards and Technology) Guidelines⁶: While not IEEE-specific, NIST provides guidelines on security and privacy, including those applicable to distributed systems. NIST's cybersecurity framework and privacy framework offer valuable insights for building secure and privacy-preserving distributed learning systems.
- IEEE Code of Ethics⁷: While not a policy or regulation in the legal sense, the IEEE Code of Ethics guides professionals working in technical fields, including distributed learning. It encourages ethical behavior, respect for privacy, and responsible decision-making.

VIII. CONCLUSIONS

As an important and emerging technology, distributed ML has the capability to leverage the incremental amount of data in UEs to the maximum extent. However, this emergence raises increased concerns about privacy and security. In this survey, we have proposed a new framework, which divides distributed ML into four levels for the purpose of understanding privacy and security issues. Moreover, we have discussed and summarized the state-of-the-art related to these issues and revealed the particular characteristics of adversaries at each level. In addition, several research challenges and future directions have also been discussed.

REFERENCES

- [1] J. Chen and X. Ran, "Deep learning with edge computing: A review," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1655–1674, 2019.
- [2] E. Elsebakhi, F. Lee, E. Schendel, A. Haque, N. Kathireason, T. Pathare, N. Syed, and R. Al-Ali, "Large-scale machine learning based on functional networks for biomedical big data with high performance computing platforms," *Journal of Computational Science*, vol. 11, pp. 69–81, 2015.
- [3] P. A. Bernstein and E. Newcomer, *Principles of Transaction Processing (Second Edition)*. Morgan Kaufmann, 2009.
- [4] I. Raicu, I. Foster, A. Szalay, and G. Turcu, "Astroportal: A science gateway for large-scale astronomy data analysis," in *Proc. TeraGrid conference*, 2016, pp. 12–15.
- [5] R. Gu, S. Yang, and F. Wu, "Distributed machine learning on mobile devices: A survey," *Arxiv*, 2019. [Online]. Available: <http://arxiv.org/abs/1909.08329>
- [6] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *Arxiv*, 2016. [Online]. Available: <http://arxiv.org/abs/1610.05492>
- [7] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.
- [8] C. Ma, J. Li, M. Ding, H. H. Yang, F. Shu, T. Q. S. Quek, and H. V. Poor, "On safeguarding privacy and security in the framework of federated learning," *IEEE Network*, vol. 34, no. 4, pp. 242–248, 2020.
- [9] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 2, pp. 1–19, 2019.

³https://standards.ieee.org/standard/3652_1-2020.html

⁴https://ec.europa.eu/info/law/law-topic/data-protection_en

⁵<https://www.hhs.gov/hipaa/index.html>

⁶<https://csrc.nist.gov/>

⁷<https://www.ieee.org/about/ieee-code-of-ethics.html>

- [10] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," *Arxiv*, 2016. [Online]. Available: <https://arxiv.org/abs/1610.05755>
- [11] Z.-K. Zhang, M. C. Y. Cho, C.-W. Wang, C.-W. Hsu, C.-K. Chen, and S. Shieh, "IoT security: Ongoing challenges and research opportunities," in *Proc. IEEE 7th International Conference on Service-Oriented Computing and Applications*, 2014, pp. 230–234.
- [12] Y. Yang, L. Wu, G. Yin, L. Li, and H. Zhao, "A survey on security and privacy issues in internet-of-things," *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1250–1258, 2017.
- [13] D. Peteiro-Barral and B. Guijarro-Berdiñas, "A survey of methods for distributed machine learning," *Progress in Artificial Intelligence*, vol. 2, no. 1, pp. 1–11, 2013.
- [14] J. Verbraeken, M. Wolting, J. Katzy, J. Kloppenburg, T. Verbelen, and J. S. Rellermeyer, "A survey on distributed machine learning," *ACM Computing Surveys*, vol. 53, no. 2, Mar. 2020.
- [15] E. De Cristofaro, "An overview of privacy in machine learning," *Arxiv*, 2020. [Online]. Available: <https://arxiv.org/abs/2005.08679>
- [16] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin, "When machine learning meets privacy: A survey and outlook," *ACM Computing Surveys*, vol. 54, no. 2, Mar. 2021.
- [17] P. Vepakomma, T. Swedish, R. Raskar, O. Gupta, and A. Dubey, "No peek: A survey of private distributed deep learning," *Arxiv*, 2018. [Online]. Available: <http://arxiv.org/abs/1812.03288>
- [18] M. Gong, Y. Xie, K. Pan, K. Feng, and A. K. Qin, "A survey on differentially private machine learning," *IEEE Computational Intelligence Magazine*, vol. 15, no. 2, pp. 49–64, 2020.
- [19] M. Amiri-Zarandi, R. Dara, and E. D. G. Fraser, "A survey of machine learning-based solutions to protect privacy in the internet of things," *Computers & Security*, vol. 96, 2020.
- [20] C. Briggs, Z. Fan, and P. Andras, "A review of privacy preserving federated learning for private IoT analytics," *Arxiv*, 2020. [Online]. Available: <https://arxiv.org/abs/2004.11794>
- [21] D. Enthoven and Z. Al-Ars, *An Overview of Federated Deep Learning Privacy Attacks and Defensive Strategies*. Springer Cham, 2021.
- [22] L. Lyu, H. Yu, and Q. Yang, "Threats to federated learning: A survey," *Arxiv*, 2020. [Online]. Available: <https://arxiv.org/abs/2003.02133>
- [23] G. Xu, H. Li, H. Ren, K. Yang, and R. H. Deng, "Data security issues in deep learning: Attacks, countermeasures, and opportunities," *IEEE Communications Magazine*, vol. 57, no. 11, pp. 116–122, 2019.
- [24] R. Xu, N. Baracaldo, and J. Joshi, "Privacy-preserving machine learning: Methods, challenges and directions," *Arxiv*, 2021. [Online]. Available: <http://arxiv.org/abs/2108.04417>
- [25] R. J. Schalkoff, *Pattern Recognition*. American Cancer Society, 2007.
- [26] W. contributors, "Training, validation, and test sets — Wikipedia, the free encyclopedia," 2021, [Online; accessed 29-July-2021].
- [27] E. P. Xing, Q. Ho, P. Xie, and D. Wei, "Strategies and principles of distributed machine learning on big data," *Engineering*, vol. 2, no. 2, pp. 179–195, 2016.
- [28] M. Zinkevich, M. Weimer, L. Li, and A. Smola, "Parallelized stochastic gradient descent," in *Advances in Neural Information Processing Systems*, vol. 23, 2010.
- [29] Q. Ho, J. Cipar, H. Cui, S. Lee, J. K. Kim, P. B. Gibbons, G. A. Gibson, G. Ganger, and E. P. Xing, "More effective distributed ML via a stale synchronous parallel parameter server," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2013, pp. 1223–1231.
- [30] X. Lian, Y. Huang, Y. Li, and J. Liu, "Asynchronous parallel stochastic gradient for nonconvex optimization," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., 2015, pp. 2737–2745.
- [31] B. Recht, C. Re, S. Wright, and F. Niu, "Hogwild: A lock-free approach to parallelizing stochastic gradient descent," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2011, pp. 693–701.
- [32] M. Li, D. G. Andersen, A. J. Smola, and K. Yu, "Communication efficient distributed machine learning with the parameter server," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 19–27.
- [33] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: A system for large-scale machine learning," in *Proc. USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, Savannah, GA, Nov. 2016, pp. 265–283.
- [34] T. Chilimbi, Y. Suzue, J. Apacible, and K. Kalyanaraman, "Project adam: Building an efficient and scalable deep learning training system," in *Proc. USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, Broomfield, CO, Oct. 2014, pp. 571–582.
- [35] K. Hsieh, A. Harlap, N. Vijaykumar, D. Konomis, G. R. Ganger, P. B. Gibbons, and O. Mutlu, "Gaia: Geo-distributed machine learning approaching LAN speeds," in *Proc. USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, Boston, MA, mar. 2017, pp. 629–647.
- [36] M. Li, Z. Liu, A. J. Smola, and Y.-X. Wang, "DiFacto: Distributed factorization machines," in *Proc. ACM International Conference on Web Search and Data Mining*, San Francisco, California, USA, 2016, pp. 377–386.
- [37] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017, pp. 1273–1282.
- [38] J. Konečný, B. McMahan, and D. Ramage, "Federated optimization: Distributed optimization beyond the datacenter," *Arxiv*, 2015. [Online]. Available: <http://arxiv.org/abs/1511.03575>
- [39] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečný, S. Mazzocchi, H. B. McMahan *et al.*, "Towards federated learning at scale: System design," *Arxiv*, 2019. [Online]. Available: <https://arxiv.org/abs/1902.01046>
- [40] D. L. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms," *Communications of the ACM*, vol. 24, no. 2, pp. 84–90, Feb. 1981.
- [41] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Communications Surveys Tutorials*, vol. 22, no. 3, pp. 2031–2063, 2020.
- [42] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *Proc. IEEE Symposium on Security and Privacy (SP)*, 2019, pp. 691–706.
- [43] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS)*, Denver, CO, USA, Oct. 2015, pp. 1322–1333.
- [44] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep models under the GAN: Information leakage from collaborative deep learning," in *Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2017, pp. 603–618.
- [45] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020, pp. 2938–2948.
- [46] C. He, M. Annamalai, and S. Avestimehr, "Group knowledge transfer: Federated learning of large CNNs at the edge," in *Proc. International Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada, 2020.
- [47] P. Vepakomma, O. Gupta, T. Swedish, and R. Raskar, "Split learning for health: Distributed deep learning without sharing raw patient data," *Arxiv*, 2018. [Online]. Available: <http://arxiv.org/abs/1812.00564>
- [48] P. Vepakomma, O. Gupta, A. Dubey, and R. Raskar, "Reducing leakage in distributed deep learning for sensitive health data," *Arxiv*, 2019. [Online]. Available: <https://arxiv.org/abs/1812.00564>
- [49] O. Gupta and R. Raskar, "Distributed learning of deep neural network over multiple agents," *Journal of Network and Computer Applications*, vol. 116, pp. 1–8, 2018.
- [50] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch SGD: Training imagenet in 1 hour," *Arxiv*, 2017. [Online]. Available: <http://arxiv.org/abs/1706.02677>
- [51] J. Chen, X. Pan, R. Monga, S. Bengio, and R. Jozefowicz, "Revisiting distributed synchronous SGD," *Arxiv*, 2016. [Online]. Available: <http://arxiv.org/abs/1604.00981>
- [52] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2015, pp. 1310–1321.
- [53] W. Li, B. Jin, X. Wang, J. Yan, and H. Zha, "F2A2: Flexible fully-decentralized approximate actor-critic for cooperative multi-agent reinforcement learning," *Arxiv*, 2020. [Online]. Available: <https://arxiv.org/abs/2004.11145>
- [54] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 3–18.
- [55] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning,"

