

Survey on Privacy-Preserving Techniques for Federated Learning

Jiaqi Lu ^a

College of Electronics and Information Engineering, Beibu Gulf University, Qinzhou, Guangxi, 535011, China

Keywords: Federated Learning, Privacy Protection, Privacy Threats, Artificial Intelligence.

Abstract: Federated learning (FL) is a process that allows multiple participants to train a model locally and share the model parameters. This approach reduces the risk of data leakage. To improve privacy protection, researchers have proposed a range of techniques that are designed to preserve privacy. These include differential privacy, homomorphic encryption, and secure multi-party computation, which enhance privacy protection by operating at different levels. Nevertheless, further research is required to achieve a balance between privacy and model performance in the context of FL. The present paper commences with an exposition of the notion of FL, clarifying its definition and rationale. Subsequently, a comprehensive review of extant architectures and classifications of FL is presented. The subsequent discussion focuses on the root causes of privacy threats in FL and analyses the risks that may be caused by data sharing and other links. On this basis, the advantages of privacy-preserving techniques such as differential privacy, homomorphic encryption, and secure multi-party computation in combination with FL are described in detail. These advantages include enhanced data security and privacy protection. The limitations of these techniques are also discussed. Finally, it comprehensively discusses the challenges of privacy protection in FL, such as the contradictory nature of ensuring both high model accuracy and efficient algorithms and the lack of unified quantitative standards. The paper also provides an outlook on future development directions, which will doubtless serve as a reference for subsequent research and practice.


1 INTRODUCTION

Artificial intelligence has been widely applied in many different industries in recent years. Federated learning (FL) is a new distributed machine learning system that addresses data privacy issues by distributing model parameter updates among several computers without transferring the original data. However, the data's sensitivity creates significant difficulties for model training. There is mounting concern among the public, the media, and governments worldwide regarding data leaks. For example, Facebook suffered a massive data breach, which has led to increased public attention to privacy and security issues. The European Union (EU), to strengthen data security and privacy, has issued the General Data Protection Regulation (GDPR), which provides for the protection of the personal data and privacy rights of EU citizens, as well as regulating the processing and use of personal data by businesses (Goddard, 2017). China has also introduced laws to emphasise the importance of personal privacy

(National People's Congress of the People's Republic of China, 2021).

Recent years have seen considerable progress in the application of privacy-preserving techniques in the context of FL. For example, there has been a notable advancement in the field of differential privacy techniques, particularly in the integration of local and global differential privacy strategies. A recent study proposes an approach that combines local differential privacy (LDP) and global differential privacy (GDP) in a manner that safeguards data privacy while minimising the impact on model performance (Zhu, &Chen, 2025). In addition, techniques such as secure multi-party computation and homomorphic encryption have been further explored in FL. These techniques ensure that the model update information of the client is not leaked even in an untrusted server environment by encrypting the process.

Based on this, this paper aims to provide a comprehensive theoretical framework and practical guidance for privacy protection in FL. By deeply analysing the privacy threats and their protection

^a <https://orcid.org/0009-0002-9840-0581>

techniques in FL, it reveals its potential and challenges in data privacy protection and provides valuable references for researchers and practitioners.

2 FEDERATED LEARNING

2.1 Introduction to Federated Learning

To solve the privacy protection issue, Google developed federated learning techniques (McMahan, Moore, & Ramage, 2017). The technology effectively protects user privacy and security by keeping the data storage and model training process on the local device and exchanging model update information only with a central server.

In a practical application scenario, Suppose N clients $\{E_1, \dots, E_i\}$ hold their own dataset $\{D_1, \dots, D_i\}$ and cannot access each other's data directly. It contains four basic steps:

- The server sends an initial model to each client;
- Client E_i does not need to share its resource data and trains its model M_i with local data D_i ;
- The server aggregates the collected individual local models $\{M_1, \dots, M_i\}$ into a global model M' ;
- Issues updates of the global model M' to each client's local model.

