Future Generation Computer Systems 148 (2023) 408–424

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Review article

Differential privacy in deep learning: Privacy and beyond

Yanling Wang [a,b,c], Qian Wang [b,*], Lingchen Zhao [b], Cong Wang [c]

[a] Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, No. 299 Bayi Road, Wuhan, 430072, Hubei, China
[b] School of Cyber Science and Engineering, Wuhan University, No. 299 Bayi Road, Wuhan, 430072, Hubei, China
[c] Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, 999077, Hong Kong, China

ARTICLE INFO

Article history:
Received 29 November 2022
Received in revised form 25 March 2023
Accepted 9 June 2023
Available online 12 June 2023

Keywords:
Deep learning
Differential privacy
Stochastic gradient descent
Lower bound
Fairness
Robustness

ABSTRACT

Motivated by the security risks of deep neural networks, such as various membership and attribute inference attacks, differential privacy has emerged as a promising approach for protecting the privacy of neural networks. As a result, it is crucial to investigate the frontier intersection of differential privacy and deep learning, which is the main motivation behind this survey. Most of the current research in this field focuses on developing mechanisms for combining differentially private perturbations with deep learning frameworks. We provide a detailed summary of these works and analyze potential areas for improvement in the near future. In addition to privacy protection, differential privacy can also play other critical roles in deep learning, such as fairness, robustness, and prevention of over-fitting, which have not been thoroughly explored in previous research. Accordingly, we also discuss future research directions in these areas to offer practical suggestions for future studies.

Contents

[*] Corresponding author.
E-mail addresses: yanling@whu.edu.cn (Y. Wang), qianwang@whu.edu.cn (Q. Wang), lczhaocs@whu.edu.cn (L. Zhao), congwang@cityu.edu.hk (C. Wang).

https://doi.org/10.1016/j.future.2023.06.010
0167-739X/© 2023 Elsevier B.V. All rights reserved.

1. Introduction

A vast amount of personal information is collected and analyzed by enterprises and organizations to provide more personalized services or to access integrated information. During the processing of these raw data, deep learning, which brings the machine closer to the way of human thinking to learn the underlying laws of sample data, is one of the most famous feature investigation methods [1–3]. However, the use of personal data in deep learning can pose legal risks due to the potential for information leakage [4–7]. One notorious attack that leads to information leakage is the membership inference attack [8]. Differential privacy (DP), a privacy constraint originally applied to query results [9,10], is becoming an increasingly important tool in deep learning, as it offers a way to protect personal data while still allowing organizations to extract valuable insights. By adding noise to query results, DP ensures that no single piece of data can be traced back to an individual, making it much harder for attackers to carry out privacy-invasive attacks [11]. Such a measurable privacy mechanism is exactly what deep learning desires, so researchers have tried to think of deep learning as a complex data query to allow DP to demonstrate its effectiveness in the process. There has been a growing amount of research on DP with deep learning in both academia [12–17] and industry [18,19]. For the latter, Google, as a noteworthy example, has provided DP support in Tensorflow, one of the most common open-source frameworks for machine learning [19].

The implications of DP in deep learning also go beyond privacy guarantees, such as fairness, over-fitting, and robustness, which are essential but have rarely been discussed in previous works. These effects are not all positive but can help us understand more about the role of DP in neural networks. In terms of fairness, initial research suggested that DP could exacerbate network unfairness [16]. However, subsequent work has shown that this problem can be mitigated with appropriate techniques [20,21]. Regarding over-fitting, DP has been shown to significantly improve the generalization performance of neural networks, which can help mitigate over-fitting [22–24]. For robustness, DP gave neural networks a certificated scheme against adversarial attacks, which ignored the specific manner the attacks are carried out [15, 25,26]. The model with DP noise is shown to be more robust to poisoning attacks [27]. Additionally, DP noise can make poisoning attacks more covert [27–29]. Overall, the implications of DP in deep learning are wide-ranging, and researchers should carefully consider these effects in their work. This survey provides a detailed and in-depth analysis of the existing research and offers recommendations for future work in this area.

1.1. Privacy risks in deep learning

Neural networks are known to process voluminous amounts of user data, thereby posing a significant privacy risk. Malicious adversaries can gain access to sensitive user information (e.g., vehicle trajectory privacy [30,31], power system privacy [32], and spatial data privacy [33]). This risk is especially severe when adversaries possess prior knowledge of the dataset and have API access to the neural network model [34,35]. The following are two primary forms of attacks on neural networks that are linked to DP.

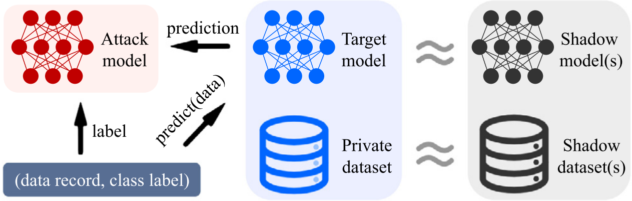


Fig. 1. Membership inference attack in the black-box setting [8]. The adversary submits a record to the target model, which generates a prediction vector consisting of the probabilities of the record being associated with each class. The prediction vector, along with the true label of the record, is then utilized to develop an attack model that determines whether the record belongs to the training dataset. However, in real-world scenarios, obtaining the prediction vector of the target model is typically challenging. Consequently, this approach entails training shadow models that emulate the behavior of the target model. The training data used for the shadow models and the target model are separate but have the same underlying distribution.

Membership Inference Attack [8]. Membership inference attack is a prominent privacy-invasive method that is specifically targeted by DP. Its main objective is to identify if a specific record is included in the training dataset of a neural network. Typically, this attack is conducted through the use of shadow models in a black-box setting, as illustrated in Fig. 1. By training a set of shadow models, adversaries can achieve similar decision boundaries as the target model and, subsequently, obtain an attack model that can infer the membership of a specific record in the training dataset. While membership inference attacks can be carried out without the use of shadow models, their precision rates are often lower than those that employ shadow models [36].

Attribute Inference Attack [37]. Attribute inference attacks, which are also referred to as model inversion attacks [38], are a type of attack used to infer sensitive attributes of an individual from a trained model [39]. These attacks enable adversaries to recover hidden sensitive attributes in a dataset. In this scenario, adversaries with access to the non-sensitive attributes of a record manage to learn their mathematical correlation to the model output. They then use this correlation to infer sensitive attributes by maximizing the posterior probability estimates.

1.2. Differential privacy with deep learning

Differential Privacy is a privacy-preserving technique that aims to mitigate the risk of privacy leakage resulting from small changes in a dataset [9,11]. For instance, if Alice's COVID-19

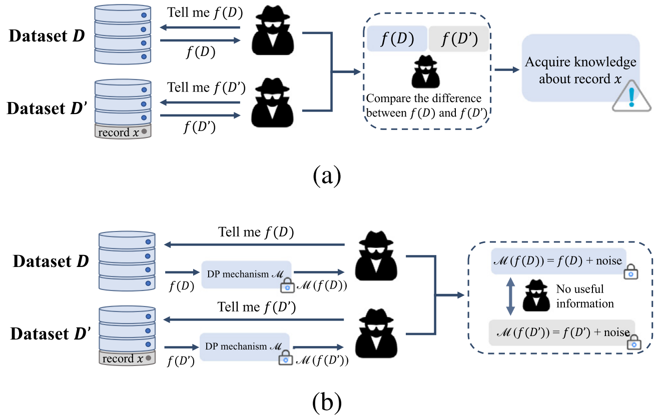


Fig. 2. An example of a DP mechanism. Datasets D and D' are neighboring datasets. (a) The adversary sent query instructions to D and D' and got the exact return value, thereby obtaining private information. (b) The adversary sent query instructions to D and D'. The DP algorithm has processed the exact return value, so the adversary obtains ambiguous results without inferring private information.

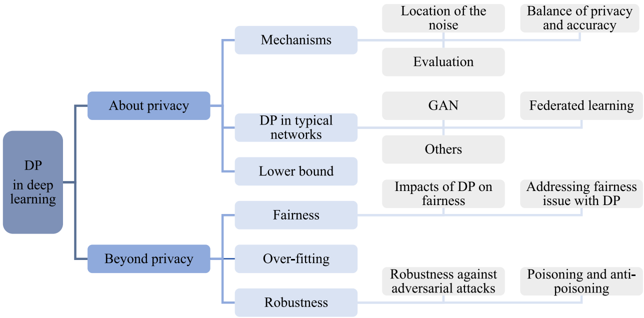


Fig. 3. Hierarchal diagram of this survey.

diagnosis was recorded in a hospital dataset after 20 other cases were already recorded, an adversary can infer Alice's diagnosis by submitting two queries to the dataset, one before and one after Alice's registration. Datasets before and after Alice's registration are called neighboring datasets. Neighboring datasets are defined as two datasets that differ by only one record, and one can obtain the neighboring dataset by adding, deleting, or changing a single record in the original dataset. Queries submitted to neighboring datasets may lead to accurate return values, allowing adversaries to infer sensitive information about the different records in the neighboring dataset, as in Fig. 2. To prevent such attacks, DP aims to obfuscate the accurate return value, ensuring that any response from the dataset includes added noise, thereby preventing the adversary from learning the exact private information. The definition and properties of DP will be presented in Section 2.

Prior to its use in deep learning, extensive research has been conducted on how to apply DP to statistical machine learning tasks and algorithms, including clustering [40–43], Combinatorial Public Projects (CPP) problem [44], network diagrams [45–48], natural language processing (NLP) [49,50], geometric concepts [51], equivalence testing [52], pairwise learning [53], online learning [54–57], ensemble learning [58], boosting [59,60], Empirical Risk Minimization (ERM) [61,62], sparse linear regression [63], sliced Wasserstein distance [64], etc.

Given the severity of membership inference and attribute inference attacks, adding noise to the response of a neural network is an effective defense strategy to obfuscate the results (just like adding noise to the return value of a query). The DP framework offers a useful approach to implement such defenses in protecting deep learning neural networks from privacy threats [65–69]. To assess the privacy guarantees of neural networks, researchers are concerned about how much privacy guarantees are actually achieved by neural networks. Thus they focused on two main parameters [70]: the upper and lower bound. The former represents the most severe privacy leakage that a neural network can have theoretically, while the latter refers to the private information that an adversary can actually obtain (that is, the actual privacy leakage that occurs). Researchers are interested in exploring whether the upper bound of DP is tight and how the lower bound varies among real-world adversaries of different abilities. Two recent works have provided a comprehensive analysis of these questions. Jayaraman et al. [13] showed that neural networks could only achieve acceptable privacy guarantees when the privacy budget ϵ is in the range of $(0, 1)$. However, to achieve acceptable utility, ϵ needs to be much greater than 1 (e.g., $\epsilon = 100$). This gap makes it impossible for the neural network to achieve both utility and privacy simultaneously. Nevertheless, Nasr et al. [17] argued that the lower bounds in [13] are sufficient because the adversary has too many unrealistic capabilities. After limiting the adversary's capabilities according to the actual situation, the network can achieve a large degree of privacy protection, even if ϵ is large.

In addition to privacy concerns, it is important to consider the potential side effects (both good and bad) that may arise from adding noise to a model. These effects broadly include:

- Fairness: Bagdasaryan et al. [16] discovered that adding noise can lead to the "weaker of the weak" phenomenon, where subgroups with poor performance accuracy are more adversely affected by DP noise. This represents that DP noise increases the unfairness of the network. Subsequent research has been carried out on the fairness problem in networks [20,71]. They found that this unfairness is significantly mitigated under special fine-tuning [21].
- Over-fitting: It represents a reduction in the generalization ability of the neural network, which can be mitigated by injecting well-designed noise, such as DP noise [22].
- Robustness: DP can improve the robustness of neural networks against adversarial attacks by preventing small changes in input data from disproportionately impacting the network [15]. In addition to the ability to defend against adversarial attacks, DP also performs well in detecting poisoning attacks by treating poisoned data as "outliers" of the input data [27]. On the other hand, there is also research on using DP for more covert poisoning attacks [29].

Our discussion will proceed from the perspectives of privacy and beyond privacy, as shown in Fig. 3.

1.3. Related work

Most previous surveys investigating this topic have primarily focused on analyzing DP mechanisms. In this scope, some researchers focused on the mechanism itself: Ji et al. [72] analyzed the algorithms of the combination of DP on machine learning and verified the upper bounds of loss functions for DP algorithms. Zhao et al. [66] classified the layers that DP deployed in the network (input, hidden, and output layer). Other surveys have focused on specific neural networks and application scenarios, exploring the potential applications of DP in various fields: Fan [73] studied differentially private generative adversarial network (GAN), and Zhao et al. [74] investigated the potential applications of local differential privacy (LDP) in protecting the privacy of the internet of connected vehicles (IoV). These papers mainly focused on the accuracy and consumed budget of DP mechanisms.