- in *Proc. ACM on Asia conference on computer and communications security (ASIACSS)*, 2017, pp. 506–519.
- [56] C. Ma, J. Li, M. Ding, K. Wei, W. Chen, and H. V. Poor, “Federated learning with unreliable clients: Performance analysis and mechanism design,” *IEEE Internet of Things Journal*, vol. 8, no. 24, pp. 17308–17319, 2021.
- [57] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *Arxiv*, 2015. [Online]. Available: <https://arxiv.org/abs/1412.6572>
- [58] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, “Manipulating machine learning: Poisoning attacks and countermeasures for regression learning,” in *Proc. IEEE Symposium on Security and Privacy (SP)*, 2018, pp. 19–35.
- [59] O. Suciu, R. Marginean, Y. Kaya, H. D. III, and T. Dumitras, “When does machine learning FAIL? Generalized transferability for evasion and poisoning attacks,” in *Proc. USENIX Security Symposium (USENIX Security)*, Baltimore, MD, Aug. 2018, pp. 1299–1316.
- [60] X. Yuan, P. He, Q. Zhu, and X. Li, “Adversarial examples: Attacks and defenses for deep learning,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2805–2824, 2019.
- [61] H. Kwon, Y. Kim, H. Yoon, and D. Choi, “Selective audio adversarial example in evasion attack on speech recognition system,” *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 526–538, 2020.
- [62] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, “Robust physical-world attacks on deep learning visual classification,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [63] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, “Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing,” in *Proc. USENIX Security Symposium (USENIX Security)*, 2014, pp. 17–32.
- [64] J. A. Calandrino, A. Kilzer, A. Narayanan, E. W. Felten, and V. Shmatikov, ““you might also like:” Privacy risks of collaborative filtering,” in *Proc. IEEE Symposium on Security and Privacy (SP)*, 2011, pp. 231–246.
- [65] C. Song, T. Ristenpart, and V. Shmatikov, “Machine learning models that remember too much,” in *Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2017, pp. 587–601.
- [66] G. Ateniese, L. V. Mancini, A. Spognardi, A. Villani, D. Vitali, and G. Felici, “Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers,” *International Journal of Security and Networks*, vol. 10, no. 3, pp. 137–150, 2015.
- [67] Y. Aono, T. Hayashi, L. Wang, S. Moriai *et al.*, “Privacy-preserving deep learning: Revisited and enhanced,” in *Proc. International Conference on Applications and Techniques in Information Security (ATIS)*, 2017, pp. 100–110.
- [68] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, “Inference attacks against collaborative learning,” *Arxiv*, 2018. [Online]. Available: <http://arxiv.org/abs/1805.04049>
- [69] S. Truex, L. Liu, M. E. Gursoy, L. Yu, and W. Wei, “Towards demystifying membership inference attacks,” *Arxiv*, 2018. [Online]. Available: <http://arxiv.org/abs/1807.09173>
- [70] M. Backes, P. Berrang, M. Humbert, and P. Manoharan, “Membership privacy in microRNA-based studies,” in *Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2016, pp. 319–330.
- [71] N. Homer, S. Szlinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig, “Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays,” *PLoS genetics*, vol. 4, no. 8, 2008.
- [72] A. Pyrgelis, C. Troncoso, and E. De Cristofaro, “Knock knock, who’s there? Membership inference on aggregate location data,” *Arxiv*, 2017. [Online]. Available: <http://arxiv.org/abs/1708.06145>
- [73] C. Dwork, A. Smith, T. Steinke, J. Ullman, and S. Vadhan, “Robust traceability from trace amounts,” in *Proc. IEEE 56th Annual Symposium on Foundations of Computer Science (FCS)*, 2015, pp. 650–669.
- [74] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, “Stealing machine learning models via prediction APIs,” in *Proc. USENIX Security Symposium (USENIX Security)*, 2016, pp. 601–618.
- [75] B. Wang and N. Z. Gong, “Stealing hyperparameters in machine learning,” in *Proc. IEEE Symposium on Security and Privacy (SP)*, 2018, pp. 36–52.
- [76] S. J. Oh, B. Schiele, and M. Fritz, *Towards Reverse-Engineering Black-Box Neural Networks*. Cham: Springer International Publishing, 2019.
- [77] T. Orekondy, B. Schiele, and M. Fritz, “Knockoff nets: Stealing functionality of black-box models,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4954–4963.
- [78] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, “Practical secure aggregation for privacy-preserving machine learning,” in *Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS)*, Dallas, Texas, USA, 2017, pp. 1175–1191.
- [79] J. Liu, M. Juuti, Y. Lu, and N. Asokan, “Oblivious neural network prediction via miniONN transformations,” in *Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS)*, Dallas, Texas, USA, 2017, pp. 619–631.
- [80] P. Mohassel and Y. Zhang, “SecureML: A system for scalable privacy-preserving machine learning,” in *Proc. IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 19–38.
- [81] V. Nikolaenko, U. Weinsberg, S. Ioannidis, M. Joye, and N. Taft, “Privacy-preserving ridge regression on hundreds of millions of records,” in *Proc. IEEE Symposium on Security and Privacy (SP)*, 2013.
- [82] W. Du, Y. S. Han, and S. Chen, “Privacy-preserving multivariate statistical analysis: Linear regression and classification,” in *Proc. SIAM International Conference on Data Mining (SDM)*, 2004, pp. 222–233.
- [83] R. Bost, R. A. Popa, S. Tu, and S. Goldwasser, “Machine learning classification over encrypted data,” in *Proc. Network and Distributed System Security Symposium (NDSS)*, 2014.
- [84] K. Lauter and M. Naehrig, “Private predictive analysis on encrypted medical data,” *Journal of Biomedical Informatics*, 2014.
- [85] T. Graepel, K. Lauter, and M. Naehrig, “ML confidential: Machine learning on encrypted data,” in *Proc. International Conference on Information Security and Cryptology (ICISC)*, 2012.
- [86] N. Kilbertus, A. Gascón, M. J. Kusner, M. Veale, K. P. Gummadi, and A. Weller, “Blind justice: Fairness with encrypted sensitive attributes,” *Arxiv*, 2018. [Online]. Available: <http://arxiv.org/abs/1806.03281>
- [87] Y. Li, Y. Duan, Y. Yu, S. Zhao, and W. Xu, “PrivPy: Enabling scalable and general privacy-preserving machine learning,” *Arxiv*, 2018. [Online]. Available: <http://arxiv.org/abs/1801.10117>
- [88] P. Mohassel and P. Rindal, “ABY3: A mixed protocol framework for machine learning,” in *Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2018, pp. 35–52.
- [89] T. Araki, J. Furukawa, Y. Lindell, A. Nof, and K. Ohara, “High-throughput semi-honest secure three-party computation with an honest majority,” in *Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2016, pp. 805–817.
- [90] P. Mohassel, M. Rosulek, and Y. Zhang, “Fast and secure three-party computation: The garbled circuit approach,” in *Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2015, pp. 591–602.
- [91] J. Furukawa, Y. Lindell, A. Nof, and O. Weinstein, “High-throughput secure three-party computation for malicious adversaries and an honest majority,” in *Proc. Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT)*, 2017, pp. 225–255.
- [92] A. Acar, H. Aksu, A. S. Uluagac, and M. Conti, “A survey on homomorphic encryption schemes: Theory and implementation,” *ACM Computing Surveys (CSUR)*, vol. 51, no. 4, pp. 1–35, 2018.
- [93] Q. Zhang, L. T. Yang, and Z. Chen, “Privacy preserving deep computation model on cloud for big data feature learning,” *IEEE Transactions on Computers*, vol. 65, no. 5, pp. 1351–1362, 2015.
- [94] J. Yuan and S. Yu, “Privacy preserving back-propagation neural network learning made practical with cloud computing,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 1, pp. 212–221, 2013.
- [95] A. Acar, H. Aksu, A. S. Uluagac, and M. Conti, “A survey on homomorphic encryption schemes: Theory and implementation,” *ACM Computing Surveys*, vol. 51, no. 4, July 2018.
- [96] W. Lu, S. Kawasaki, and J. Sakuma, “Using fully homomorphic encryption for statistical analysis of categorical, ordinal and numerical data,” in *Proc. Annual Network and Distributed System Security Symposium (NDSS)*, San Diego, California, USA, 2017.
- [97] L. Lamport, R. E. Shostak, and M. C. Pease, *The Byzantine Generals Problem*, 2019.
- [98] P. Blanchard, E. M. E. Mhamdi, R. Guerraoui, and J. Stainer, “Machine learning with adversaries: Byzantine tolerant gradient descent,” in *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 119–129.
- [99] E. M. El Mhamdi, R. Guerraoui, and S. Rouault, “The hidden vulnerability of distributed learning in byzantium,” in *Proc. International*

- Conference on Machine Learning (ICML)*, vol. 80, Jul. 2018, pp. 3521–3530.
- [100] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han, “Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation,” in *Proc. ACM SIGMOD International Conference on Management of Data (SIGMOD)*, 2014.
- [101] K. Pillutla, S. M. Kakade, and Z. Harchaoui, “Robust aggregation for federated learning,” *IEEE Transactions on Signal Processing*, vol. 70, pp. 1142–1154, 2022.
- [102] H. Chang, V. Shejwalkar, R. Shokri, and A. Houmansadr, “Cronus: Robust and heterogeneous collaborative learning with black-box knowledge transfer,” *Arxiv*, 2019. [Online]. Available: <https://arxiv.org/abs/1912.11279>
- [103] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart, “Being robust (in high dimensions) can be practical,” in *Proc. International Conference on Machine Learning (ICML)*, vol. 70, Aug. 2017, pp. 999–1008.
- [104] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, “QSGD: Communication-efficient SGD via gradient quantization and encoding,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
- [105] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, “signSGD: Compressed optimisation for non-convex problems,” in *Proc. International Conference on Machine Learning (ICML)*, vol. 80, Jul. 2018, pp. 560–569.
- [106] W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li, “TernGrad: Ternary gradients to reduce communication in distributed deep learning,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 1509–1519.
- [107] D. Alistarh, T. Hoefler, M. Johansson, N. Konstantinov, S. Khirirat, and C. Renggli, “The convergence of sparsified gradient methods,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, Montréal, Canada, 2018, pp. 5977–5987.
- [108] S. U. Stich, J. Cordonnier, and M. Jaggi, “Sparsified SGD with memory,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, Montréal, Canada, 2018, pp. 4452–4463.
- [109] M. Lin, Q. Chen, and S. Yan, “Network in network,” *Arxiv*, 2013. [Online]. Available: <http://arxiv.org/abs/1312.4400>
- [110] E. Jeong, S. Oh, H. Kim, J. Park, M. Bennis, and S.-L. Kim, “Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data,” *Arxiv*, 2018. [Online]. Available: <http://arxiv.org/abs/1811.11479>
- [111] Z. Liu, M. Sun, T. Zhou, G. Huang, and T. Darrell, “Rethinking the value of network pruning,” in *Proc. International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA, 2019.
- [112] Y. Jiang, S. Wang, B. Ko, W. Lee, and L. Tassiulas, “Model pruning enables efficient federated learning on edge devices,” *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- [113] F. Haddadpour, B. Karimi, P. Li, and X. Li, “FedSKETCH: Communication-efficient and private federated learning via sketching,” *Arxiv*, 2020. [Online]. Available: <https://arxiv.org/abs/2008.04975>
- [114] J. Jiang, F. Fu, T. Yang, and B. Cui, “SketchML: Accelerating distributed machine learning with data sketches,” in *Proc. ACM SIGMOD International Conference on Management of Data (SIGMOD)*, Houston, TX, USA, 2018, pp. 1269–1284.
- [115] T. Li, Z. Liu, V. Sekar, and V. Smith, “Privacy for free: Communication-efficient learning with differential privacy using sketches,” *Arxiv*, 2019. [Online]. Available: <https://arxiv.org/abs/1911.00972>
- [116] J. Yoon, W. Jeong, G. Lee, E. Yang, and S. J. Hwang, “Federated continual learning with adaptive parameter communication,” 2020. [Online]. Available: <https://arxiv.org/abs/2003.03196>
- [117] L. T. Phong, Y. Aono, T. Hayashi, L. Wang, and S. Moriai, “Privacy-preserving deep learning via additively homomorphic encryption,” *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 5, pp. 1333–1345, 2018.
- [118] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, “FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization,” in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 108, Aug. 2020, pp. 2021–2031.
- [119] D. Rothchild, A. Panda, E. Ullah, N. Ivkin, I. Stoica, V. Braverman, J. Gonzalez, and R. Arora, “FetchSGD: Communication-efficient federated learning with sketching,” in *Proc. International Conference on Machine Learning (ICML)*, vol. 119, Jul. 2020, pp. 8253–8265.
- [120] R. Jin, Y. Huang, X. He, H. Dai, and T. Wu, “Stochastic-sign SGD for federated learning with theoretical guarantees,” *Arxiv*, 2020. [Online]. Available: <https://arxiv.org/abs/2002.10940>
- [121] H. Li and T. Han, “An end-to-end encrypted neural network for gradient updates transmission in federated learning,” *Arxiv*, 2019. [Online]. Available: <https://arxiv.org/abs/1908.08340>
- [122] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” *Arxiv*, 2012. [Online]. Available: <http://arxiv.org/abs/1207.0580>
- [123] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [124] C. Dwork, “Differential privacy: A survey of results,” in *Proc. International conference on theory and applications of models of computation (TAMC)*, 2008, pp. 1–19.
- [125] B. I. Rubinstein, P. L. Bartlett, L. Huang, and N. Taft, “Learning in a large function space: Privacy-preserving mechanisms for SVM learning,” *Arxiv*, 2009. [Online]. Available: <http://arxiv.org/abs/0911.5708>
- [126] J. Zhang, Z. Zhang, X. Xiao, Y. Yang, and M. Winslett, “Functional mechanism: Regression analysis under differential privacy,” *Arxiv*, 2012. [Online]. Available: <http://arxiv.org/abs/1208.0219>
- [127] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, “Deep learning with differential privacy,” in *Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2016, pp. 308–318.
- [128] I. Mironov, “Rényi differential privacy,” in *Proc. IEEE Computer Security Foundations Symposium (CSF)*, 2017, pp. 263–275.
- [129] C. Dwork and G. N. Rothblum, “Concentrated differential privacy,” *Arxiv*, 2016. [Online]. Available: <http://arxiv.org/abs/1603.01887>
- [130] M. Nasr, S. Song, A. Thakurta, N. Papernot, and N. Carlini, “Adversary instantiation: Lower bounds for differentially private machine learning,” in *Proc. IEEE Symposium on Security and Privacy (SP)*, San Francisco, CA, USA, 2021, pp. 866–882.
- [131] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, “Local privacy and statistical minimax rates,” in *Proc. IEEE 54th Annual Symposium on Foundations of Computer Science (FCS)*, 2013, pp. 429–438.
- [132] T. Wang, J. Blocki, N. Li, and S. Jha, “Locally differentially private protocols for frequency estimation,” in *Proc. USENIX Security Symposium (USENIX Security)*, Vancouver, BC, Aug. 2017, pp. 729–745.
- [133] R. C. Wong and A. W. Fu, *Privacy-Preserving Data Publishing: An Overview*, ser. Synthesis Lectures on Data Management, 2010.
- [134] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, “Boosting adversarial attacks with momentum,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 9185–9193.
- [135] S. Ji, P. Mittal, and R. A. Beyah, “Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey,” *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 1305–1326, 2017.
- [136] S. H. Huang, N. Papernot, I. J. Goodfellow, Y. Duan, and P. Abbeel, “Adversarial attacks on neural network policies,” in *Proc. International Conference on Learning Representations (ICLR)*, Toulon, France, Apr. 2017.
- [137] X. Pan, C. Xiao, W. He, S. Yang, J. Peng, M. Sun, M. Liu, B. Li, and D. Song, “Characterizing attacks on deep reinforcement learning,” in *Proc. International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Virtual Event, New Zealand, 2022, pp. 1010–1018.
- [138] C. Zhong, F. Wang, M. C. Guroy, and S. Velipasalar, “Adversarial jamming attacks on deep reinforcement learning based dynamic multichannel access,” in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, 2020, pp. 1–6.
- [139] A. Narayanan and V. Shmatikov, “Robust de-anonymization of large sparse datasets,” in *Proc. IEEE Symposium on Security and Privacy (SP)*, Oakland, California, USA, May 2008, pp. 111–125.
- [140] Y. Dong, Z. Deng, T. Pang, J. Zhu, and H. Su, “Adversarial distributional training for robust deep learning,” in *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, virtual, Dec. 2020.
- [141] L. Sweeney, “K-anonymity: A model for protecting privacy,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.
- [142] S. Shaham, M. Ding, B. Liu, S. Dang, Z. Lin, and J. Li, “Privacy preserving location data publishing: A machine learning approach,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 9, pp. 3270–3283, 2021.
- [143] M. Malekzadeh, R. G. Clegg, A. Cavallaro, and H. Haddadi, “Mobile sensor data anonymization,” in *Proc. International Conference on*