One recent work [68] has discussed the effects of DP on the stability and fairness of machine learning. Motivated by yet different from their work, we focus on deep learning and explore the benefits of DP in enhancing the robustness of neural networks against adversarial and poisoning attacks. Moreover, we suggest that DP could be used for more covert poisoning attacks. Additionally, we summarize solutions proposed to mitigate the unfairness in DP, which causes subgroups with poor performance accuracy to be more adversely affected by DP noise, leading to unfairness in the network [16].

1.4. Contributions

In this paper, we present a comprehensive evaluation of the existing research on DP with deep learning. Our analysis covers three primary aspects:

- Mechanisms: from the mechanism perspective, we conduct a more in-depth analysis than previous surveys and categorize the existing DP mechanisms' improvements into three categories: (1) Designing noise addition mechanisms independent of the training epoch; (2) Improving the method of calculating the privacy budget; and (3) Fine-tuning the size of noise on different parameters according to the network's actual situation. Additionally, we discuss the fine-tuning of DP in various application scenarios and suggest future improvements in this area.
- Upper and lower bounds: as a supplement to previous surveys, we provide a more comprehensive analysis by focusing on both the upper and lower bounds of DP. The latter represents the level of privacy protection achieved by networks under actual attacks. We summarize the reasons for the apparent gap between the lower and upper bounds and emphasize that analyzing the lower bound is as crucial as studying the upper bound.
- Beyond privacy: apart from privacy, we also summarize other effects of DP, including fairness, over-fitting, and robustness.

The remainder of this paper is organized as follows. Section 2 introduces the related background knowledge. Section 3 analyzes work on privacy protection. Section 4 lists the role of DP beyond privacy guarantee. Section 5 concludes our work and briefly discusses the possible future work.

2. Preliminary

This section provides an overview of the theoretical foundations of differential privacy, including its mathematical definition and fundamental properties.

2.1. Definition of differential privacy

Differential privacy is a rigorous privacy-preserving algorithm that obfuscates the output by adding noise to the data and has gained prominence in the field of data mining privacy protection [75–77]. The foundational work of Dwork et al. [11,78] introduced the following mathematical definitions.

Definition 1 (*Neighboring Datasets*)**.** $d(D, D')$ denotes the distance between dataset D and dataset D', which means the minimum number of sample changes that are required to change D into D'. Dataset D and Dataset D' are neighboring datasets if:

$$d(D, D') = 1. \tag{1}$$

The definition of neighboring datasets in the context of DP is that only one record of a dataset D and its neighbor dataset D' is different. A neighboring dataset can be obtained by adding, deleting, or changing a single record in the original dataset.

Definition 2 (*(ϵ, δ)-DP*)**.** Given a pair of neighboring datasets D and D', for every set of outcomes $\mathbf{S}$, a mechanism $\mathcal{M}$ satisfies DP if the following holds:

$$Pr[\mathcal{M}(D) \in \mathbf{S}] \leq Pr[\mathcal{M}(D') \in \mathbf{S}] \times e^{\epsilon} + \delta, \tag{2}$$

where ϵ is the privacy budget ($\epsilon \geq 0$), which represents the degree of privacy protection. δ is the failure probability, which represents the degree of relaxation. A smaller value of ϵ and δ indicates a higher level of privacy protection. When δ is set to 0, privacy protection is stricter. A smaller value of ϵ implies a stronger privacy guarantee, which makes it more challenging for an adversary to infer information about the dataset. However, the introduction of excessive noise to the query output results in reduced accuracy due to increased variability around the correct center value.

Definition 3 (*Rényi DP (RDP) [79]*)**.** In addition to the above traditional definition, DP can also be defined by the concept of Rényi divergence, which is the basis of Rényi DP. For two probability distributions $\mathcal{P}$ and $\mathcal{Q}$ defined over $\mathcal{R}$, the Rényi divergence of order $\alpha > 1$ is defined as ($\mathcal{P}(x)$ is the probability density of $\mathcal{P}$ at x):

$$D_{\alpha}(\mathcal{P} \| \mathcal{Q}) \triangleq \frac{1}{\alpha - 1} \log E_{x \sim \mathcal{Q}} \left(\frac{\mathcal{P}(x)}{\mathcal{Q}(x)} \right)^{\alpha}. \tag{3}$$

On the basis of this, when we set $\alpha = \infty$, a new definition in the form of DP is obtained. A randomized mechanism $\mathcal{M}$ is said to have ϵ-Rényi DP of order α (α, ϵ)-RDP), if for any adjacent dataset D and dataset D':

$$D_{\alpha}(\mathcal{M}(D) \| \mathcal{M}(D')) \leq \epsilon. \tag{4}$$

Mechanism $\mathcal{M}$ satisfies $(\epsilon + \frac{\log 1/\delta}{\alpha - 1}, \delta)$-DP for any $0 < \delta < 1$. Rényi DP bounds one moment at a time, providing a quantitatively accurate way to perform a precise analysis of privacy loss and thus can provide a tight upper bound guarantee on the privacy budget for neural networks.

2.2. Basic DP mechanisms

Laplace mechanism [80]. The Laplace mechanism is a commonly employed technique for preserving DP in numerical datasets, wherein random noise is added to the numerical results. The Laplace mechanism provides a stringent guarantee of (ϵ, 0)-DP.

Definition 4 (*Sensitivity of Laplace Mechanism*)**.** Given a pair of neighboring datasets D and D', the sensitivity is the maximum variation range of the query result function:

$$\Delta f = \max_{D,D'} \|f(D) - f(D')\|_1, \tag{5}$$

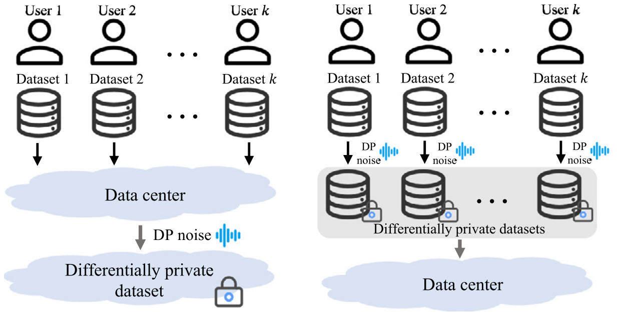
where $\| * \|_1$ represents ℓ_1-norm.

Definition 5 (*Laplace Mechanism*)**.**

$$\mathcal{M}(D) = f(D) + Y, \tag{6}$$

where $f(D)$ represents the query function, $\mathcal{M}(D)$ represents the final return value, and Y represents Laplace random noise, which satisfies $Y \sim L(0, \frac{\Delta f}{\epsilon})$. $\mathcal{M}$ satisfies (ϵ,0)-DP.

Gaussian mechanism [81]. The Gaussian mechanism is another common approach for satisfying DP by adding random noise to data.



(a) Centralized differential privacy (b) Local differential privacy

Fig. 4. The difference between centralized differential privacy and local differential privacy. In the former, users send private data to the data center, where the dataset is then noised. In local differential privacy, data owners add noise to their data before uploading it to the data center.

Definition 6 (*Sensitivity of Gaussian Mechanism*)**.**

$$\Delta f = \max_{D,D'} \|f(D) - f(D')\|_2, \tag{7}$$

where $\| * \|_2$ represents ℓ_2-norm.

Definition 7 (*Gaussian Mechanism*)**.** For any $\delta \in (0, 1)$ and $\sigma > \frac{\sqrt{2ln(1.25/\delta)}\Delta f}{\epsilon}$, there is $Y \sim N(0, \sigma^2)$ makes mechanism $\mathcal{M}$ satisfies (ϵ, δ)-DP:

$$\mathcal{M}(D) = f(D) + Y, \tag{8}$$

where σ is the standard deviation of the Gaussian mechanism, which determines the scale of noise, and the privacy budget ϵ is negatively correlated with the level of noise. Relaxation, represented by δ, refers to the probability of violating strict DP. However, in the high privacy mechanism ($\epsilon \to 0$), the variance formula of the original Gaussian mechanism is not tight. Balle et al. [82] proposed an improvement scheme for this phenomenon.

2.3. Local differential privacy (LDP)

Local differential privacy (LDP) is a commonly used mechanism in the industry to protect individual privacy during data collection [83]. Traditional DP adds noise to the original data after it has been aggregated in a trusted data center. However, in situations where the data center cannot be trusted, LDP is proposed as an alternative, where noise is added directly to the user's dataset before being transmitted to the data center for processing (shown in Fig. 4).

Security concerns arise not only from untrustworthy data centers but also from dishonest participants who may provide fake data that can be falsely labeled as real data with DP noise. For example, Cao et al. [84] have demonstrated the feasibility of data poisoning attacks against the LDP protocol. Besides, when the dataset contains highly sensitive information or has a large input domain, adversaries can more easily achieve the goal of obscuring the true distribution of the dataset by manipulating a small subset of data providers in the LDP protocol [85]. Another topic of interest is the allocation of privacy budgets in LDP [86]. The original definition of LDP assumes that all data providers' data are equally sensitive, requiring the same amount of noise. However, in practice, some data may be more sensitive than others, and privacy budgets can be conserved by applying more noise to those parts of the dataset that require greater protection.

2.4. Properties of differential privacy

2.4.1. Composition property

Deep learning is a computationally intensive task that involves many training steps, each of which consumes a certain amount of privacy budget. To obtain tighter upper bounds on the privacy budget required for deep learning, it is necessary to study the composition properties of DP applicable to neural networks (more details in Section 3.1.2). Simple composition properties, such as sequential and parallel composition property [87,88], assume that queries are independent of each other. However, in practice, queries are often correlated, which can be exploited to obtain tighter upper bounds on the privacy budget. Researchers have proposed various methods for computing privacy budgets, resulting in increasingly clear theoretical bounds on DP. To validate the tightness of these bounds, researchers have also developed new attack schemes to identify instances of privacy violations [89,90]. These efforts have contributed to a better understanding of the theoretical limits of DP in deep learning.

Sequential composition property [87]. Given a dataset D, use algorithm $\{\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_k\}$ that satisfies the DP mechanism to add noise to dataset D separately. The privacy budget is $\{\epsilon_1, \epsilon_2, \dots, \epsilon_k\}$. The output sequence $\{O_1, O_2, \dots, O_k\}$ satisfies $\sum_{i=1}^{k} \epsilon_i$-DP.

Parallel composition property [88]. For disjoint datasets $\{D_1, D_2, \dots, D_k\}$, use algorithm $\{\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_k\}$ that satisfies the DP mechanism to add noise to dataset $\{D_1, D_2, \dots, D_k\}$ separately. The privacy budget is $\{\epsilon_1, \epsilon_2, \dots, \epsilon_k\}$. The output sequence $\{O_1, O_2, \dots, O_k\}$ at this time satisfies $\max_{1 \le i \le k} \epsilon_i$-DP. According to what we discussed before, a higher ϵ means a weaker privacy guarantee.

Advanced composition property. Privacy loss is a random variable that has a range of its own value and a range of its expectation. Considering that the above two things may be different, Dwork et al. [91] pointed out a new way to get a tighter upper bound. The basic idea of the advanced composition is to exchange a small bit of δ for a lot of ϵ. The definition of the advanced composition is: For every $\epsilon \ge 0$, $\delta, \delta' \ge 0$, if $\mathcal{M}_i$ is a random function of (ϵ, δ)-DP, $\{\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_k\}$ satisfies $(\epsilon', k\delta + \delta')$-DP, where

$$\epsilon' = \sqrt{2kln(1/\delta')}\epsilon + k\epsilon(e^{\epsilon} - 1). \tag{9}$$

Compared to the simple composition properties, the advanced composition property can obtain significantly better performance when k is large, as the total privacy budget grows only linearly with $\sqrt{k}$ instead of k.

2.4.2. Post-processing immunity

Post-processing immunity provides the theoretical underpinnings for combining DP and neural networks. Specifically, post-processing immunity asserts that a result satisfying DP remains so regardless of any subsequent processing, such as arbitrary calculations or combination with other modules [82]. With this property in mind, researchers have explored adding DP noise to different parts of the neural network, such as the dataset, stochastic gradient descent, or adding a DP noise layer. By adding a small front differentially private network, it is possible to deploy differential privacy on large networks. However, while post-processing immunity provides convenience, some researchers have noted that it can increase variance and/or introduce bias [92].

Table 1
Main challenges and selected works.

Model	Challenge	Work
Traditional neural network	Noise addition methodfor better accuracy	[12,93–96]
	DP noise mechanism forthe tighter upper bound	[97–101]
FL	Balance of convergenceperformance and privacy	[102–105]
GAN	Appropriate noise preventsmemorizing samples	[106–109]
Specific network or framework	Noise addition applicableto the framework	[14,110–112]

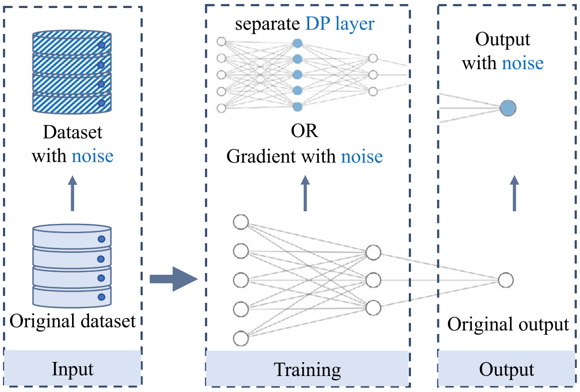


Fig. 5. Locations to add DP noise. Note that the scheme of adding noise on the output layer is generally applied to a particular model architecture.

3. Differential privacy in deep learning: About privacy

In terms of privacy, the trade-off between privacy guarantees and network performance has been a topic of focus for researchers, resulting in the continuous evolution of algorithms and mechanisms in this area. This section delves into the mechanisms that serve to ensure privacy in neural networks. Specifically, we examine the location where noise is added, techniques for enhancing privacy, and special schemes tailored to different networks, such as Generative Adversarial Networks (GANs) and Federated Learning (FL). Of particular significance is the lower bound, which is just as crucial as the upper bound but has received less attention in previous research.

3.1. DP mechanisms in deep learning

We summarize the main challenges of deploying DP mechanisms on diverse neural network architectures (outlined in Table 1). We subsequently provide an in-depth discussion in this section.

3.1.1. DP mechanisms—location of the noise

How to add noise in DP mechanisms does not have a universally optimal solution, as the ideal approach varies depending on the specific scenario. Previous studies have employed various techniques to add DP noise, such as incorporating it into the input layer, output layer, a separate DP layer, or stochastic gradient descent (SGD) (shown in Fig. 5).

Adding DP noise on the input layer. Adding noise at the input layer, whereby noise is directly added to the input data, is not a frequently utilized approach, as the resulting noise-induced data may cause the neural network to learn inappropriate decision boundaries. However, the effective deployment of DP mechanisms at the input layer requires prior knowledge to enable adaptive noise injection and prevent excessive influence on the decision boundary. To facilitate adaptive noise injection, Phan et al. [95] added DP noise to the input layers of the network through the use of Layer-wise Relevance Propagation (LRP) on another neural network that was trained with the dataset to learn the association between the features and the model output. The noise is added adaptively based on the significance of the features, which results in an entirely uncorrelated privacy budget with the number of training epochs, allowing for the efficient design and control of the privacy upper bound in advance. However, this technique incurs additional deployment and computational overhead due to the pre-training process.

Adding DP noise on the output layer. Directly injecting noise into the output of a neural network in a typical classification network lacks significance in the scope of neural network research, and is similar to adding noise to the query result of a database. Consequently, this approach requires a specific framework, and fewer studies have been conducted in this area. The Private Aggregation of Teacher Ensembles (PATE), proposed by Papernot et al. [14], serves as a typical example of such a framework. In PATE, noise is added to the voting results of the output layer to prevent an attacker from learning private information when two classified votes are close to each other. In their approach, sensitive data is partitioned into N disjoint datasets to train N teacher models, and the neural network predictions are derived through voting. If the majority of the teachers agree on one classification, it indicates that the classification result does not rely on a particular record. However, when two or more classifications receive very similar votes, there is a risk of private data leakage. To mitigate the privacy leakage, they added DP noise to the voting results. But adding DP noise alone is insufficient, as more queries to the teacher model can lead to more privacy leakage. To address this issue, they deploy a student model on the device that has access to public data. The input to the student model is a public dataset and a private dataset with privacy-preserving labels. With this configuration, even if adversaries gain access to the data sources of the student model, they can only obtain the privacy-preserved labels. PATE has a widely applicable architecture and robust privacy safeguards, but its high deployment cost limits its application scenarios. Despite the difficulty of deploying PATE, its strong generality makes it highly scalable on special neural networks. A series of works have proposed new mechanisms based on the design of PATE. For example, Jordon et al. [113] applied the PATE framework to GANs and obtained excellent performance.

Adding a separate DP noise layer. A separate DP layer is commonly used to enhance the robustness of neural networks against adversarial attacks (which will be discussed in detail in Section 4.3.1). This approach enables the entire network to satisfy DP through the post-processing immunity of DP. For example, Lecuyer et al. [15] proposed to insert a separate DP layer on the already designed neural network. The separate DP layer is added after the first layer of the DNN with noise having a mean of 0 injected. The size of the noise is proportional to the parameter p in the p-norm attack. Furthermore, the post-processing immunity allows them to extend their algorithm to large networks by splicing the large neural network behind a small encoder neural network that satisfies DP. When the preceding encoder layer adds bounded noise that satisfies DP, the entire large network model also has DP guarantees. Although this design results in a significant accuracy loss, it reduces the overhead of modifying the neural network deployment. Compared to traditional approaches for defending against adversarial attacks (e.g., training adversarial samples), this design provides a scheme for the defense that is not limited to a specific attack and guarantees robustness under any p-norm attack.

Adding DP noise on SGD. Currently, the predominant research focus is on integrating DP with stochastic gradient descent (SGD) in neural networks. Since a single data point can have a significant impact on the network, any changes in the SGD algorithm can be highly consequential. By adding DP noise to the SGD algorithm, the network's generalization performance can be improved, making it less sensitive to individual data points and better able to resist inference attacks. This solution is especially popular for lightweight networks as it incurs minimal additional deployment overhead. However, it is important to note that simply adding DP noise into SGD comes with some drawbacks, such as increased computational overhead and privacy budget proportional to the number of training epochs. The baseline method in this field is DPSGD proposed by Abadi et al. [12] that allows for privacy guarantees without significant noise addition. Variants of this approach have been explored in subsequent studies. DPSGD can be formulated as the following steps: (1) Sampling the input dataset; (2) For each sampling group, calculate the gradient; (3) Clip gradients above a certain threshold, which is a function of the norm bound and gradient; (4) Add noise to the gradient (the variance of Gaussian noise is proportional to the clipping bound); (5) Use the privacy calculation method (moments accountant) to calculate whether the privacy loss exceeds the privacy budget. In addition to DPSGD, there are other methods that possess similar properties. For example, Damaskinos et al. [114] have proposed integrating DP with Stochastic Coordinate Descent (SCD) as another solution that requires only minor tuning and yields competitive performance.

Summarization. The optimal choice of where to add DP noise in a neural network depends on the specific application scenario, and there is no inherently superior or inferior method. The input layer is a suitable location for adding noise if there is sufficient equipment available to pre-train the input data. When the data owner desires to maintain as much distance as possible between adversaries and the real data, a new generic framework like PATE that adds noise to the voting results of the teacher network can be designed. A separate DP noise layer is often appropriate when the data owner needs to protect against adversarial attacks, and modifying the network structure is challenging. Finally, the stochastic gradient descent (SGD) approach is the easiest to implement and most widely applicable. It allows for accurate control of the privacy budget and is thus preferred in many cases.

3.1.2. DP mechanisms—balance of privacy and accuracy

Deep learning requires many training epochs to produce reliable results, which is in contrast to the limited privacy budget of DP. If each epoch adds enough noise to mask training data features, combining privacy budgets sequentially (as mentioned in Section 2.4.1) would lead to a total privacy budget explosion as the number of epochs increases, resulting in lower model utility. On the other hand, if the total privacy budget is restricted to ensure model utility, the network may not provide an adequate privacy guarantee. Existing research mainly addresses this problem in the following ways.

Noise addition mechanism independent of the training epoch. In this scenario, DP noise is commonly added to the input or output layer of the neural network, which often results in high additional deployment costs, making this solution unsuitable for many cases. To address this issue, Phan et al. [95] obtained the proportion of noise added to the training data from the pre-training results. In [14], noise is added to the voting results outside the training process. These approaches bypass the issue of exploding privacy budgets associated with the training epochs, but they require platforms that can handle the substantial deployment costs and computational overhead.

Improve the calculation of budget. In the DPSGD method, DP noise is added to each epoch. However, this approach faces a contradiction: when the privacy budget allocated to each epoch is too small, weak privacy guarantees are achieved; conversely, when the privacy budget per epoch is appropriate, the traditional sequential composition method leads to a large total noise, rendering the network useless. The above contradiction can be solved by changing the composition method [12]: researchers have discovered that the sequential composition property has a very robust definition of privacy. Since it assumes that each query is completely uncorrelated, it gives a very loose upper bound, which will be significantly smaller in practice. To address this issue, researchers have proposed the advanced composition property, which exchanges a tiny bit of δ for a lot of ϵ and results in privacy budgets growing linearly with $\sqrt{k}$ instead of k when the number of serial k is large. However, this method cannot consider special noise distributions. To overcome this limitation, Abadi et al. [12] proposed moments accountant to compute tighter upper bounds.

Moments accountant [12] is a method applied to the DPSGD method. Compared to advanced composition, moments accountant focuses on higher moments of variables. More moments allow them to obtain tighter upper bounds. The definition of moments accountant is: given the probability of SGD random selection $q = L/N$ (N is the total number of samples and L is the number of randomly-selected samples), and the number of training epoch is T, there exist constants c_1, c_2, for any $\epsilon < c_1 q^2 T$, the composition consists of $\{\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_T\}$ satisfies (ϵ, δ)-DP for any $\delta > 0$ if:

$$\sigma \geq c_2 \frac{q\sqrt{Tlog(1/\delta)}}{\epsilon}. \tag{10}$$

Compared to previous methods, the moments accountant requires less privacy budget to achieve the same level of privacy protection.

However, it is important to note that using (ϵ, δ)-DP relaxation is not without cost. A small portion of the complete records of the dataset may be compromised, creating a significant privacy risk [115]. Although tools such as the moments accountant can provide very accurate upper bounds, existing neural networks may not be able to maintain sufficient privacy guarantees while sacrificing only a small amount of accuracy [13]. Therefore, this result suggests that it may be worthwhile to consider more practical application scenarios by focusing on the lower bound of privacy and taking into account the actual attacker's potential real-world limitations [17]. We will discuss this in more detail in Section 3.3.

Fine-tuning the noise on different parameters. By adding noise to parameters with less impact on the neural network, this approach achieves a lower loss of accuracy [99]. Although similar to the idea in [95], this approach shifts the focus from the input dataset to the SGD. One of the advantages of this solution is that it can be easily incorporated into an existing DPSGD framework without incurring additional deployment costs.

3.1.3. DP mechanisms—evaluation

The biggest challenge faced by existing research is striking the right balance between accuracy and privacy guarantees. Because the MNIST and CIFAR-10 datasets are relatively lightweight, many researchers tend to use them to validate the effectiveness of their mechanisms. Abadi et al. [12] achieved 90%, 95%, and 97% accuracy on the MNIST dataset for the test samples when reaching $(0.5, 10^{-5})$, $(2, 10^{-5})$, and $(8, 10^{-5})$-DP, and 67%, 70%, and 73% accuracy on the CIFAR-10 dataset when reaching $(2, 10^{-5})$, $(4, 10^{-5})$, and $(8, 10^{-5})$-DP. These results served as benchmarks for most of the subsequent work. We present the evaluation results in Table 2.

Accuracy. The core idea behind the DP mechanism is to add noise to different components of the neural network. Regardless of how the noise is introduced, such as through stochastic gradient descent, hidden layer parameters, or an independent DP layer, the added noise will inevitably result in reduced accuracy during convergence compared to what would have been achieved without the noise. As a result, it is essential for researchers to carefully consider the impact of noise on neural network accuracy. Some work attempt to protect model utility by weakening privacy guarantees (e.g., using DP variants), but Papernot et al. [96] argue that such approaches are unnecessary. Instead, they propose using tempered sigmoids instead of ReLU, which is prone to gradient explosion, as a way to effectively enhance network utility.

Privacy budget. The privacy budget value represents the level of privacy protection, that is, the ability to resist inference attacks. Privacy protection and neural network accuracy are often seen as being in conflict, as satisfying one can lead to a decrease in the other. These two things usually cannot both be satisfied (Mcmahan et al. [94] proposed that in federated scenarios with enough users, the same level of privacy protection can be achieved by increasing the cost of calculation rather than reducing the network accuracy). Researchers are therefore seeking more efficient ways to add noise, such as adaptively adding noises of varying sizes during the neural network learning process [116]. Some researchers are also working on improving existing DP noise mechanisms to strike a balance between neural network privacy and utility. For example, the heterogeneous Gaussian mechanism has been used to enhance network robustness [25], while a new privacy definition known as f-DP [117] has been considered instead of the traditional definition to take advantage of composition and subsampling [101]: the central limit theorem of f-DP more accurately captures the privacy loss in model iterations to achieve stronger privacy guarantees.