- Internet of Things Design and Implementation (IoTDI)*, Montreal, QC, Canada, Apr. 2019, pp. 49–58.
- [144] M. Maximov, I. Elezi, and L. Leal-Taixé, “CIAGAN: Conditional identity anonymization generative adversarial networks,” in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 5446–5455.
- [145] P. Samarati and L. Sweeney, “Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression,” technical report, SRI International, Tech. Rep., 1998.
- [146] P. Samarati, “Protecting respondents identities in microdata release,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010–1027, Nov. 2001.
- [147] K. El Emam and F. K. Dankar, “Protecting privacy using k-anonymity,” *Journal of the American Medical Informatics Association*, vol. 15, no. 5, pp. 627–637, 2008.
- [148] N. Li, T. Li, and S. Venkatasubramanian, “t-closeness: Privacy beyond k-anonymity and l-diversity,” in *Proc. IEEE 23rd International Conference on Data Engineering (ICDE)*, 2007, pp. 106–115.
- [149] S. Shaham, M. Ding, B. Liu, S. Dang, Z. Lin, and J. Li, “Privacy preservation in location-based services: A novel metric and attack model,” *IEEE Transactions on Mobile Computing*, vol. 20, no. 10, pp. 3006–3019, 2021.
- [150] A. M. Shaker, M. M. Tantawi, H. A. Shedeed, and M. F. Tolba, “Generalization of convolutional neural networks for ECG classification using generative adversarial networks,” *IEEE Access*, vol. 8, pp. 35 592–35 605, 2020.
- [151] Z. Ji, Z. C. Lipton, and C. Elkan, “Differential privacy and machine learning: A survey and review,” *Arxiv*, 2014. [Online]. Available: <http://arxiv.org/abs/1412.7584>
- [152] Ú. Erlingsson, V. Pihur, and A. Korolova, “Rappor: Randomized aggregatable privacy-preserving ordinal response,” in *Proc. ACM SIGSAC conference on computer and communications security (CCS)*, 2014, pp. 1054–1067.
- [153] G. Fanti, V. Pihur, and Ú. Erlingsson, “Building a rappor with the unknown: Privacy-preserving learning of associations and data dictionaries,” *Proceedings on Privacy Enhancing Technologies*, vol. 2016, no. 3, pp. 41–61, 2016.
- [154] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating noise to sensitivity in private data analysis,” in *Proc. Theory of cryptography conference (TCC)*, 2006, pp. 265–284.
- [155] M. Balcan, A. Blum, S. Fine, and Y. Mansour, “Distributed learning, communication complexity and privacy,” in *Proc. Annual Conference on Learning Theory (COLT)*, vol. 23, Edinburgh, Scotland, Jun. 2012, pp. 26.1–26.22.
- [156] O. Fagbohunge, S. R. Reza, X. Dong, and L. Qian, “Efficient privacy preserving edge computing framework for image classification,” *Arxiv*, 2020. [Online]. Available: <https://arxiv.org/abs/2005.04563>
- [157] D. Froelicher, J. R. Troncoso-Pastoriza, A. Pyrgelis, S. Sav, J. S. Sousa, J. Bossuat, and J. Hubaux, “Scalable privacy-preserving distributed learning,” *Proceedings on Privacy Enhancing Technologies*, vol. 2021, no. 2, pp. 323–347, 2021.
- [158] K. Xu, H. Yue, L. Guo, Y. Guo, and Y. Fang, “Privacy-preserving machine learning algorithms for big data systems,” in *Proc. IEEE International Conference on Distributed Computing Systems (ICDCS)*, Columbus, OH, USA, Jun. 2015, pp. 318–327.
- [159] R. Zhang and Q. Zhu, “Secure and resilient distributed machine learning under adversarial environments,” in *Proc. International Conference on Information Fusion (FUSION)*, Washington, DC, USA, Jul. 2015, pp. 644–651.
- [160] J. Hur, “Improving security and efficiency in attribute-based data sharing,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 10, pp. 2271–2282, 2013.
- [161] S. L. Warner, “Randomized response: A survey technique for eliminating evasive answer bias,” *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965.
- [162] “Learning with privacy at scale-apple machine learning research,” 2022. [Online]. Available: <https://machinelearning.apple.com/research/learning-with-privacy-at-scale>
- [163] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, “How to backdoor federated learning,” in *Proc. Twenty Third International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 108, Aug. 2020, pp. 2938–2948.
- [164] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, “Analyzing Federated Learning through an Adversarial Lens,” in *Proc. International Conference on Machine Learning (ICML)*, vol. 97, Long Beach, California, USA, Jun. 2019, pp. 634–643.
- [165] M. Fang, X. Cao, J. Jia, and N. Z. Gong, “Local model poisoning attacks to byzantine-robust federated learning,” in *Proc. Usenix Security Symposium (USENIX Security)*, Aug. 2020.
- [166] X. Pan, W. Wang, X. Zhang, B. Li, J. Yi, and D. Song, “How you act tells a lot: Privacy-leaking attack on deep reinforcement learning,” in *Proc. International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, Montreal QC, Canada, 2019, pp. 368–376.
- [167] Y. Zhao, I. Shumailov, H. Cui, X. Gao, R. Mullins, and R. Anderson, “Blackbox attacks on reinforcement learning agents using approximated temporal information,” in *Proc. Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, 2020, pp. 16–24.
- [168] Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, and H. Qi, “Beyond inferring class representatives: User-level privacy leakage from federated learning,” in *Proc. IEEE Conference on Computer Communications (INFOCOM)*, 2019, pp. 2512–2520.
- [169] J. Ge, Z. Wang, M. Wang, and H. Liu, “Minimax-optimal privacy-preserving sparse pca in distributed systems,” in *Proc. Twenty-First International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 84, Apr. 2018, pp. 1589–1598.
- [170] Z. Huang, R. Hu, Y. Guo, E. Chan-Tin, and Y. Gong, “DP-ADMM: ADMM-based distributed learning with differential privacy,” *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1002–1012, 2020.
- [171] R. C. Geyer, T. Klein, and M. Nabi, “Differentially private federated learning: A client level perspective,” *Arxiv*, 2017. [Online]. Available: <http://arxiv.org/abs/1712.07557>
- [172] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. S. Quek, and H. V. Poor, “Federated learning with differential privacy: Algorithms and performance analysis,” *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020.
- [173] C. X. Wang, Y. Song, and W. P. Tay, “Arbitrarily strong utility-privacy tradeoff in multi-agent systems,” *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 671–684, 2021.
- [174] D. Ye, T. Zhu, W. Zhou, and P. S. Yu, “Differentially private malicious agent avoidance in multiagent advising learning,” *IEEE Transactions on Cybernetics*, vol. 50, no. 10, pp. 4214–4227, 2020.
- [175] V. Gandikota, D. Kane, R. K. Maity, and A. Mazumdar, “vqSGD: Vector quantized stochastic gradient descent,” in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 130, Virtual Event, Apr. 2021, pp. 2197–2205.
- [176] G. Xu, H. Li, S. Liu, K. Yang, and X. Lin, “VerifyNet: Secure and verifiable federated learning,” *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 911–926, 2020.
- [177] L. Muñoz-González, K. T. Co, and E. C. Lupu, “Byzantine-robust federated machine learning through adaptive model averaging,” *Arxiv*, 2019. [Online]. Available: <https://arxiv.org/abs/1909.05125>
- [178] S. Fu, C. Xie, B. Li, and Q. Chen, “Attack-resistant federated learning with residual-based reweighting,” *Arxiv*, 2019. [Online]. Available: <https://arxiv.org/abs/1912.11464>
- [179] L. Zhao, S. Hu, Q. Wang, J. Jiang, C. Shen, X. Luo, and P. Hu, “Shielding collaborative learning: Mitigating poisoning attacks through client-side detection,” *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 5, pp. 2029–2041, 2021.
- [180] Y. Zhao, J. Chen, J. Zhang, D. Wu, J. Teng, and S. Yu, “PDGAN: A novel poisoning defense method in federated learning using generative adversarial network,” in *Proc. International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP)*, Cham, 2020, pp. 595–609.
- [181] Y. Wang and M. P. Singh, “Formal trust model for multiagent systems,” in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, M. M. Veloso, Ed., Hyderabad, India, Jan. 2007, pp. 1551–1556.
- [182] K. Danilov, R. Rezin, I. Afanasyev, and A. Kolotov, “Towards blockchain-based robonomics: Autonomous agents behavior validation,” in *Proc. IEEE International Conference on Intelligent Systems (IS)*, Funchal, Madeira, Portugal, Sep. 2018, pp. 222–227.
- [183] V. Strobel, E. C. Ferrer, and M. Dorigo, “Managing byzantine robots via blockchain technology in a swarm robotics collective decision making scenario,” in *Proc. International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, Stockholm, Sweden, Jul. 2018, pp. 541–549.
- [184] D. Calvaresi, Y. Mualla, A. Najjar, S. Galland, and M. Schumacher, “Explainable multi-agent systems through blockchain technology,” in *Proc. International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS)*, vol. 11763, Montreal, QC, Canada, May 2019, pp. 41–58.

- [185] L. Xiao, A. Peet, P. H. Lewis, S. Dashmapatra, C. Sáez, M. Croitoru, J. Vicente, H. González-Vélez, and M. L. i Ariet, "An adaptive security model for multi-agent systems and application to a clinical trials environment," in *Proc. Annual International Computer Software and Applications Conference (COMPSAC)*, vol. 2, Beijing, China, Jul. 2007, pp. 261–268.
- [186] S. Ahmad and M. U. Bokhari, "A new approach to multi-agent based architecture for secure and effective E-learning," *International Journal of Computer Applications*, vol. 46, no. 22, May 2012.
- [187] S. Bosse, "Mobile multi-agent systems for the internet-of-things and clouds using the javascript agent machine platform and machine learning as a service," in *Proc. IEEE International Conference on Future Internet of Things and Cloud (FiCloud)*, Vienna, Austria, Aug. 2016, pp. 244–253.
- [188] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.
- [189] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in *Proc. ACM Workshop on Security and Artificial Intelligence (AISec)*, Chicago, Illinois, USA, 2011, pp. 43–58.
- [190] G. Baruch, M. Baruch, and Y. Goldberg, "A little is enough: Circumventing defenses for distributed learning," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 8632–8642.
- [191] X. Gong, Y. Chen, H. Huang, Y. Liao, S. Wang, and Q. Wang, "Coordinated backdoor attacks against federated learning with model-dependent triggers," *IEEE Network*, vol. 36, no. 1, pp. 84–90, 2022.
- [192] X. Luo, Y. Wu, X. Xiao, and B. C. Ooi, "Feature inference attack on model predictions in vertical federated learning," in *Proc. IEEE 37th International Conference on Data Engineering (ICDE)*, 2021, pp. 181–192.
- [193] Y.-C. Lin, Z.-W. Hong, Y.-H. Liao, M.-L. Shih, M.-Y. Liu, and M. Sun, "Tactics of adversarial attack on deep reinforcement learning agents," in *Proc. the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, 2017, pp. 3756–3762.
- [194] V. Behzadan and A. Munir, "Vulnerability of deep reinforcement learning to policy induction attacks," in *Proc. International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM)*, 2017, pp. 262–275.
- [195] A. Russo and A. Proutière, "Optimal attacks on reinforcement learning policies," *Arxiv*, 2019. [Online]. Available: <http://arxiv.org/abs/1907.13548>
- [196] Y. Ma, X. Zhang, W. Sun, and X. Zhu, "Policy poisoning in batch reinforcement learning and control," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada, Dec. 2019, pp. 14 543–14 553.
- [197] P. Kiourt, K. Wardaga, S. Jha, and W. Li, "TrojDRL: Evaluation of backdoor attacks on deep reinforcement learning," in *Proc. ACM/EDAC/IEEE Design Automation Conference (DAC)*, Virtual Event, USA, Jul.
- [198] X. Zhang, Y. Ma, A. Singla, and X. Zhu, "Adaptive reward-poisoning attacks against reinforcement learning," in *Proc. International Conference on Machine Learning (ICML)*, Jul. 2020, pp. 11 225–11 234.
- [199] M. Huai, J. Sun, R. Cai, L. Yao, and A. Zhang, "Malicious attacks against deep reinforcement learning interpretations," in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, Virtual Event, CA, USA, 2020, pp. 472–482.
- [200] J. Sun, T. Zhang, X. Xie, L. Ma, Y. Zheng, K. Chen, and Y. Liu, "Stealthy and efficient adversarial attacks against deep reinforcement learning," in *Proc. Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, Feb. 2020, pp. 5883–5891.
- [201] A. Rakhsha, G. Radanovic, R. Devidze, X. Zhu, and A. Singla, "Policy teaching via environment poisoning: Training-time adversarial attacks against reinforcement learning," in *Proc. International Conference on Machine Learning (ICML)*, vol. 119, Jul. 2020, pp. 7974–7984.
- [202] A. M. Lonzetta, P. Cope, J. Campbell, B. J. Mohd, and T. Hayajneh, "Security vulnerabilities in bluetooth technology as used in IoT," *Journal of Sensor and Actuator Networks*, vol. 7, no. 3, 2018.
- [203] A. Gleave, M. Dennis, N. Kant, C. Wild, S. Levine, and S. Russell, "Adversarial policies: Attacking deep reinforcement learning," in *Proc. International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, Apr. 2020.
- [204] A. N. Bhagoji, W. He, B. Li, and D. Song, "Practical black-box attacks on deep neural networks using efficient query mechanisms," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- [205] M. Inkawhich, Y. Chen, and H. Li, "Snooping attacks on deep reinforcement learning," in *Proc. International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, Auckland, New Zealand, 2020, pp. 557–565.
- [206] J. Liu, J. Lou, L. Xiong, J. Liu, and X. Meng, "Projected federated averaging with heterogeneous differential privacy," *Proc. VLDB Endow.*, vol. 15, no. 4, p. 828840, Apr. 2022.
- [207] C. Zhang, S. Ekanut, L. Zhen, and Z. Li, "Augmented multi-party computation against gradient leakage in federated learning," *IEEE Transactions on Big Data*, pp. 1–10, 2022.
- [208] R. Kanagavelu, Z. Li, J. Samsudin, Y. Yang, F. Yang, R. S. Mong Goh, M. Cheah, P. Wiwatphonthana, K. Akkarajitsakul, and S. Wang, "Two-phase multi-party computation enabled privacy-preserving federated learning," in *Proc. 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID)*, 2020, pp. 410–419.
- [209] E. Sotthiwat, L. Zhen, Z. Li, and C. Zhang, "Partially encrypted multi-party computation for federated learning," in *Proc. IEEE/ACM 21st International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, 2021, pp. 828–835.
- [210] W. Mou, C. Fu, Y. Lei, and C. Hu, "A verifiable federated learning scheme based on secure multi-party computation," in *Proc. 16th International Conference on Wireless Algorithms, Systems, and Applications (WASA)*, Jun. 2021, pp. 198–209.
- [211] A. S. Shamsabadi, A. Gascn, H. Haddadi, and A. Cavallaro, "PrivEdge: From local to distributed private training and prediction," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3819–3831, 2020.
- [212] C. Xie, M. Chen, P.-Y. Chen, and B. Li, "CRFL: Certifiably robust federated learning against backdoor attacks," in *Proc. International Conference on Machine Learning*, vol. 139, Jul. 2021, pp. 11 372–11 382.
- [213] X. Cao, J. Jia, Z. Zhang, and N. Gong, "FedRecover: Recovering from poisoning attacks in federated learning using historical information," in *Proc. IEEE Symposium on Security and Privacy (SP)*, May 2023, pp. 326–343.
- [214] X. Cao, Z. Zhang, J. Jia, and N. Z. Gong, "FLCert: Provably secure federated learning against poisoning attacks," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3691–3705, 2022.
- [215] Z. Zhang, X. Cao, J. Jia, and N. Z. Gong, "FLDetector: Defending federated learning against model poisoning attacks via detecting malicious clients," in *Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2022, p. 25452555.
- [216] F. Tahmasebian, J. Lou, and L. Xiong, "RobustFed: A truth inference approach for robust federated learning," in *Proc. ACM International Conference on Information & Knowledge Management (CIKM)*, 2022, pp. 1868–1877.
- [217] P. Novák, M. Rollo, J. Hodík, and T. Vlcek, "Communication security in multi-agent systems," in *Proc. International Central and Eastern European Conference on Multi-Agent Systems (CEEMAS)*, vol. 2691, Prague, Czech Republic, Jun. 2003, pp. 454–463.
- [218] T. D. Huynh, "A personalized framework for trust assessment," in *Proc. ACM Symposium on Applied Computing (SAC)*, Honolulu, Hawaii, USA, Mar. 2009, pp. 1302–1307.
- [219] "Ai solution for self-learning emr," 2022. [Online]. Available: <https://www.basebit.me/en/solution1.aspx>
- [220] O. Li, J. Sun, X. Yang, W. Gao, H. Zhang, J. Xie, V. Smith, and C. Wang, "Label leakage and protection in two-party split learning," in *Proc. International Conference on Learning Representations (ICLR)*, 2022.
- [221] S. Abuadbba, K. Kim, M. Kim, C. Thapa, S. A. Çamtepe, Y. Gao, H. Kim, and S. Nepal, "Can we use split learning on 1d CNN models for privacy preserving training?" in *Proc. ACM Asia Conference on Computer and Communications Security (ASIACCS)*, Taipei, Taiwan, Oct. 2020, pp. 305–318.
- [222] H. Weng, J. Zhang, F. Xue, T. Wei, S. Ji, and Z. Zong, "Privacy leakage of real-world vertical federated learning," 2020. [Online]. Available: <https://arxiv.org/abs/2011.09290>
- [223] S. Itahara, T. Nishio, Y. Koda, M. Morikura, and K. Yamamoto, "Distillation-based semi-supervised federated learning for communication-efficient collaborative training with non-IID private data," *IEEE Trans. Mob. Comput.*, Early Access 2021.
- [224] D. Romanini, A. J. Hall, P. Papadopoulos, T. Titcombe, A. Ismail, T. Ceber, R. Sandmann, R. Roehm, and M. A. Hoeh, "PyVertical: A vertical federated learning framework for multi-headed splitNN," in *Proc. ICLR Workshop on Distributed and Private Machine Learning (DPML)*, 2021.