3.1.4. DP mechanisms—summarization and future work

Table 2 presents a summary of related work on DP mechanisms in deep learning, with a focus on the following aspects. (1) The location where DP noise is introduced. In most studies, DP is combined with stochastic gradient descent. Researchers adapt the combination mechanism to various application scenarios and employ new DP definitions to improve accuracy while maintaining a certain level of privacy guarantee. (2) The evaluation of trade-offs between accuracy and privacy. The accuracy of neural networks and privacy protection are two major concerns in DP mechanisms. In general, privacy leakage is inversely proportional to the accuracy loss. For evaluating the accuracy, the DPSGD scheme's performance is used as the baseline due to its landmark experimental results, which serve as a reference for most subsequent studies. (3) The type of DP noise being added. Laplace and Gaussian noise are commonly used by researchers. Some researchers propose new noise mechanisms to achieve better performance [25,101].

We have summarized the discussions as follows.

- Most studies on DP in neural networks use gradient-based noise, with DPSGD being the most widely used method. Yu et al. [118] explored the impact of privacy noise on optimization property and showed the combination of gradient perturbation and composition methods is generally superior to other perturbation methods.
- Improved DPSGD schemes can be classified into two categories: those focused on neural networks [95] and those focused on DP properties [12]. The former deals with where privacy noise is added to the network and how the noise can be unevenly distributed to reduce the privacy budget. The latter deals with improved noise distribution functions and composition methods that are suitable for neural networks. These two categories are often combined to develop new differentially private deep neural network algorithms [15].
- As neural networks are not fully explainable, an exciting research direction is to challenge the upper bound of the theoretical privacy budget by designing attack schemes [17], aiming to make the lower bound approach the upper bound. To date, all relevant studies [13,119] have shown that the upper bound of DPSGD is tight.

In light of the above discussions, we propose the following potential research directions.

Improvements of DPSGD. As DPSGD is a prevalent technique, maintaining an equilibrium between privacy and accuracy based on DPSGD is an ongoing research topic. Its quantitative definition and adaptive algorithms require more in-depth study. Potential future directions in these areas include: (1) Techniques to improve network accuracy within the same privacy budget. (2) Methods to achieve tighter privacy upper bounds.

Design new DP mechanisms. Although the upper bound has now proven to be very tight, a better privacy budget calculation mechanism is still desirable to achieve even tighter upper bounds. Moreover, in some real-world scenarios, we may be willing to trade more deployment cost and computational overhead for improved accuracy, necessitating the development of entirely new frameworks. Potential future directions in these areas include: (1) DP noise addition schemes that have minimal impact on network accuracy, including approaches unrelated to training epoch and methods with adaptive noise (such as those in [95]). (2) Novel composition methods that are more suitable for deep learning neural networks (to achieve tighter upper bounds). (3) Widely applicable solutions rather than being limited to a particular type of network.

3.2. DP mechanisms in gan, fl, and beyond

Besides works focusing on the algorithms related to SGD, some researchers have turned their attention to exploring DP mechanisms in different neural network scenarios. The most popular are GAN and FL [107,120].

Generative Adversarial Networks (GANs) [121] allow the generator to create samples different from the training set that are difficult for the discriminator to distinguish whether they are real or not. From the perspective of privacy protection, GANs are used to generate high-quality fake samples to protect the private information of the real samples. However, this is not entirely secure: the information of the real samples can be reconstructed by the fake samples [122], so we need to obscure the information of the real samples, and the DP framework is suitable for this purpose. The introduction of DP noise may cause the generator and discriminator in GANs to have non-convergence or converge to a noisy equilibrium [73], which needs to be addressed by researchers.

Table 2
Summarizing DP mechanisms in deep learning. We focus on the location of the DP noise combined with the neural network, the accuracy, and the type of noise distribution. For accuracy evaluation, we use the DPSGD method as a baseline, which achieves 90%, 95%, and 97% accuracy on the MNIST dataset for the test samples when reaching $(0.5, 10^{-5})$, $(2, 10^{-5})$, and $(8, 10^{-5})$-DP. We define the accuracy of the DPSGD method as MEDIUM, and the accuracy of methods that are similar to it (more accurate in some conditions and less accurate in others) are also classified as MEDIUM; methods that are less accurate than DPSGD are classified as LOW; and methods that are more accurate than DPSGD is classified as HIGH. Specifically, there are some works that are not performed on the MINST dataset (we have marked the datasets they used in the table). For these works, we follow the above classification rule if it is compared with DPSGD. While it is not compared with DPSGD, we compare the accuracy loss of the privacy scheme to the non-privacy scheme around $\epsilon = 2$, and classify it as HIGH if the accuracy loss $\leq$ 5%; if 5% $<$ accuracy loss $\leq$ 10%, it is classified as MEDIUM; the other case is classified as LOW.

Ref.	Position	Accuracy	Noise	Advantages and disadvantages
[93]	SGD	LOW	Laplace	Early work on adding DP noise to neural networks. Lack of consideration for the balance of accuracy and privacy loss.
[12]	SGD	MEDIUM	Gaussian	Propose moments accountant for tighter privacy upper bound. (ϵ, δ)-DP leaks a small portion of dataset and may cause a disproportionate privacy risk [115].
[95]	Input	HIGH	Laplace	Adaptively add noise on model inputs, making the privacy budget independent of epochs. Require significant additional deployment and computational overhead.
[94]	SGD	HIGH (Reddit)	Gaussian	Implement differential privacy on user-level data with negligible accuracy loss. Suitable for federated datasets with many users, and the computational overhead is high.
[14]	Output	HIGH	Laplace	Design a framework for isolating public and private datasets and apply DP in it. From the perspective of DP, this noise addition scheme is only suitable for specific frameworks.
[97]	SGD	LOW	Binomial	Use Binomial mechanism to achieve higher communication efficiency. Applicable to federated learning scenarios.
[98]	SGD	-	Gaussian	General modular DP scheme that allows for minor variation in the training algorithm. It is excellent in terms of scalability for practical deployment.
[99]	SGD	HIGH	Gaussian	Add more DP noise to parameters that have less impact on the output to reduce accuracy loss. Potentially increase the unfairness of the model.
[25]	Hidden layer	HIGH	Heterogeneous Gaussian	Use heterogeneous Gaussian mechanism to enhance the robustness against adversarial attacks. Excessive noise may be added against specific types of adversarial attacks.
[100]	SGD	HIGH	Gaussian	They perform estimation of privacy loss under two different data batching methods. They have high privacy budget, which may result in meaningless privacy guarantees [13].
[116]	SGD	HIGH	Gaussian	Reduce privacy overhead by increasing the convergence rate through an adaptive method. Need a little more computational overhead.
[101]	SGD	HIGH	Gaussian	Consider f-DP, which has clearer privacy constraints on privacy analysis. A large ϵ is still required to achieve an acceptable accuracy loss.
[114]	SCD	HIGH (YearPredictionMSD)	Gaussian	Replace SGD with SCD, which requires less tuning and has similar performance with SGD. Consider the practicality of the model, but does not address the balance of privacy and accuracy.
[96]	SGD	HIGH	Gaussian	Replace ReLU with tempered sigmoids to avoid gradient explosion and thus ensure accuracy. It is simple and effective in practice.

Federated Learning (FL) [123] is of interest to researchers because the structure of FL is compatible with the idea of local DP. In FL, the network parameters are high-dimensional, continuous, and high-precision values provided by each participant, so making the existing LDP algorithm applicable to FL is challenging [104].

Meanwhile, some special neural networks require the introduction of DP for privacy preservation (e.g., teacher–student networks [14]). These special neural networks require unique DP deployment schemes, and the privacy budget allocation scheme for DP in these special cases needs a particular design. Therefore, relevant work needs to be analyzed and summarized.

3.2.1. GAN with DP

Generative adversarial network (GAN) [121] has become one of the most popular frameworks for unsupervised learning on complex distributions in recent years. The main structure of GAN includes a generator $\mathcal{D}$ and a discriminator $\mathcal{G}$ (shown in Fig. 6). The generator produces fake samples with noise satisfying the latent distribution, and then these fake samples are fed into the discriminator for discrimination. The discriminator receives both real and fake samples, and its role is to distinguish the fake samples generated by the generator as fake and to recognize the real samples as true. The goal of the generator is to produce fake samples that can fool the discriminator into thinking they are real. The generator and the discriminator are trained against each other to produce sufficiently realistic fake data.

In a nutshell, GAN learns the features of real data and then generates high-quality fake data to prevent unauthorized access to real data, which could lead to privacy leakage. However, the generated distribution of data generated by GANs is often concentrated around the training data, making it possible for adversaries

Table 3
Summarizing GAN with DP. We are concerned with whether DP noise is added to the discriminator or the generator. $\mathcal{G}$ is the generator and $\mathcal{D}$ is the discriminator.

Work	Position		Statement
	$\mathcal{G}$	$\mathcal{D}$	
Xie *et al.* [106]	✗	SGD	Differentially private discriminator + Computation of generator → Differentially private generator.
Zhang *et al.* [107]	✗	SGD	Only Discriminator can access the actual data. Structure of Discriminator is simpler.
Xu *et al.* [108]	✗	SGD	New gradient clipping strategy.
Jordon *et al.* [113]	PATE	✗	ApplyL PATE [14] to GAN.
Torkzadehmahani *et al.* [109]	✗	SGD	New clipping and disturbance strategies. Use Rényi DP accountant to track privacy budget.

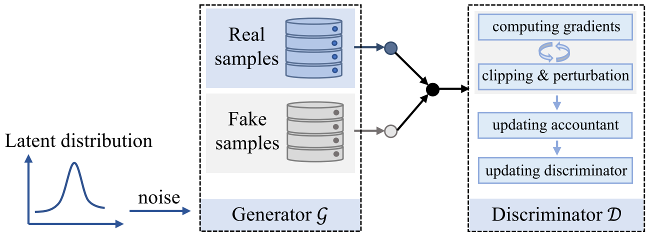


Fig. 6. A simple example of a Generative Adversarial Network (GAN) combined with DP [107]. The DP mechanism is added to the discriminator because only the discriminator has access to real-world data. Meanwhile, the discriminator has fewer parameters (allowing for a tighter estimate of privacy loss).

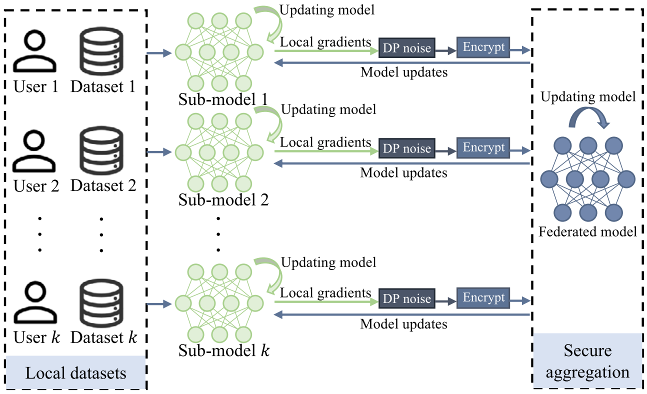


Fig. 7. A simple example of FL combined with DP [120]. The data owners add DP noise to their gradients, encrypt them, and upload them to the federated model.

to reconstruct real samples from fake ones [122]. To prevent such attacks, noise needs to be introduced to the generative process. Differential privacy provides a suitable framework for adding noise, prompting researchers to explore combining DP mechanisms with GANs. Most of these works are based on the idea of DPSGD, which involves improving the DPSGD mechanism and combining it with the Discriminator of GANs [106–109]. These works focus on improving gradient clipping strategies and privacy budget tracking methods to obtain better experimental results. The specific solutions are not vastly different from those mentioned in Section 3.1. It is worth noting that nearly all researchers choose to add noise to the Discriminator rather than the Generator. This is because the Discriminator has access to real data, and its structure is simpler, making it easier to add noise to it. In particular, Jordon et al. [113] proposed a scheme that deviates from basic DPSGD. They combined Private Aggregation of Teacher Ensembles (PATE) [14] with GANs, modifying PATE's framework to apply to generators, and trained a differentially private generative model. An overview of GANs with DP is provided in Table 3.

3.2.2. FL with DP

FL [123] is widely used in distributed learning scenarios, as it enables multiple data providers to collaborate without revealing their specific information to each other or the cloud server. Given k data owners $\{F_1, F_2, \dots, F_k\}$ who have datasets $\{D_1, D_2, \dots, D_k\}$. The traditional method is to get the union of all the datasets to obtain the total dataset and then use the total dataset to train deep learning model M_{SUM}. In FL, all data owners train the model M_{FED} together. In this structure, no data owner F_i will disclose his dataset F_i to others (shown in Fig. 7).

However, the framework of FL, while facilitating data silos to jointly unleash the power of their respective data, is not without security threats. The adversaries can be untrustworthy participants (including users themselves and curious cloud servers) or external adversaries (shown in Table 4). As a dishonest data provider, an adversary can infer private information about user datasets by analyzing distributed models [124]. Moreover, an unscrupulous data collector can infer private data information from their model updates. In addition to this, there is the possibility of cloud servers colluding with multiple dishonest users [125]. To address these risks, DP, which is often combined with the parameters uploaded by users, has become one of the schemes to guarantee the privacy of FL [126]. This technique has received significant attention, and [127] describes a unified vision of the FL framework that supports DP.