- [225] L. Sun and L. Lyu, "Federated model distillation with noise-free differential privacy," in *Proc. Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, Aug. 2021, pp. 1563–1570.
- [226] E. Erdoğan, A. Kıpççı, and A. E. Çiçek, "UnSplit: Data-oblivious model inversion, model stealing, and label inference attacks against split learning," in *Proc. 21st Workshop on Privacy in the Electronic Society (WPES)*, 2022, p. 115124.
- [227] C. Fu, X. Zhang, S. Ji, J. Chen, J. Wu, S. Guo, J. Zhou, A. X. Liu, and T. Wang, "Label inference attacks against vertical federated learning," in *Proc. 31st USENIX Security Symposium (USENIX Security)*, Aug. 2022, pp. 1397–1414.
- [228] L. Junlin and L. Xinchen, "Clustering label inference attack against practical split learning," *Arxiv*, 2022. [Online]. Available: <https://arxiv.org/abs/2203.05222>
- [229] P. Ye, Z. Jiang, W. Wang, B. Li, and B. Li, "Feature reconstruction attacks and countermeasures of DNN training in vertical federated learning," *arXiv*, 2022. [Online]. Available: <https://arxiv.org/abs/2210.06771>
- [230] P. Qiu, X. Zhang, S. Ji, T. Du, Y. Pu, J. Zhou, and T. Wang, "Your labels are selling you out: Relation leaks in vertical federated learning," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–16, 2022.
- [231] X. Jin, P.-Y. Chen, C.-Y. Hsu, C.-M. Yu, and T. Chen, "CAFE: Catastrophic data leakage in vertical federated learning," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, 2021, pp. 994–1006.
- [232] M. Pathak, S. Rane, and B. Raj, "Multipart differential privacy via aggregation of locally trained classifiers," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2010, pp. 1876–1884.
- [233] Y. Wu, S. Cai, X. Xiao, G. Chen, and B. C. Ooi, "Privacy preserving vertical federated learning for tree-based models," *Proceedings of the VLDB Endowment*, vol. 13, no. 11, pp. 2090–2103, 2020.
- [234] Y. Zhang and H. Zhu, "Additively homomorphical encryption based deep neural network for asymmetrically collaborative machine learning," *Arxiv*, 2020. [Online]. Available: <https://arxiv.org/abs/2007.06849>
- [235] Z. Tian, R. Zhang, X. Hou, J. Liu, and K. Ren, "FederBoost: Private federated learning for GBDT," *Arxiv*, 2020. [Online]. Available: <https://arxiv.org/abs/2011.02796>
- [236] A. Li, Y. Duan, H. Yang, Y. Chen, and J. Yang, "TIPRDC: Task-independent privacy-respecting data crowdsourcing framework for deep learning with anonymized intermediate representations," in *Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, Virtual Event, CA, USA, Aug. 2020, pp. 824–832.
- [237] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, *Federated Learning*. Springer Cham, 2019.
- [238] "Overviewfate," 2022. [Online]. Available: <https://fate.fedai.org/overview/>
- [239] B. Trevizan, J. Chamby-Diaz, A. L. Bazzan, and M. Recamonde-Mendoza, "A comparative evaluation of aggregation methods for machine learning over vertically partitioned data," *Expert Systems with Applications*, vol. 152, 2020.
- [240] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.
- [241] D. Peteiro-Barral and B. Guijarro-Berdiñas, "A Survey of Methods for Distributed Machine Learning," *Progress in Artificial Intelligence*, vol. 2, no. 1, pp. 1–11, 2013.
- [242] P. K. Chan, S. J. Stolfo *et al.*, "Toward parallel and distributed learning by meta-learning," in *Proc. AAAI workshop in Knowledge Discovery in Databases*, 1993, pp. 227–240.
- [243] F. L. da Silva, R. Glatt, and A. H. R. Costa, "Simultaneously learning and advising in multiagent reinforcement learning," in *Proc. International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, São Paulo, Brazil, May 2017, pp. 1100–1108.
- [244] U. F. Minhas, J. Zhang, T. Tran, and R. Cohen, "A multifaceted approach to modeling agent trust for effective communication in the application of mobile ad hoc vehicular networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 3, pp. 407–420, 2011.
- [245] J. Hayes, L. Melis, G. Danezis, and E. D. Cristofaro, "LOGAN: Evaluating privacy leakage of generative models using generative adversarial networks," 2017. [Online]. Available: <http://arxiv.org/abs/1705.07663>
- [246] B. Hilprecht, M. Härterich, and D. Bernau, "Monte carlo and reconstruction membership inference attacks against generative models," *Proceedings on Privacy Enhancing Technologies*, vol. 2019, no. 4, pp. 232–249, 2019.
- [247] D. Chen, N. Yu, Y. Zhang, and M. Fritz, "GAN-leaks: A taxonomy of membership inference attacks against generative models," in *Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS)*, Virtual Event, CA, USA, Nov. 2020, pp. 343–362.
- [248] W. Ren, X. Zhou, J. Liu, and N. B. Shroff, "Multi-armed bandits with local differential privacy," *Arxiv*, 2020. [Online]. Available: <https://arxiv.org/abs/2007.03121>
- [249] J. Zhao, "Distributed deep learning under differential privacy with the teacher-student paradigm," in *Proc. The Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, New Orleans, Louisiana, USA, Feb. 2018, pp. 404–408.
- [250] M. Christopher and B. Franziska, "Personalized PATE: Differential privacy for machine learning with individual privacy guarantees," *Arxiv*, 2022. [Online]. Available: <https://arxiv.org/abs/2202.10517v2>
- [251] Y. Long, S. Lin, Z. Yang, C. A. Gunter, and B. Li, "Scalable differentially private generative student model via PATE," *Arxiv*, 2019. [Online]. Available: <http://arxiv.org/abs/1906.09338>
- [252] M. Abadi, U. Erlingsson, I. Goodfellow, H. B. McMahan, I. Mironov, N. Papernot, K. Talwar, and L. Zhang, "On the protection of private information in machine learning systems: Two recent approaches," in *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, 2017, pp. 1–6.
- [253] X. Tang, F. Zhang, Z. Qin, Y. Wang, D. Shi, B. Song, Y. Tong, H. Zhu, and J. Ye, "Value function is all you need: A unified learning framework for ride hailing platforms," in *Proc. ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD)*, 2021, p. 36053615.
- [254] Y. Wang, Y. Tong, C. Long, P. Xu, K. Xu, and W. Lv, "Adaptive dynamic bipartite graph matching: A reinforcement learning approach," in *Proc. IEEE 35th International Conference on Data Engineering (ICDE)*, 2019, pp. 1478–1489.
- [255] T. Zhu, D. Ye, W. Wang, W. Zhou, and P. S. Yu, "More than privacy: Applying differential privacy in key areas of artificial intelligence," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 6, pp. 2824–2843, 2022.
- [256] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surv.*, vol. 42, no. 4, 2010.
- [257] Y. Cheng *et al.*, "Federated learning for privacy-preserving AI," *Communications of the ACM*, vol. 63, no. 12, pp. 33–36, Nov. 2020.
- [258] C. Ma, J. Li, L. Shi, M. Ding, T. Wang, Z. Han, and H. V. Poor, "When federated learning meets blockchain: A new distributed learning paradigm," *IEEE Computational Intelligence Magazine*, vol. 17, no. 3, pp. 26–33, 2022.
- [259] X. Luo *et al.*, "Feature inference attack on model predictions in vertical federated learning," in *Proc. IEEE International Conference on Data Engineering (ICDE)*, Chania, Greece, Apr. 2021, pp. 181–192.
- [260] H. Hu, Z. Salic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, "Membership inference attacks on machine learning: A survey," *ACM Comput. Surv.*, vol. 54, no. 11s, 2022.
- [261] M. Balunovic, D. I. Dimitrov, R. Staab, and M. T. Vechev, "Bayesian framework for gradient leakage," in *Proc. International Conference on Learning Representations (ICLR)*, Virtual Event, 2022.
- [262] Y. Dong, Q.-A. Fu, X. Yang, T. Pang, H. Su, Z. Xiao, and J. Zhu, "Benchmarking adversarial robustness on image classification," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 318–328.
- [263] C. Xie, K. Huang, P.-Y. Chen, and B. Li, "DBA: Distributed backdoor attacks against federated learning," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.
- [264] K. Wei, J. Li, C. Ma, M. Ding, S. Wei, F. Wu, G. Chen, and T. Ranbaduge, "Covert model poisoning against federated learning: Algorithm design and optimization," *Arxiv*, 2021. [Online]. Available: <https://arxiv.org/abs/2101.11799>
- [265] D. Zhong, H. Sun, J. Xu, N. Gong, and W. H. Wang, "Understanding disparate effects of membership inference attacks and their countermeasures," in *Proc. ACM on Asia Conference on Computer and Communications Security (ASIACCS)*, 2022, p. 959974.
- [266] H. Liu, J. Jia, W. Qu, and N. Z. Gong, "EncoderMI: Membership inference against pre-trained encoders in contrastive learning," in *Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2021, p. 20812095.
- [267] F. Mo, Z. Tarkhani, and H. Haddadi, "SoK: Machine learning with confidential computing," *Arxiv*, 2022. [Online]. Available: <https://arxiv.org/abs/2112.05423>

- [268] E. Bao, Y. Zhu, X. Xiao, Y. Yang, B. C. Ooi, B. H. M. Tan, and K. M. M. Aung, "Skellam mixture mechanism: A novel approach to federated learning with differential privacy," *Proc. VLDB Endow.*, vol. 15, no. 11, pp. 2348–2360, 2022.
- [269] W. Ruan, M. Xu, W. Fang, L. Wang, L. Wang, and W. Han, "Private, efficient, and accurate: Protecting models trained by multi-party learning with differential privacy," in *Proc. IEEE Symposium on Security and Privacy (SP)*, 2022.
- [270] S. Chen and B. Li, "Towards optimal multi-modal federated learning on non-IID data with hierarchical gradient blending," in *Proc. IEEE Conference on Computer Communications (INFOCOM)*, 2022, pp. 1469–1478.
- [271] L. Zong, Q. Xie, J. Zhou, P. Wu, X. Zhang, and B. Xu, "FedCMR: Federated cross-modal retrieval," in *Proc. 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2021, p. 16721676.
- [272] B. Xiong, X. Yang, F. Qi, and C. Xu, "A unified framework for multi-modal federated learning," *Neurocomputing*, vol. 480, pp. 110–118, 2022.
- [273] C. Wu, F. Wu, L. Lyu, Y. Huang, and X. Xie, "Communication-efficient federated learning via knowledge distillation," *Nature Communications*, 2022.
- [274] X. Gong, A. Sharma, S. Karanam, Z. Wu, T. Chen, D. Doermann, and A. Innanje, "Ensemble attention distillation for privacy-preserving federated learning," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 15076–15086.
- [275] L. Zhang, L. Shen, L. Ding, D. Tao, and L.-Y. Duan, "Fine-tuning global model via data-free knowledge distillation for non-IID federated learning," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10174–10183.
- [276] J. Li, Y. Shao, K. Wei, M. Ding, C. Ma, L. Shi, Z. Han, and H. V. Poor, "Blockchain assisted decentralized federated learning (BLADE-FL): Performance analysis and resource allocation," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 10, pp. 2401–2415, 2022.
- [277] D. C. Nguyen, S. Hosseinalipour, D. J. Love, P. N. Pathirana, and C. G. Brinton, "Latency optimization for blockchain-empowered federated learning in multi-server edge computing," *IEEE Journal on Selected Areas in Communications*, Early Access 2022.
- [278] X. Deng, J. Li, C. Ma, K. Wei, L. Shi, M. Ding, W. Chen, and H. V. Poor, "Blockchain assisted federated learning over wireless channels: Dynamic resource allocation and client scheduling," *IEEE Transactions on Wireless Communications*, Early Access 2022.
- [279] L. Cui, X. Su, and Y. Zhou, "A fast blockchain-based federated learning framework with compressed communications," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 12, pp. 3358–3372, 2022.
- [280] Q. Zhang *et al.*, "AsySQN: Faster vertical federated learning algorithms with better computation resource utilization," in *Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, Virtual Event, Singapore, Aug. 2021, pp. 3917–3927.
- [281] F. Fu *et al.*, "VF²Boost: Very fast vertical federated gradient boosting for cross-enterprise learning," in *Proc. International Conference on Management of Data (SIGMOD)*, Virtual Event, China, Jun. 2021, pp. 563–576.
- [282] K. Wei, J. Li, C. Ma, M. Ding, S. Wei, F. Wu, G. Chen, and T. Ranbaduge, "Vertical federated learning: Challenges, methodologies and experiments," *Arxiv*, 2022. [Online]. Available: <https://arxiv.org/abs/2202.04309>
- [283] N. Agarwal, A. T. Suresh, F. X. Yu, S. Kumar, and B. McMahan, "CPSGD: Communication-efficient and differentially-private distributed SGD," in *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, Montréal, Canada, Dec. 2018, pp. 7575–7586.
- [284] L. Yu, L. Liu, C. Pu, M. E. Gursoy, and S. Truex, "Differentially private model publishing for deep learning," in *Proc. IEEE Symposium on Security and Privacy (SP)*, San Francisco, CA, USA, May 2019, pp. 332–349.
- [285] Y. Nie, W. Yang, L. Huang, X. Xie, Z. Zhao, and S. Wang, "A utility-optimized framework for personalized private histogram estimation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 4, pp. 655–669, 2019.
- [286] X. Gu, M. Li, L. Xiong, and Y. Cao, "Providing input-discriminative protection for local differential privacy," in *Proc. IEEE 36th International Conference on Data Engineering (ICDE)*, 2020, pp. 505–516.
- [287] Y. Tao, S. Chen, F. Li, D. Yu, J. Yu, and H. Sheng, "A distributed privacy-preserving learning dynamics in general social networks," *Arxiv*, 2020. [Online]. Available: <https://arxiv.org/abs/2011.09845>
- [288] R. W. Baldwin and W. C. Gramlich, "Cryptographic protocol for trustable match making," in *Proc. IEEE Symposium on Security and Privacy (SP)*, 1985, pp. 92–92.
- [289] C. Hazay and Y. Lindell, "Constructions of truly practical secure protocols using standards smartcards," in *Proc. ACM Conference on Computer and Communications Security (CCS)*, 2008, pp. 491–500.
- [290] C. Meadows, "A more efficient cryptographic matchmaking protocol for use in the absence of a continuously available third party," in *Proc. IEEE Symposium on Security and Privacy (SP)*, 1986, pp. 134–134.
- [291] M. J. Freedman, K. Nissim, and B. Pinkas, "Efficient private matching and set intersection," in *Proc. Advances in Cryptology (EUROCRYPT)*, Berlin, Heidelberg, 2004, pp. 1–19.
- [292] Y. Huang, D. Evans, and J. Katz, "Private set intersection: Are garbled circuits better than custom protocols?" in *Proc. Network and Distributed System Security Symposium (NDSS)*, 2012.
- [293] C. Dong, L. Chen, and Z. Wen, "When private set intersection meets big data: An efficient and scalable protocol," in *Proc. ACM SIGSAC Conference on Computer Communications Security (CCS)*, 2013, pp. 789–800.
- [294] J. Li and J. Li, "Brief paper-adaptive iterative learning control for consensus of multi-agent systems," *IET Control Theory & Applications*, vol. 7, no. 1, pp. 136–142, 2013.
- [295] D. Meng and Y. Jia, "Finite-time consensus for multi-agent systems via terminal feedback iterative learning," *IET control theory & applications*, vol. 5, no. 18, pp. 2098–2110, 2011.
- [296] —, "Iterative learning approaches to design finite-time consensus protocols for multi-agent systems," *Systems & Control Letters*, vol. 61, no. 1, pp. 187–194, 2012.
- [297] T. D. Nguyen, L. H. Pham, and J. Sun, "SGUARD: Towards fixing vulnerable smart contracts automatically," in *Proc. IEEE Symposium on Security and Privacy (SP)*, San Francisco, CA, USA, 2021, pp. 1215–1229.
- [298] D. C. Nguyen, M. Ding, Q.-V. Pham, P. N. Pathirana, L. B. Le, A. Seneviratne, J. Li, D. Niyato, and H. V. Poor, "Federated learning meets blockchain in edge computing: Opportunities and challenges," *IEEE Internet of Things Journal*, vol. 8, no. 16, pp. 12806–12825, 2021.
- [299] S. Awan, F. Li, B. Luo, and M. Liu, "Poster: A reliable and accountable privacy-preserving federated learning framework using the blockchain," in *Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2019, pp. 2561–2563.
- [300] Y. Zhan and J. Zhang, "An incentive mechanism design for efficient edge learning by deep reinforcement learning approach," in *Proc. IEEE Conference on Computer Communications (INFOCOM)*, 2020, pp. 2489–2498.
- [301] R. H. L. Sim, Y. Zhang, M. C. Chan, and B. K. H. Low, "Collaborative machine learning with incentive-aware model rewards," in *Proc. International Conference on Machine Learning (ICML)*, vol. 119, Jul. 2020, pp. 8927–8936.
- [302] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and robust federated learning through personalization," in *Proc. International Conference on Machine Learning (ICML)*, Jul. 2021, pp. 6357–6368.
- [303] I. Bistriz, T. Baharav, A. Leshem, and N. Bambos, "My fair bandit: Distributed learning of max-min fairness with multi-player bandits," in *Proc. International Conference on Machine Learning (ICML)*, vol. 119, Jul. 2020, pp. 930–940.
- [304] L. Lyu, X. Xu, Q. Wang, and H. Yu, *Collaborative Fairness in Federated Learning*. Cham: Springer International Publishing, 2020.
- [305] H. Hasan, T. Salah, D. Shehadeh, M. J. Zemerly, C. Y. Yeun, M. Al-Qutayri, and Y. Al-Hammadi, "Secure lightweight ECC-based protocol for multi-agent IoT systems," in *Proc. IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, Rome, Italy, Oct. 2017, pp. 1–8.
- [306] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Communications Magazine*, vol. 58, no. 1, pp. 106–112, 2020.
- [307] L. Deng, G. Li, S. Han, L. Shi, and Y. Xie, "Model compression and hardware acceleration for neural networks: A comprehensive survey," *Proceedings of the IEEE*, vol. 108, no. 4, pp. 485–532, 2020.
- [308] Y. Liu, Z. Ma, X. Liu, S. Ma, and K. Ren, "Privacy-preserving object detection for medical images with faster R-CNN," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 69–84, 2022.
- [309] J. Chao, A. A. Badawi, B. Unnikrishnan, J. Lin, C. F. Mun, J. M. Brown, J. P. Campbell, M. F. Chiang, J. Kalpathy-Cramer, V. R.