Due to the natural compatibility between the two concepts, LDP is a popular choice in FL. However, centralized differential privacy (CDP) also has its advantages. Naseri et al. [128] examined the differences between LDP and CDP in FL and found that LDP is effective against membership inference attacks, while CDP performs better under backdoor attacks (a special type of poisoning attack, which will be discussed in Section 4.3.2). Unfortunately, both mechanisms are not effective against attribute inference attacks because naive DP definitions are record-level [128].

For example, Geyer et al. [124] proposed a solution for situations where some participants have low-quality data. Besides, Truex et al. [104] found that the existing LDP protocol is not suitable for the situation where the parameter updates of FL are collected repeatedly from each participant and consist of high-precision high-dimensional continuous values. Accordingly, they proposed the LDP-fed system to solve the problem of uncontrolled noise in parameter updating for large and complex FL models. There are also some other studies focusing on particular types of data fields, such as the combination of DP and FL in medical datasets [129].

Similar to the study of mechanisms on DP in neural networks, research on FL also focuses on the trade-offs between the

Table 4
Application domains of FL with DP. We classify them by whether or not external adversaries are considered. When external adversaries are present, the noised data often has to be encrypted, which increases the overhead.

Work	Application Domains		Type of DP
	Untrustworthy Participant	External Adversary	
Geyer *et al.* [124]	✓	✗	CDP
Jiang *et al.* [102]	✓	✗	LDP
Wei *et al.* [103]	✓	✓	LDP
Hu *et al.* [131]	✓	✗	LDP
Sun *et al.* [132]	✓	✗	LDP
Hao *et al.* [125]	✓	✓	LDP
Truex *et al.* [104]	✓	✗	LDP
Naseri *et al.* [128]	✓	✗	LDP&CDP
Truex *et al.* [120]	✓	✓	LDP
Wang *et al.* [105]	✓	✗	LDP
Sun *et al.* [130]	✓	✗	NFDP

level of privacy assurance and model performance. Most of the studies that focus on model performance are devoted to allocating and improving the local DP models appropriately. Besides, The vast majority of previous work chose to use the traditional normal DP definition, while Sun et al. [130] considered using a noise-free differential privacy (NFDP) mechanism (NFDP is DP of sampling without replacement, which eliminates the dependence on the number of queries on the public dataset) and resolved the problem of privacy cost explosion.

3.2.3. Other types of neural networks with DP

Differential privacy combined with particular types of neural networks [110,111] is also primarily based on the design intuitions of DPSGD. However, the direct application of DP without any optimization can significantly reduce the network's utility, and researchers often fine-tune the mechanisms according to the network's nature. Although the post-processing nature of DP allows it to be added to various neural networks in a simple way, whether the direct introduction of the DPSGD mechanism without any optimization is an appropriate solution is a question worthy of consideration by researchers. It is important to optimize the mechanisms to ensure reasonable utility. Phan et al. [111] studied how to reasonably allocate the privacy budget in the hidden unit group of CDBN, which is a well-known energy-based neural network with DP. They focused on how to allocate the privacy budget reasonably in the hidden unit group. Li et al. [112] considered the parameter transfer in the meta-learning method, and they used DP noise to protect the transferred parameters. Phan et al. [133] studied DP in adversarial learning of deep neural networks (DNN) and tried to make their method applicable to large DNNs and datasets.

In addition to Federated Learning, other network constructs for distributed learning, such as the Alternating Direction Method of Multipliers (ADMM), have also been explored. Huang et al. [110] investigated the differentially private protection of distributed deep learning based on ADMM and achieved higher utility by using the approximately enhanced Lagrangian function and time-varying Gaussian noise in the iterative process. In addition, from a practical point of view, Farokhi et al. [134] had more realistic considerations: in a distributed scenario, it is difficult for the central learner to communicate with all data providers simultaneously. Therefore, they addressed the challenge of communication in distributed scenarios by focusing on the DP protection algorithm when the central learner communicates asynchronously with data owners.

3.2.4. DP in special application scenarios

Building on the DPSGD approach, researchers have explored variants of the DPSGD approach on different types of neural networks. In addition to these designs mentioned above on particular networks, researchers have found some inspiration for new applications of DP in other fields. For example, since DP can be used for outlier detection [27], and poisoned data can be considered as outliers of the input data, it is not surprising that DP can also be used to detect poisoned data [27] (more discussion in Section 4.3.2).

Differential privacy is widely used as a tool for designing frameworks with privacy guarantees. For example, consider a real-world scenario where we cannot trust data aggregators because they may expose private data (sensitive datasets such as healthcare and finance). In this setting, Hynes et al. [135] proposed Myelin, a framework that protects the privacy of model providers and data providers. In the medical field, Beaulieu et al. [136] focused on distributed clinical data collection under the DP mechanism. Meanwhile, in the visual and image domain, Fan et al. [137] explored the protection of image data by DP to reduce the success rate of re-identification attacks effectively. And Tramèr et al. [138] also studied DP protection for visual tasks. They suggest that in order for visual DP machine learning tasks to perform better, we need more private data or better features. In terms of the Internet of Things (IoT), users' data is often collected in the cloud for learning. DP in IoT has also received attention from researchers. Arachchige et al. [139] proposed a practical local DP scheme in a cloud-based environment driven by the Internet of Things.

These particular application scenarios are partly privacy-related and partly beyond privacy. For the privacy-related ones, researchers usually focus on constructing new frameworks applicable to their scenarios of interest rather than on the improvement of the DPSGD algorithm. In fact, the DPSGD algorithm can be used for training most of the datasets that require privacy preservation, such as medical and local IoT device data. As for research beyond privacy, some research has identified the impact of DP (both positive and negative) on neural networks outside of privacy [16,23]. This will be discussed in Section 4.

3.3. Regarding the lower bound of DP mechanisms

In the previous section, we discussed how the mathematical definition of DP provides an upper bound on privacy guarantees that cannot be broken under arbitrary attacks. However, in practice, when only the upper bound is considered, almost all algorithms fail to simultaneously achieve meaningful privacy guarantees and acceptable utility losses [13]. Nevertheless, this does not imply that introducing DP into neural networks is useless. Note that the upper bound is a theoretical privacy guarantee, which means such privacy loss may not be incurred in real-world situations. In fact, real-world attackers face various constraints that prevent them from achieving the theoretical maximum privacy leakage [17]. Therefore, to evaluate the privacy leakage of a DP scheme under realistic attacks, we need to explore the lower bound, which is the privacy leakage that has been generated under actual attacks. Unfortunately, even though DP is a neat mathematical definition, we can only obtain the lower bound by running real attacks rather than calculating it in advance [17,119].

Differential privacy uses strong adversary assumptions when calculating the upper bound. These powerful adversaries have white-box access to models and can also modify the dataset. However, in most practical cases, the adversary only has black-box model access (they do not know the exact architecture and parameters of the model). Simultaneously, it is difficult for read-world adversaries to arbitrarily change the training dataset.

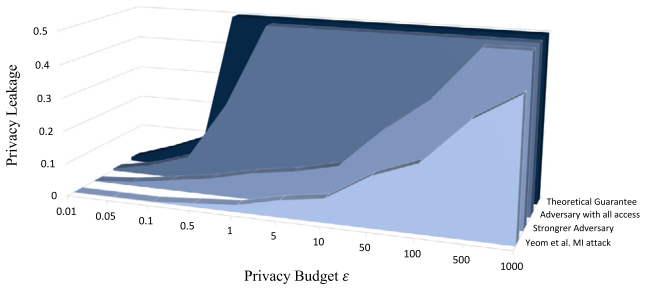


Fig. 8. Schematic diagram of privacy leakage under different attacks. Privacy leakage is defined by Yeom et al. [23] as the difference between the true positive rate (TPR) and the false positive rate (FPR) under attack (e.g., there are 100 records belonging to the training data and 100 records that do not belong to the training data, then the adversary identifies 90 of the former and 40 of the latter as belonging to the training data, the privacy leakage here is 0.5). (1) When the model is under the threat of membership inference attack (proposed by [23]), the privacy leakage increases as the privacy budget increases, but this has a significant gap with the theoretical upper bound [13]. (2) We inject virtual data points to simulate the situation where the model faces a stronger adversary and adversary with all access: the adversary with more access leads to less noise that can be injected (to guarantee the accuracy of the network) [17]. (3) Finally, we simulate the theoretical guarantee: the theoretical upper bound is tight as far as the current study is concerned. The adversary can approach but cannot exceed the theoretical upper bound [17].

This difference in assumptions of the adversary's capabilities may make the theoretical privacy guarantees too strong for a real-world adversary. As a result, researchers have found a non-negligible gap between the upper and lower bounds [13]. This means that under certain known assumptions about the adversary, the privacy guarantees that were thought to be weak due to accuracy compromises are actually stronger than expected. On this basis, exploring the lower bound under different assumptions about the adversary can give us more options to adjust the noise scheme to achieve better accuracy when training differentially private neural networks [17]. For ease of presentation, we will use the example shown in Fig. 8 to illustrate the differences in privacy leakage between theoretical guarantee and actual attack assumptions.

How to get the lower bound? For understanding the size of the gap between the upper and lower bound in a deep learning model with DP protection, the primary challenge is how to access the lower bound. One direct approach is to utilize an existing attack scheme to obtain privacy leakage. Jayaraman et al. [13] conducted comprehensive research on this subject, utilizing existing membership inference attack methods to target differentially private neural networks. Among these attacks, Yeom et al. [23] (which has full white-box access) performed the best but still fell short of the theoretical privacy leakage. They attributed this phenomenon to the weakness of existing attacks, suggesting that stronger attacks could approach the upper bound of DP. Under this assumption, they analyzed the balance between privacy guarantees and the accuracy of current mechanisms and found that no prior work had achieved meaningful privacy guarantees with acceptable utility loss. According to their experimental results, under the Rényi differential privacy (RDP) setting, the neural network on CIFAR-100 has 0.53 accuracy loss (which represents the degradation of the network accuracy compared to that without noise, for example, if a network has 80% accuracy without noise and 60% accuracy after noise addition, then its accuracy loss is 0.25) with $\epsilon = 10$ while privacy leakage is 0.07 (the larger the privacy leakage, the more capable the attacker is, as explained in detail in the description of Fig. 8). In the above experimental result, the privacy leakage of the network was acceptable, but the accuracy loss was deemed unacceptable. Meanwhile, the accuracy loss of other DP variants studied in their research is even more than that of RDP under the same ϵ. However, when the network suffers an acceptable loss of accuracy, the value of ϵ becomes too large, resulting in excessive privacy leakage. From the experiments of [13], it seems that no deep learning model can use DP to achieve utility and privacy at the same time, but the gap of lower and upper bound gives us such a possibility.

As the privacy loss obtained by existing attack methods is typically far from the upper bound, researchers have begun to consider instantiating attackers instead of relying on existing attack methods and assumptions. Jagielski et al. [119] computed the lower bound by instantiating a black-box adversary. Specifically, they selected a small poisoning set from the dataset and constructed a binary classifier to measure the actual privacy leakage. Their results showed that this instantiation method approximated the upper bound more closely than using existing attacks. For differentially private neural networks, instantiated attackers can effectively quantify privacy leakage. Nasr et al. [17] further formalized this phenomenon, explaining that the gap between the upper and lower bounds arises from differing assumptions about the attacker's capabilities. They demonstrated that the inability to approach the theoretical bounds with existing attacks was due to the limitations of real-world adversaries. They also use the adversary instantiation method, which actually constructs an attacker, to calculate the level of privacy loss. They found that when the attacker only has access to the API or black box, the actual privacy leakage is significantly smaller than the theoretical one. And when an adversary has white-box access, or even further, the ability to insert poisoned data at every step of the training process (which is difficult to achieve in a real attack), the privacy leakage gets even closer to the theoretical guarantee. Further, the ability to contribute poisoned gradients to the data center is also favorable for the adversary (in the FL setting). Finally, in the ideal situation where the adversary had all capabilities and could construct the most advantageous dataset for themselves (although such attackers are almost non-existent), the actual privacy leakage was very close to the upper bound of the theory. This work formatted the study of lower bounds by categorizing the different permissions possessed by the attacker from the actual situation to the ideal situation. Their explanation of the gap between theoretical and actual privacy leakage suggests that adding only a little noise could achieve sufficient privacy in practical applications (thereby avoiding excessive accuracy loss).

Summarization. Our findings based on the current research are as follows. (1) In real-world application scenarios, networks that add only a small amount of noise (to ensure network utility), which were previously considered insufficient to provide adequate privacy guarantees under strong adversary assumptions of DP, may actually be effective in defending against real-world adversaries who do not meet the strict criteria of DP assumptions. (2) The latest research endeavors to construct attackers beyond the practical limitations of the real world and thereby approximates the theoretical upper bound, which is proven to be tight.