- Chandrasekhar, P. Krishnaswamy, and K. M. M. Aung, "CaRENets: Compact and resource-efficient CNN for homomorphic inference," 2019. [Online]. Available: <http://arxiv.org/abs/1901.10074>
- [310] J. Feng, L. T. Yang, Q. Zhu, and K.-K. R. Choo, "Privacy-preserving tensor decomposition over encrypted data in a federated cloud environment," *IEEE Transactions on Dependable and Secure Computing*, vol. 17, no. 4, pp. 857–868, 2020.
- [311] R. Zhang and P. Venkitasubramaniam, "Optimal local differentially private quantization," *IEEE Transactions on Signal Processing*, vol. 68, pp. 6509–6520, 2020.
- [312] Z. Luo, D. J. Wu, E. Adeli, and L. Fei-Fei, "Scalable differential privacy with sparse network finetuning," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 5059–5068.
- [313] D. Quarta, M. Pogliani, M. Polino, F. Maggi, A. M. Zanchettin, and S. Zanero, "An experimental security analysis of an industrial robot controller," in *Proc. IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 268–286.
- [314] B. Breiling, B. Dieber, and P. Schartner, "Secure communication for the robot operating system," in *Proc. Annual IEEE International Systems Conference (SysCon)*, 2017, pp. 1–6.
- [315] K. Jokinen and G. Wilcock, "Do you remember me? Ethical issues in long-term social robot interactions," in *Proc. IEEE International Conference on Robot Human Interactive Communication (RO-MAN)*, 2021, pp. 678–683.
- [316] A. Sharkey and N. Sharkey, "Granny and the robots: Ethical issues in robot care for the elderly," *Ethics and information technology*, vol. 14, no. 1, pp. 27–40, 2012.
- [317] D. Adams, A. Bah, C. Barwulor, N. Musaby, K. Pitkin, and E. M. Redmiles, "Ethics emerging: The story of privacy and security perceptions in virtual reality," in *Proc. Fourteenth Symposium on Usable Privacy and Security (SOUPS)*, Baltimore, MD, Aug. 2018, pp. 427–442.
- [318] A. Gulhane, A. Vyas, R. Mitra, R. Oruche, G. Hoefer, S. Valluripally, P. Calyam, and K. A. Hoque, "Security, privacy and safety risk assessment for virtual reality learning environment applications," in *Proc. IEEE Annual Consumer Communications Networking Conference (CCNC)*, 2019, pp. 1–9.
- [319] D. Maloney, S. Zamanifard, and G. Freeman, "Anonymity vs. familiarity: Self-disclosure and privacy in social virtual reality," in *Proc. ACM Symposium on Virtual Reality Software and Technology (VRST)*, 2020, pp. 1–9.
- [320] M. Roetteler and K. M. Svore, "Quantum computing: Codebreaking and beyond," *IEEE Security & Privacy*, vol. 16, no. 5, pp. 22–36, 2018.
- [321] K. A. Fisher, A. Broadbent, L. Shalm, Z. Yan, J. Lavoie, R. Prevedel, T. Jennewein, and K. J. Resch, "Quantum computing on encrypted data," *Nature communications*, vol. 5, no. 1, pp. 1–7, 2014.
- [322] Q. Yang, Y. Zhao, H. Huang, Z. Xiong, J. Kang, and Z. Zheng, "Fusing blockchain and AI with metaverse: A survey," *IEEE Open Journal of the Computer Society*, vol. 3, pp. 122–136, 2022.
- [323] Y. Jiang, J. Kang, D. Niyato, X. Ge, Z. Xiong, C. Miao, and X. Shen, "Reliable distributed computing for metaverse: A hierarchical game-theoretic approach," *IEEE Transactions on Vehicular Technology*, pp. 1–16, 2022.
- [324] Y. Lu, X. Huang, K. Zhang, S. Maharjan, and Y. Zhang, "Communication-efficient federated learning for digital twin edge networks in industrial IoT," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 8, pp. 5709–5718, 2021.
- [325] R. Dong, C. She, W. Hardjawana, Y. Li, and B. Vucetic, "Deep learning for hybrid 5G services in mobile edge computing systems: Learn from a digital twin," *IEEE Transactions on Wireless Communications*, vol. 18, no. 10, pp. 4692–4707, 2019.
- [326] E. Karaarslan and M. Babiker, "Digital twin security threats and countermeasures: An introduction," in *Proc. International Conference on Information Security and Cryptology (ISCURKEY)*, 2021, pp. 7–11.
- [327] C. Chen, L. Zhang, Y. Li, T. Liao, S. Zhao, Z. Zheng, H. Huang, and J. Wu, "When digital economy meets web3.0: Applications and challenges," *IEEE Open Journal of the Computer Society*, vol. 3, pp. 233–245, 2022.
- [328] C. Huang, J. H. Lim, and A. C. Courville, "A variational perspective on diffusion-based generative models and score matching," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, Virtual, 2021, pp. 22 863–22 876.



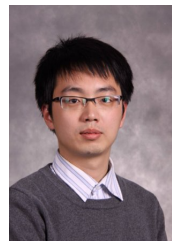
Chuan Ma (M'19) received the B.S. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2013 and the Ph.D. degree from the University of Sydney, Australia, in 2018. From 2018 to 2022, he worked as a lecturer at the Nanjing University of Science and Technology, and now he is a principal investigator at Zhejiang Lab, Hangzhou, China. He has published more than 40 journal and conference papers, including the best paper in WCNC 2018, and the best paper award in IEEE Signal Processing Society 2022. His research interests include stochastic geometry, wireless caching networks, and distributed machine learning, and now focuses on big data analysis and privacy-preserving.



Jun Li (SM'16) received Ph. D degree in Electronic Engineering from Shanghai Jiao Tong University, Shanghai, P. R. China in 2009. From January 2009 to June 2009, he worked in the Department of Research and Innovation, Alcatel Lucent Shanghai Bell as a Research Scientist. From June 2009 to April 2012, he was a Postdoctoral Fellow at the School of Electrical Engineering and Telecommunications, at the University of New South Wales, Australia. From April 2012 to June 2015, he was a Research Fellow at the School of Electrical Engineering, the University of Sydney, Australia. From June 2015 to now, he is a Professor at the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China. He was a visiting professor at Princeton University from 2018 to 2019. His research interests include network information theory, game theory, distributed intelligence, multiple agent reinforcement learning, and their applications in ultra-dense wireless networks, mobile edge computing, network privacy and security, and the industrial Internet of things. He has co-authored more than 200 papers in IEEE journals and conferences and holds 1 US patent and more than 10 Chinese patents in these areas. He is serving as an editor of IEEE Transactions on Wireless Communication and TPC member for several flagship IEEE conferences.

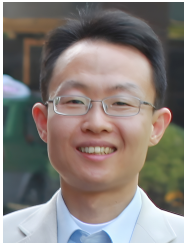


Kang Wei (M'23) received his Ph.D. degree from Nanjing University of Science and Technology. Before that, he received the B.S. degree in information engineering from Xidian University, Xian, China, in 2014. He is currently a postdoctoral fellow at The Hong Kong Polytechnic University. He mainly focuses on privacy protection and optimization techniques for edge intelligence, including federated learning, differential privacy, and network resource allocation.



Bo Liu received the BEng degree from the Department of Computer Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing, China, in 2004. He then received the MEng. and Ph.D. Degrees from the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2007 and 2010, respectively. He is currently an Associate Professor at the University of Technology Sydney, Australia. His research interests include cybersecurity and privacy, location privacy and image privacy, privacy protection, and machine

learning.



Ming Ding (IEEE M'12-S'17) received the B.S. and M.S. degrees (with first-class Hons.) in electronics engineering from Shanghai Jiao Tong University (SJTU), Shanghai, China, and the Doctor of Philosophy (Ph.D.) degree in signal and information processing from SJTU, in 2004, 2007, and 2011, respectively. From April 2007 to September 2014, he worked at Sharp Laboratories of China in Shanghai, China as a Researcher/Senior Researcher/Principal Researcher. Currently, he is a Principal Research Scientist at Data61, CSIRO, in Sydney, NSW, Australia.

His research interests include information technology, data privacy and security, and machine learning and AI. He has authored more than 200 papers in IEEE journals and conferences, all in recognized venues, and around 20 3GPP standardization contributions, as well as two books, i.e., "Multi-point Cooperative Communication Systems: Theory and Applications" (Springer, 2013) and Fundamentals of Ultra-Dense Wireless Networks (Cambridge University Press, 2022). Also, he holds 21 US patents and has co-invented another 100+ patents on 4G/5G technologies. Currently, he is an editor of IEEE Transactions on Wireless Communications and IEEE Communications Surveys and Tutorials. Besides, he has served as a guest editor/co-chair/co-tutor/TPC member for multiple IEEE top-tier journals/conferences and received several awards for his research work and professional services, including the prestigious IEEE Signal Processing Society Best Paper Award in 2022.



H. Vincent Poor (S'72, M'77, SM'82, F'87) received the Ph.D. degree in EECS from Princeton University in 1977. From 1977 until 1990, he was on the faculty of the University of Illinois at Urbana-Champaign. Since 1990 he has been on the faculty at Princeton, where he is currently the Michael Henry Strater University Professor. From 2006 to 2016, he served as the dean of Princeton's School of Engineering and Applied Science. He has also held visiting appointments at several other universities, including most recently at Berkeley and Cambridge.

His research interests are in the areas of information theory, machine learning, and network science, and their applications in wireless networks, energy systems, and related fields. Among his publications in these areas is the recent book *Machine Learning and Wireless Communications*. (Cambridge University Press, 2022). Dr. Poor is a member of the National Academy of Engineering and the National Academy of Sciences and is a foreign member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies. He received the IEEE Alexander Graham Bell Medal in 2017.



Long Yuan is currently a professor at the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. He received his Ph.D. degree from the database group of the University of New South Wales, Australia, M.S. degree and B.S. degree both from Sichuan University, China. His research interests include graph data management and analysis. He has published papers in conferences and journals including VLDB, ICDE, WWW, The VLDB Journal, and TKDE.



Zhu Han (S'01CM'04-SM'09-F'14) received the B.S. degree in electronic engineering from Tsinghua University, in 1997, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, in 1999 and 2003, respectively. From 2000 to 2002, he was an R&D Engineer of JDSU, Germantown, Maryland. From 2003 to 2006, he was a Research Associate at the University of Maryland. From 2006 to 2008, he was an assistant professor at Boise State University, Idaho. Currently, he is a John and Rebecca Moores

Professor in the Electrical and Computer Engineering Department as well as in the Computer Science Department at the University of Houston, Texas. His research interests include wireless resource allocation and management, wireless communications and networking, game theory, big data analysis, security, and smart grid. Dr. Han received an NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011, the EURASIP Best Paper Award for the Journal on Advances in Signal Processing in 2015, IEEE Leonard G. Abraham Prize in the field of Communications Systems (best paper award in IEEE JSAC) in 2016, and several best paper awards in IEEE conferences. Dr. Han was an IEEE Communications Society Distinguished Lecturer from 2015-2018, AAAS fellow since 2019, and ACM distinguished Member since 2019. Dr. Han is a 1% highly cited researcher since 2017 according to Web of Science. Dr. Han is also the winner of the 2021 IEEE Kiyo Tomiyasu Award, for outstanding early to mid-career contributions to technologies holding the promise of innovative applications, with the following citation: "for contributions to game theory and distributed management of autonomous communication networks."