Future work. We anticipate that research on lower bounds will continue to be a prevalent topic in the coming years because using only upper bounds often fails to provide sufficient privacy guarantees while maintaining network utility (even a slight decrease in accuracy is unacceptable in practice). The primary focus of the research on lower bounds is on how to obtain a more accurate lower bound. The current state-of-the-art approach involves constructing virtual attackers and using multiple attempts to launch attacks to obtain statistical results. Therefore, future research in this field will pursue the following directions. (1) Exploring more suitable computational schemes to narrow the gap between upper and lower bounds by instantiating the adversary. The current method of constructing an adversary requires a large number of attacks to obtain statistical results, which necessitates high computing power. Are there any alternative, lightweight approaches? (2) Investigating different assumptions about real-world adversaries and obtaining lower bounds under various circumstances. Nasr et al. [17] has proposed six classifications for adversary capabilities, and we expect to see more diverse classification assumptions in future studies. This will enable us to better deploy real-world networks.

4. Differential privacy in deep learning: Beyond privacy

Apart from the standard privacy guarantees against inference attacks, the implementation of DP involves adding noise, which can have particular effects on neural networks. These effects include certain benefits, such as mitigating over-fitting and enhancing robustness, as well as some side effects, such as reduced fairness. We will provide a detailed discussion of these effects in this section.

4.1. Fairness

Neural networks often exhibit lower accuracy when dealing with under-represented subgroups in real-world datasets. Various factors can contribute to model unfairness. For instance, a loan prediction system with a low proportion of women in the training dataset may discriminate against women's loan applications. Interestingly, the addition of DP noise has been shown to exacerbate model unfairness [16]. As a result, mitigating discrimination caused by DP has emerged as a recent, critical research topic.

4.1.1. Impact of DP on fairness

As mentioned earlier, introducing DP into the SGD of neural networks inevitably results in decreased accuracy. Bagdasaryan et al. [16] conducted experiments suggesting that this accuracy decrease may be discriminatory. They observed that networks using DP exhibited the "weaker of the weaker" phenomenon, meaning that under-represented subgroups experienced a more significant decrease in accuracy. This work differed from previous research that only provided overall results on this issue, such as the overall accuracy drop of the MNIST dataset. Instead, this study analyzed the accuracy of various data subgroups. For example, in a sentiment analysis dataset of Twitter articles, the accuracy of individuals who wrote in African–American English (AAE) was lower than that of those who wrote in Standard American English (SAE) on the non-DP model. Furthermore, on the DP model, the accuracy of AAE decreased more than that of SAE. After presenting this general phenomenon, the authors explained the reason behind it. The DPSGD method clips gradients higher than a certain threshold, which disproportionately affects subgroups with larger gradients during training (under-represented subgroups are more likely to produce larger gradients). Given this phenomenon, the next critical task for researchers is to develop methods to mitigate or even eliminate this unfairness.

Based on Bagdasaryan's work [16], Farrand et al. [20] provided some additional insights on the impact of DP on fairness in cases of slightly imbalanced data and low privacy guarantee. This suggests the urgent need for mitigation mechanisms to address fairness issues. Given that the clipping step is a primary source of unfairness in DP, Tran et al. [21] explored the effect of fine-tuning the clipping parameter on model fairness and accuracy. They discovered that the network achieves the best accuracy-fairness trade-off when the clipping parameter is set to a moderate value. If the clipping parameter is too small, the resulting model provides some fairness guarantees, but accuracy is further reduced due to the loss of significant information. On the other hand, if the clipping parameter is set too high, the resulting model achieves poor fairness performance because clipping is rarely performed.

4.1.2. Recent efforts addressing fairness issue with DP

Even though neural networks often face conflicts between privacy, fairness, and accuracy, researchers strive to find a balance between the three to improve the network's usability. Some representative solutions are shown as follows.

Relaxing definitions of fairness. As traditional notions of fairness and privacy are challenging to achieve simultaneously, Cummings et al. [140] proposed using approximate fairness on a finite sample access setting. After losing the notion of fairness, they successfully propose a classifier that satisfies privacy and approximate fairness.

Fine-tuning the distribution of noise. In Section 3.1.2, we mentioned that fine-tuning the noise on network parameters can reduce the loss of accuracy. Similarly, the discrimination of the neural network can be mitigated by fine-tuning the noise. Ding et al. [141] assigned different coefficients to standard and protected attributes to reduce discrimination while achieving a relaxed (ϵ, δ)-DP.

Separating privacy and fairness. Combining privacy and fairness in one step can create conflicts, so it is beneficial to consider separating the processing of privacy and fairness guarantees. Padala et al. [142] proposed a framework based on FL, where each agent trains the fairness model first and then the privacy model. The fairness and privacy schemes in this framework are replaceable, making it adaptable to different situations.

4.1.3. Potential future directions

The past few years have seen extensive research on the impact of the DPSGD approach on network fairness and its causes [16,20], with subsequent work investigating more detailed phenomena based on these findings. However, there is still much to explore in terms of algorithms that can mitigate, eliminate, or even reverse this unfairness [141,142]. We believe that future work on DP and fairness should focus on the following areas. (1) Developing non-clipping adjusting algorithms to mitigate discrimination in traditional DPSGD settings, such as fine-tuning the noise distribution [141]. (2) Designing adaptive clipping schemes that can find trade-offs between accuracy and fairness in DPSGD settings, building on the research of Tran et al. [21]. (3) Developing mechanisms to improve fairness in special networks or non-DPSGD schemes. (4) Creating general frameworks that can separate fairness and privacy guarantees, as the framework proposed by Padala et al. [142]. (5) Investigating the influence of DP schemes on fairness in various networks in more detail.

4.2. Over-fitting

Over-fitting is a well-known troubling phenomenon in machine learning, where the model performs well on the training data but poorly on new data, limiting its usefulness in real-world scenarios. To tackle this problem, researchers have found that introducing well-designed noise into the neural network training process can mitigate over-fitting, and DP noise is a suitable option [22]. This means that DP noise not only protects privacy but also has the added benefit of mitigating over-fitting.

On the other hand, can models with poor generalization be more susceptible to membership and attribute inference attacks? The answer is yes. Yeom et al. [23] analyzed various learning algorithms and model properties and demonstrated that models with more over-fitting are weaker under membership and attribute inference attacks. Therefore, using DP can be a two-in-one solution to the over-fitting and privacy problems. Wu et al. [24] proposed a DP-based SGD scheme for pathology image classification that simultaneously addresses the challenges of over-fitting and privacy leakage.

4.3. Robustness

In addition to its traditional role of defending against inference attacks, DP has been found to enhance network robustness, particularly in the context of adversarial attacks and poisoning attacks. These benefits are derived from the original definition of DP, which involves concealing differences in adjacent datasets. Adversarial attacks are designed to deceive the network by introducing subtle noise, whereas DP effectively prevents even minor changes in input data from significantly affecting the output, as demonstrated in [15].

4.3.1. Robustness against adversarial attacks

Adversarial attacks on images involve adding imperceptible noise to the input image, causing the network to classify it incorrectly. The crucial aspect of such attacks is that the added perturbation is small, but the impact on the network is significant. The DP framework can be an effective defense against these attacks by using DP noise to mask slight differences that have a disproportionate effect.

Inspired by this potential, researchers have conducted several studies on using DP to protect against adversarial attacks [15,25,26]. One notable work is PixelDP, proposed by Lecuyer et al. [15]. In their approach, they added noise with a mean value of 0 after the first layer of DNN, where the size of the noise is proportional to the p of the p-norm attack. This technique enabled them to obtain the confidence level of the network under p-norm attack. The novelty of their approach is that it departs from the best-effort defense approach and proposes a scheme that can defend against all p-norm attacks rather than being limited to a specific attack. This breakthrough removes the defense method from the previous offensive and defensive arms race and increases the usefulness of the network. Moreover, post-processing immunity plays a crucial role in their work, where they spliced small PixelDP networks in front of large networks, providing large networks with some immunity to adversarial attacks. However, this approach comes at the cost of losing a significant amount of accuracy, although it has the advantage of not requiring any changes to the structure of large networks.

4.3.2. Poisoning and anti-poisoning

DP can also defend against poisoning attacks, where the attacker injects specially crafted poisoned samples into the training dataset to change the model behavior and degrade the model's performance. These poisoned data can be considered as "outliers" [27]. Inspired by the stability guarantees of DP, Du et al. [27] demonstrated that DP could improve the effectiveness of poisoning attack detection.

Moreover, the use of DP in poisoning is not limited to detection only. Giraldo et al. [28] proposed an innovative idea of masking poisoning attacks using DP noise. The research is based on the scenario where data is collected by IoT devices and sent to the data center for processing. In this approach, adversaries inject DP noise into some of these IoT devices, with the distribution of noise carefully designed to make the post-poisoning statistical distribution and the original distribution challenging to distinguish by the data center. Although the data distribution collected by the data center is close to the real data distribution, it can affect the judgment results of the data center. Such an attack is more concealed than previous attacks because it can make more significant changes to the results without the data collectors noticing it. Researchers have transferred this idea to the field of deep learning, where Hossain et al. [29] developed a covert model poisoning scheme based on DP noise for attacking FL networks.

5. Conclusion

In recent years, the combination of DP and neural networks has gained much attention. While previous reviews have primarily focused on privacy concerns, they have lacked analysis of lower bounds and impacts beyond privacy. In this area, we have supplemented those reviews and provided a more detailed discussion of future research directions. While privacy-related work, such as DPSGD and its follow-up research, has received significant attention, it is important that one must focus on not only privacy but also the usability of neural networks. Recent research has highlighted that the cost of usability resulting from appropriate privacy guarantees may be too high, thus necessitating consideration of the DP lower bound. Furthermore, beyond privacy concerns, the impact of DP on fairness needs to be addressed by developing algorithms that balance privacy, accuracy, and fairness. Additionally, the potential benefits of DP on neural networks, such as reducing over-fitting and enhancing robustness, require theoretical quantification and analysis. Future research is necessary to fully comprehend the implications of DP on neural networks.

CRediT authorship contribution statement

Yanling Wang: Conceptualization, Methodology, Writing – original draft. **Qian Wang:** Writing – review & editing, Supervision. **Lingchen Zhao:** Writing – review & editing, Methodology. **Cong Wang:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This study was supported in part by the National Natural Science Foundation of China Grants U20B2049 and U21B2018, and in part by the Research Grants Council of Hong Kong under grants CityU 11218322, C2004-21G, R6021-20F, and N_CityU139/21.

References

[1] C. Xu, J. Ren, L. She, Y. Zhang, Z. Qin, K. Ren, EdgeSanitizer: Locally differentially private deep inference at the edge for mobile data analytics, IEEE Internet Things J. 6 (3) (2019) 5140–5151.

[2] W. Huang, S. Zhou, Y. Liao, H. Chen, An efficient differential privacy logistic classification mechanism, IEEE Internet Things J. 6 (6) (2019) 10620–10626.

[3] J. Jia, N.Z. Gong, Calibrate: Frequency estimation and heavy hitter identification with local differential privacy via incorporating prior knowledge, in: Proc. of IEEE INFOCOM, 2019.

[4] S. Ghane, L. Kulik, K. Ramamohanarao, TGM: A generative mechanism for publishing trajectories with differential privacy, IEEE Internet Things J. 7 (4) (2020) 2611–2621.

[5] M. Usman, M.A. Jan, D. Puthal, Paal: A framework based on authentication, aggregation, and local differential privacy for internet of multimedia things, IEEE Internet Things J. 7 (4) (2020) 2501–2508.

[6] G. Gao, M. Xiao, J. Wu, S. Zhang, L. Huang, G. Xiao, Dpdt: A differentially private crowd-sensed data trading mechanism, IEEE Internet Things J. 7 (1) (2020) 751–762.

[7] X. Nie, L.T. Yang, J. Feng, S. Zhang, Differentially private tensor train decomposition in edge-cloud computing for SDN-based internet of things, IEEE Internet Things J. 7 (7) (2020) 5695–5705.

[8] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, in: Proc. of IEEE S&P, 2017.

[9] A. Friedman, A. Schuster, Data mining with differential privacy, in: Proc. of ACM SIGKDD, 2010.

[10] C. Li, M. Hay, V. Rastogi, G. Miklau, A. McGregor, Optimizing linear counting queries under differential privacy, in: Proc. of ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, 2010.

[11] C. Dwork, Differential privacy: A survey of results, in: Proc. of Theory and Applications of Models of Computation, TAMC, 2008.

[12] M. Abadi, A. Chu, I. Goodfellow, H.B. McMahan, I. Mironov, K. Talwar, L. Zhang, Deep learning with differential privacy, in: Proc. of ACM CCS, 2016.

[13] B. Jayaraman, D. Evans, Evaluating differentially private machine learning in practice, in: Proc. of USENIX Security Symposium, 2019.

[14] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, K. Talwar, Semi-supervised knowledge transfer for deep learning from private training data, in: Proc. of ICLR, 2017.

[15] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, S. Jana, Certified robustness to adversarial examples with differential privacy, in: Proc. of IEEE S&P, 2019.

[16] E. Bagdasaryan, O. Poursaeed, V. Shmatikov, Differential privacy has disparate impact on model accuracy, in: Proc. of NIPS, 2019.

[17] M. Nasr, S. Song, A. Thakurta, N. Papernot, N. Carlini, Adversary instantiation: Lower bounds for differentially private machine learning, in: Proc. of IEEE S&P, 2021.

[18] Ú. Erlingsson, V. Pihur, A. Korolova, Rappor: Randomized aggregatable privacy-preserving ordinal response, in: Proc. of ACM CCS, 2014.

[19] Google, TensorFlow, 2019, Online at https://github.com/tensorflow/privacy.

[20] T. Farrand, F. Mireshghallah, S. Singh, A. Trask, Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy, in: Proc. of the Workshop on Privacy-Preserving Machine Learning in Practice, 2020.

[21] C. Tran, F. Fioretto, P. Van Hentenryck, Differentially private and fair deep learning: A lagrangian dual approach, in: Proc. of AAAI, 2021.

[22] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, A.L. Roth, Preserving statistical validity in adaptive data analysis, in: Proc. of ACM Symposium on Theory of Computing, 2015.

[23] S. Yeom, I. Giacomelli, M. Fredrikson, S. Jha, Privacy risk in machine learning: Analyzing the connection to overfitting, in: Proc. of IEEE Computer Security Foundations Symposium, CSF, 2018.

[24] B. Wu, S. Zhao, G. Sun, X. Zhang, Z. Su, C. Zeng, Z. Liu, P3sgd: Patient privacy preserving sgd for regularizing deep CNNs in pathological image classification, in: Proc. of IEEE/CVF CVPR, 2019.

[25] N. Phan, M. Vu, Y. Liu, R. Jin, D. Dou, X. Wu, M.T. Thai, Heterogeneous Gaussian mechanism: Preserving differential privacy in deep learning with provable robustness, in: Proc. of IJCAI, 2019.

[26] N. Xu, O. Feyisetan, A. Aggarwal, Z. Xu, N. Teissier, Differentially private adversarial robustness through randomized perturbations, 2020, arXiv:2009.12718, http://arxiv.org/abs/2009.12718.

[27] M. Du, R. Jia, D. Song, Robust anomaly detection and backdoor attack detection via differential privacy, in: Proc. of ICLR, 2020.

[28] J. Giraldo, A. Cardenas, M. Kantarcioglu, J. Katz, Adversarial classification under differential privacy, in: Proc. of NDSS, 2020.

[29] M.T. Hossain, S. Islam, S. Badsha, H. Shen, Desmp: Differential privacy-exploited stealthy model poisoning attacks in federated learning, 2021, arXiv:2109.09955, http://arxiv.org/abs/2109.09955.

[30] L. Zhou, L. Yu, S. Du, H. Zhu, C. Chen, Achieving differentially private location privacy in edge-assistant connected vehicles, IEEE Internet Things J. 6 (3) (2019) 4472–4481.

[31] X. Li, H. Zhang, Y. Ren, S. Ma, B. Luo, J. Weng, J. Ma, X. Huang, PAPU: Pseudonym swap with provable unlinkability based on differential privacy in VANETs, IEEE Internet Things J. 7 (12) (2020) 11789–11802.

[32] X. Lou, R. Tan, D.K. Yau, P. Cheng, Cost of differential privacy in demand reporting for smart grid economic dispatch, in: Proc. of IEEE INFOCOM, 2017.

[33] A. Ghosh, J. Ding, R. Sarkar, J. Gao, Differentially private range counting in planar graphs for spatial sensing, in: Proc. of IEEE INFOCOM, 2020.

[34] J. Wang, R. Zhu, S. Liu, A differentially private unscented Kalman filter for streaming data in IoT, IEEE Access 6 (2018) 6487–6495.

[35] T. Gao, F. Li, PHDP: Preserving persistent homology in differentially private graph publications, in: Proc. of IEEE INFOCOM, 2019.

[36] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, M. Backes, Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models, in: Proc. of NDSS, 2019.

[37] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, T. Ristenpart, Privacy in pharmacogenetics: An end-to-end case study of personalized Warfarin dosing, in: Proc. of USENIX Security Symposium, 2014.

[38] S. Mehnaz, S.V. Dibbo, R. De Viti, E. Kabir, B.B. Brandenburg, S. Mangard, N. Li, E. Bertino, M. Backes, E. De Cristofaro, et al., Are your sensitive attributes private? Novel model inversion attribute inference attacks on classification models, in: Proc. of USENIX Security Symposium, 2022.

[39] M. Fredrikson, S. Jha, T. Ristenpart, Model inversion attacks that exploit confidence information and basic countermeasures, in: Proc. of ACM CCS, 2015.

[40] E. Cohen, H. Kaplan, Y. Mansour, U. Stemmer, E. Tsfadia, Differentially-private clustering of easy instances, in: Proc. of ICML, 2021.

[41] M. Bun, M. Eliáš, J. Kulkarni, Differentially private correlation clustering, in: Proc. of ICML, 2021.

[42] M. Jones, H.L. Nguyen, T.D. Nguyen, Differentially private clustering via maximum coverage, in: Proc. of AAAI, 2021.

[43] H.L. Nguyen, A. Chaturvedi, E.Z. Xu, Differentially private k-means via exponential mechanism and max cover, in: Proc. of AAAI, 2021.

[44] A. Chaturvedi, H. Nguyen, L. Zakynthinou, Differentially private decomposable submodular maximization, in: Proc. of AAAI, 2021.

[45] J. Imola, T. Murakami, K. Chaudhuri, Locally differentially private analysis of graph statistics, in: Proc. of USENIX Security Symposium, 2021.

[46] A. De, S. Chakrabarti, Differentially private link prediction with protected connections, in: Proc. of AAAI, 2021.

[47] D. Nguyen, A. Vullikanti, Differentially private densest subgraph detection, in: Proc. of ICML, 2021.

[48] C. Yang, H. Wang, K. Zhang, L. Chen, L. Sun, Secure deep graph generation with link differential privacy, in: Proc. of IJCAI, 2021.

[49] S. Gopi, P. Gulhane, J. Kulkarni, J.H. Shen, M. Shokouhi, S. Yekhanin, Differentially private set union, in: Proc. of ICML, 2020.

[50] F. Zhao, X. Ren, S. Yang, Q. Han, P. Zhao, X. Yang, Latent dirichlet allocation model training with differential privacy, IEEE Trans. Inf. Forensics Secur. 16 (2021) 1290–1305.

[51] H. Kaplan, Y. Mansour, Y. Matias, U. Stemmer, Differentially private learning of geometric concepts, in: Proc. of ICML, 2019.

[52] M. Aliakbarpour, I. Diakonikolas, R. Rubinfeld, Differentially private identity and equivalence testing of discrete distributions, in: Proc. of ICML, 2018.

[53] M. Huai, D. Wang, C. Miao, J. Xu, A. Zhang, Pairwise learning with differential privacy guarantees, in: Proc. of AAAI, 2020.

[54] X. Zhou, J. Tan, Local differential privacy for Bayesian optimization, in: Proc. of AAAI, 2021.

[55] C. Li, P. Zhou, L. Xiong, Q. Wang, T. Wang, Differentially private distributed online learning, IEEE Trans. Knowl. Data Eng. 30 (8) (2018) 1440–1453.

[56] J. Abernethy, Y.H. Jung, C. Lee, A. McMillan, A. Tewari, Online learning via the differential privacy lens, in: Proc. of NIPS, 2019.

[57] A. Gonen, E. Hazan, S. Moran, Private learning implies online learning: An efficient reduction, in: Proc. of NIPS, 2019.

[58] H. Liu, J. Jia, N.Z. Gong, On the intrinsic differential privacy of bagging, in: Proc. of IJCAI, 2021.

[59] M. Bun, M.L. Carmosino, J. Sorrell, Efficient, noise-tolerant, and private learning via boosting, in: Proc. of Conference on Learning Theory, COLT, 2020.

[60] H. Nori, R. Caruana, Z. Bu, J.H. Shen, J. Kulkarni, Accuracy, interpretability, and differential privacy via explainable boosting, in: Proc. of ICML, 2021.

[61] D. Wang, J. Xu, Differentially private empirical risk minimization with smooth non-convex loss functions: A non-stationary view, in: Proc. of AAAI, 2019.

[62] K. Chaudhuri, C. Monteleoni, A.D. Sarwate, Differentially private empirical risk minimization, J. Mach. Learn. Res. 12 (3) (2011).

[63] D. Wang, J. Xu, On sparse linear regression in the local differential privacy model, in: Proc. of ICML, 2019.

[64] A. Rakotomamonjy, R. Liva, Differentially private sliced Wasserstein distance, in: Proc. of ICML, 2021.

[65] T. Ryffel, A. Trask, M. Dahl, B. Wagner, J. Mancuso, D. Rueckert, J. Passerat-Palmbach, A generic framework for privacy preserving deep learning, 2018, arXiv:1811.04017, http://arxiv.org/abs/1811.04017.

[66] J. Zhao, Y. Chen, W. Zhang, Differential privacy preservation in deep learning: Challenges, opportunities and solutions, IEEE Access 7 (2019) 48901–48911.

[67] M. Yang, L. Lyu, J. Zhao, T. Zhu, K.-Y. Lam, Local differential privacy and its applications: A comprehensive survey, 2020, arXiv:2008.03686, https://arxiv.org/abs/2008.03686.

[68] T. Zhu, D. Ye, W. Wang, W. Zhou, P.S. Yu, More than privacy: Applying differential privacy in key areas of artificial intelligence, 2020, arXiv:2008.01916, https://arxiv.org/abs/2008.01916.

[69] Q. Wang, Z. Li, Q. Zou, L. Zhao, S. Wang, Deep domain adaptation with differential privacy, IEEE Trans. Inf. Forensics Secur. 15 (2020) 3093–3106.

[70] Q. Zheng, J. Dong, Q. Long, W. Su, Sharp composition bounds for Gaussian differential privacy via edgeworth expansion, in: Proc. of ICML, 2021.

[71] F. Fioretto, C. Tran, P. Van Hentenryck, Decision making with differential privacy under a fairness lens, in: Proc. of IJCAI, 2021.

[72] Z. Ji, Z.C. Lipton, C. Elkan, Differential privacy and machine learning: A survey and review, 2014, arXiv:1412.7584, http://arxiv.org/abs/1412.7584.

[73] L. Fan, A survey of differentially private generative adversarial networks, in: Proc. of AAAI Workshop on Privacy-Preserving Artificial Intelligence, 2020.

[74] P. Zhao, G. Zhang, S. Wan, G. Liu, T. Umer, A survey of local differential privacy for securing Internet of Vehicles, J. Supercomput. 76 (11) (2020) 8391–8412.

[75] L. Zhao, L. Ni, S. Hu, Y. Chen, P. Zhou, F. Xiao, L. Wu, Inprivate digging: Enabling tree-based distributed data mining with differential privacy, in: Proc. of IEEE INFOCOM, 2018.

[76] Y. Qu, S. Yu, L. Gao, S. Peng, Y. Xiang, L. Xiao, FuzzyDP: Fuzzy-based big data publishing against inquiry attacks, in: Proc. of IEEE INFOCOM Workshops, 2017.

[77] S. Wang, L. Huang, Y. Nie, P. Wang, H. Xu, W. Yang, Privset: Set-valued data analyses with locale differential privacy, in: Proc. of IEEE INFOCOM, 2018.

[78] C. Dwork, Differential privacy, in: Proc. of International Colloquium on Automata, Languages, and Programming, 2006.

[79] I. Mironov, Rényi differential privacy, in: Proc. of IEEE Computer Security Foundations Symposium, CSF, 2017.

[80] R. Sarathy, K. Muralidhar, Evaluating Laplace noise addition to satisfy differential privacy for numeric data, Trans. Data Priv. 4 (1) (2011) 1–17.

[81] F. Liu, Generalized Gaussian mechanism for differential privacy, IEEE Trans. Knowl. Data Eng. 31 (4) (2018) 747–756.

[82] B. Balle, Y.-X. Wang, Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising, in: Proc. of ICML, 2018.

[83] B. Ding, J. Kulkarni, S. Yekhanin, Collecting telemetry data privately, in: Proc. of NIPS, 2017.

[84] X. Cao, J. Jia, N.Z. Gong, Data poisoning attacks to local differential privacy protocols, in: Proc. of USENIX Security Symposium, 2021.

[85] A. Cheu, A. Smith, J. Ullman, Manipulation attacks in local differential privacy, in: Proc. of IEEE S&P, 2021.

[86] J. Acharya, K. Bonawitz, P. Kairouz, D. Ramage, Z. Sun, Context aware local differential privacy, in: Proc. of ICML, 2020.

[87] F. McSherry, K. Talwar, Mechanism design via differential privacy, in: Proc. of IEEE Symposium on Foundations of Computer Science, FOCS, 2007.

[88] F.D. McSherry, Privacy integrated queries: An extensible platform for privacy-preserving data analysis, in: Proc. of ACM SIGMOD, 2009.

[89] B. Bichsel, T. Gehr, D. Drachsler-Cohen, P. Tsankov, M. Vechev, Dp-finder: Finding differential privacy violations by sampling and optimization, in: Proc. of ACM CCS, 2018.

[90] B. Bichsel, S. Steffen, I. Bogunovic, M. Vechev, DP-sniper: Black-box discovery of differential privacy violations using classifiers, in: Proc. of IEEE S&P, 2021.

[91] C. Dwork, G.N. Rothblum, S. Vadhan, Boosting and differential privacy, in: Proc. of IEEE Annual Symposium on Foundations of Computer Science, 2010.

[92] K. Zhu, P. Van Hentenryck, F. Fioretto, Bias and variance of post-processing in differential privacy, in: Proc. of AAAI, 2021.

[93] R. Shokri, V. Shmatikov, Privacy-preserving deep learning, in: Proc. of ACM CCS, 2015.

[94] H.B. McMahan, D. Ramage, K. Talwar, L. Zhang, Learning differentially private recurrent language models, in: Proc. of ICLR, 2018.

[95] N. Phan, X. Wu, H. Hu, D. Dou, Adaptive laplace mechanism: Differential privacy preservation in deep learning, in: Proc. of ICDM, 2017.

[96] N. Papernot, A. Thakurta, S. Song, S. Chien, U. Erlingsson, Tempered sigmoid activations for deep learning with differential privacy, in: Proc. of AAAI, 2021.

[97] N. Agarwal, A.T. Suresh, F. Yu, S. Kumar, H.B. Mcmahan, cpSGD: Communication-efficient and differentially-private distributed SGD, in: Proc. of NIPS, 2018.

[98] H.B. McMahan, G. Andrew, U. Erlingsson, S. Chien, I. Mironov, N. Papernot, P. Kairouz, A general approach to adding differential privacy to iterative training procedures, 2018, arXiv:1812.06210, http://arxiv.org/abs/1812.06210.

[99] L. Xiang, J. Yang, B. Li, Differentially-private deep learning from an optimization perspective, in: Proc. of IEEE INFOCOM, 2019.

[100] L. Yu, L. Liu, C. Pu, M.E. Gursoy, S. Truex, Differentially private model publishing for deep learning, in: Proc. of IEEE S&P, 2019.

[101] Z. Bu, J. Dong, Q. Long, W.J. Su, Deep learning with Gaussian differential privacy, Harvard Data Sci. Rev. 2020 (23) (2020).

[102] L. Jiang, X. Lou, R. Tan, J. Zhao, Differentially private collaborative learning for the IoT edge, in: Proc. of International Conference on Embedded Wireless Systems and Networks, EWSN, 2019.

[103] K. Wei, J. Li, M. Ding, C. Ma, H.H. Yang, F. Farokhi, S. Jin, T.Q. Quek, H.V. Poor, Federated learning with differential privacy: Algorithms and performance analysis, IEEE Trans. Inf. Forensics Secur. 15 (2020) 3454–3469.

[104] S. Truex, L. Liu, K.-H. Chow, M.E. Gursoy, W. Wei, LDP-Fed: Federated learning with local differential privacy, in: Proc. of ACM International Workshop on Edge Systems, Analytics and Networking, 2020.

[105] Y. Wang, Y. Tong, D. Shi, Federated latent dirichlet allocation: A local differential privacy based framework, in: Proc. of AAAI, 2020.

[106] L. Xie, K. Lin, S. Wang, F. Wang, J. Zhou, Differentially private generative adversarial network, 2018, arXiv:1802.06739, http://arxiv.org/abs/1802.06739.

[107] X. Zhang, S. Ji, T. Wang, Differentially private releasing via deep generative model (technical report), 2018, arXiv:1801.01594, http://arxiv.org/abs/1801.01594.

[108] C. Xu, J. Ren, D. Zhang, Y. Zhang, Z. Qin, K. Ren, GANobfuscator: Mitigating information leakage under GAN via differential privacy, IEEE Trans. Inf. Forensics Secur. 14 (9) (2019) 2358–2371.

[109] R. Torkzadehmahani, P. Kairouz, B. Paten, Dp-cgan: Differentially private synthetic data and label generation, in: Proc. of CVPR Workshops, 2019.

[110] Z. Huang, R. Hu, Y. Guo, E. Chan-Tin, Y. Gong, DP-ADMM: ADMM-based distributed learning with differential privacy, IEEE Trans. Inf. Forensics Secur. 15 (2020) 1002–1012.

[111] N. Phan, X. Wu, D. Dou, Preserving differential privacy in convolutional deep belief networks, Mach. Learn. 106 (9) (2017) 1681–1704.

[112] J. Li, M. Khodak, S. Caldas, A. Talwalkar, Differentially private meta-learning, in: Proc. of ICLR, 2020.

[113] J. Jordon, J. Yoon, M. Van Der Schaar, PATE-GAN: Generating synthetic data with differential privacy guarantees, in: Proc. of ICLR, 2019.

[114] G. Damaskinos, C. Mendler-Dünner, R. Guerraoui, N. Papandreou, T. Parnell, Differentially private stochastic coordinate descent, in: Proc. of AAAI, 2021.

[115] J. Domingo-Ferrer, D. Sánchez, A. Blanco-Justicia, The limits of differential privacy (and its misuse in data release and machine learning), Commun. ACM 64 (7) (2021) 33–35.

[116] Z. Xu, S. Shi, A.X. Liu, J. Zhao, L. Chen, An adaptive and fast convergent approach to differentially private deep learning, in: Proc. of IEEE INFOCOM, 2020.

[117] J. Dong, A. Roth, W.J. Su, Gaussian differential privacy, 2019, arXiv:1905.02383, http://arxiv.org/abs/1905.02383.

[118] D. Yu, H. Zhang, W. Chen, T.-Y. Liu, J. Yin, Gradient perturbation is underrated for differentially private convex optimization, in: Proc. of IJCAI, 2020.

[119] M. Jagielski, J. Ullman, A. Oprea, Auditing differentially private machine learning: How private is private SGD? in: Proc. of NIPS, 2020.

[120] S. Truex, N. Baracaldo, A. Anwar, T. Steinke, H. Ludwig, R. Zhang, Y. Zhou, A hybrid approach to privacy-preserving federated learning, in: Proc. of ACM Workshop on Artificial Intelligence and Security, 2019.

[121] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, 2014, arXiv preprint arXiv:1406.2661, http://arxiv.org/abs/1406.2661.

[122] B. Hitaj, G. Ateniese, F. Perez-Cruz, Deep models under the GAN: Information leakage from collaborative deep learning, in: Proc. of ACM CCS, 2017.

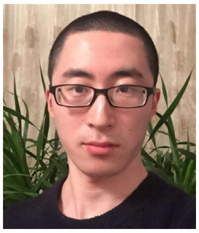
[123] J. Konečnỳ, H.B. McMahan, F.X. Yu, P. Richtárik, A.T. Suresh, D. Bacon, Federated learning: Strategies for improving communication efficiency, 2016, arXiv:1610.05492, https://arxiv.org/abs/1610.05492.

[124] R.C. Geyer, T. Klein, M. Nabi, Differentially private federated learning: A client level perspective, 2017, arXiv:1712.07557, http://arxiv.org/abs/1712.07557.

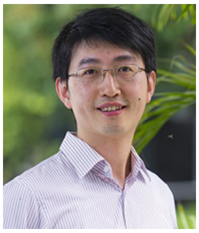
[125] M. Hao, H. Li, G. Xu, S. Liu, H. Yang, Towards efficient and privacy-preserving federated deep learning, in: Proc. of IEEE International Conference on Communications, ICC, 2019.
[126] N. Rodríguez-Barroso, D. Jiménez-López, M.V. Luzón, F. Herrera, E. Martínez-Cámara, Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges, Inf. Fusion 90 (2023) 148–173.
[127] N. Rodríguez-Barroso, G. Stipcich, D. Jiménez-López, J.A. Ruiz-Millán, E. Martínez-Cámara, G. González-Seco, M.V. Luzón, M.A. Veganzones, F. Herrera, Federated learning and differential privacy: Software tools analysis, the Sherpa. AI FL framework and methodological guidelines for preserving data privacy, Inf. Fusion 64 (2020) 270–292.
[128] M. Naseri, J. Hayes, E. De Cristofaro, Local and central differential privacy for robustness and privacy in federated learning, 2020, arXiv:2009.03561, https://arxiv.org/abs/2009.03561.
[129] O. Choudhury, A. Gkoulalas-Divanis, T. Salonidis, I. Sylla, Y. Park, G. Hsu, A. Das, Differential privacy-enabled federated learning for sensitive health data, 2019, arXiv:1910.02578, https://arxiv.org/abs/1910.02578.
[130] L. Sun, L. Lyu, Federated model distillation with noise-free differential privacy, in: Proc. of IJCAI, 2021.
[131] R. Hu, Y. Guo, H. Li, Q. Pei, Y. Gong, Personalized federated learning with differential privacy, IEEE Internet Things J. 7 (10) (2020) 9530–9539.
[132] L. Sun, J. Qian, X. Chen, P.S. Yu, Ldp-fl: Practical private aggregation in federated learning with local differential privacy, in: Proc. of IJCAI, 2021.
[133] H. Phan, M.T. Thai, H. Hu, R. Jin, T. Sun, D. Dou, Scalable differential privacy with certified robustness in adversarial learning, in: Proc. of ICML, 2020.
[134] F. Farokhi, N. Wu, D.B. Smith, M.A. Kâafar, The cost of privacy in asynchronous differentially-private machine learning, IEEE Trans. Inf. Forensics Secur. 16 (2021) 2118–2129.
[135] N. Hynes, R. Cheng, D. Song, Efficient deep learning on multi-source private data, 2018, arXiv:1807.06689, http://arxiv.org/abs/1807.06689.
[136] B.K. Beaulieu-Jones, W. Yuan, S.G. Finlayson, Z.S. Wu, Privacy-preserving distributed deep learning for clinical data, 2018, arXiv:1812.01484, http://arxiv.org/abs/1812.01484.
[137] L. Fan, Image pixelization with differential privacy, in: Proc. of IFIP Annual Conference on Data and Applications Security and Privacy, 2018.
[138] F. Tramèr, D. Boneh, Differentially private learning needs better features (or much more data), in: Proc. of ICLR, 2021.
[139] P.C.M. Arachchige, P. Bertok, I. Khalil, D. Liu, S. Camtepe, M. Atiquzzaman, Local differential privacy for deep learning, IEEE Internet Things J. 7 (7) (2020) 5827–5842.
[140] R. Cummings, V. Gupta, D. Kimpara, J. Morgenstern, On the compatibility of privacy and fairness, in: Proc. of Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization, 2019.
[141] J. Ding, X. Zhang, X. Li, J. Wang, R. Yu, M. Pan, Differentially private and fair classification via calibrated functional mechanism, in: Proc. of AAAI, 2020.
[142] M. Padala, S. Damle, S. Gujar, Federated learning meets fairness and differential privacy, 2021, arXiv:2108.09932, http://arxiv.org/abs/2108.09932.

Yanling Wang is working towards the Ph.D. degree in the School of Cyber Science and Engineering, Wuhan University, China. She is also a joint Ph.D. candidate in City University of Hong Kong, Hong Kong, China. She received the B.E. degree in Computer Science from Wuhan University, China, in 2019. Her research interests include network security and AI security.

Qian Wang is a full professor in School of Cyber Science and Engineering at Wuhan University, China. He was selected into the National High-level Young Talents Program of China, and listed among the World's Top 2% Scientists by Stanford University. He has been engaged in the research of cyberspace security, with focus on AI security, data outsourcing security and privacy, wireless systems security, and applied cryptography. He was a recipient of the 2018 IEEE TCSC Award for Excellence in Scalable Computing and the 2016 IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award. He is a Fellow of the IEEE.



Lingchen Zhao received his Ph. D. degree in Cyberspace Security in 2021, from Wuhan University, China, and his B. E. degree in Information Security in 2016, from Central South University, China. He held a postdoctoral position with the City University of Hong Kong from 2021 to 2022. He is currently an associate professor with the School of Cyber Science and Engineering, Wuhan University. His research interests include data security and AI security.



Cong Wang is a Professor at the Department of Computer Science, City University of Hong Kong. His research interests are data and network security, blockchain and decentralized applications, and privacy-enhancing technologies. At CityU, he received the Outstanding Researcher Award (2019), the Outstanding Supervisor Award (2017), and the President's Awards (2016 and 2019). He is a Founding Member of the Young Academy of Sciences of Hong Kong and a Research Fellow of the Hong Kong Research Grants Council. Since January 2023, he has been the Editor-in-Chief for IEEE TDSC. He is a Fellow of the IEEE.