From this, it appears that federated learning technologies have the following characteristics: the data of all parties are kept locally, which effectively avoids the problems of privacy leakage and violation of laws. Secondly, under the federated learning system, the identities and status of each participant are equal and no one party dominates. Finally, the modelling effect of FL is comparable to or less different from the effect of modelling the whole dataset centrally, which ensures the performance and accuracy of the model.

2.2 The Architecture of Federated Learning

2.2.1 Centralized Federated Learning Architecture

A central server acts as the coordinator in this architecture, transmitting the original model to the clients and compiling the model updates—usually model parameters or gradients—that the clients provide. Using local data, the client trains the model and communicates updates to the central server. These updates are combined by the central server to

create a global model, which is then repeated until the model converges or training is finished.

2.2.2 Decentralized Federated Learning Architecture

Parties communicate directly with each other in this architecture without a central server. Each party updates and encrypts its model and sends it to other parties. This architecture requires more cryptographic operations and all model parameter interactions are encrypted to ensure data security and privacy. At present, it can be realized by secure multi-party computation, homomorphic encryption, and other technologies.

2.3 Categorization of Federated Learning

FL can be divided into three categories based on the distribution of the data: federated transfer learning, vertical federated learning, and horizontal federated learning. The danger of privacy leakage and protection strategies is also impacted by the training and intermediate settings needed for various data division techniques.

Suppose L_u represents the data held by client u , L_p represents the data held by client p , J_u represents the sample ID of client u , J_p represents the sample ID of p , Y_u represents the dataset label information of u , Y_p represents the dataset label information of p , X_u represents the dataset feature information of u , and X_p represents the dataset feature information of p . Therefore a complete training data set D should be composed of (J, Y, X) (Yang, Liu, & Chen, 2019).

2.3.1 Horizontal Federated Learning (HFL)

HFL is suitable for data where multiple parties have different users, but these data have the same set of features. The formula is expressed as follows.

$$X_u = X_p, Y_u = Y_p, J_u \neq J_p, \forall L_u, \forall L_p, u \neq p (1)$$

In HFL, the datasets of each party are identical in the feature space but non-overlapping in the sample space, that is the datasets contain different instances or records but each instance has the same attributes or features. Gradient processing and communication in HFL may reveal users' private data. A common solution is to perform homomorphic encryption, differential privacy, and secure aggregation as a way to ensure security when exchanging gradients in HFL.

2.3.2 Vertical Federated Learning (VFL)

VFL applies when multiple participants have data about different characteristics of the same set of users. The formula is expressed as follows.

$$X_u \neq X_p, Y_u \neq Y_p, J_u = J_p, \forall L_u, \forall L_p, u \neq p \quad (2)$$

In the VFL, participants share the same sample but each possesses distinct feature data. To illustrate, consider a scenario where a bank and an e-commerce company possess a similar customer base. However, the nuances of their respective data sets diverge. While the bank holds information pertaining to the customer's financial transactions, the e-commerce company focuses on customer shopping behaviour patterns. This scenario is an exemplification of the application of VFL.

2.3.3 Federated Transfer Learning (FTL)

FTL is suitable for participants who may have a small number of overlapping or completely different datasets, but all hope to utilise the knowledge of the other participants to improve the performance of their own models. The formula is expressed as follows.

$$X_u \neq X_p, Y_u \neq Y_p, J_u \neq J_p, \forall L_u, \forall L_p, u \neq p \quad (3)$$

By integrating the decentralized training approach of FL with the knowledge transfer capabilities of transfer learning, participants can collaborate to develop a shared global model without the need to exchange raw data, thereby preserving their data privacy. The primary aim of FTL is to leverage prior knowledge to enhance the learning process for new tasks, addressing issues related to insufficient data and labels.

In the above three types of FL, the data of each participant is always kept locally and the data exists independently. The parameters exchanged during joint training are encrypted and the communication also adopts strict encryption algorithms, making it difficult to leak the original data information. Compared with traditional centralised machine learning training, it has higher privacy protection. However, FL itself does not provide comprehensive and sufficient privacy protection and still faces the threat of information leakage. Only by recognizing the risk of privacy leakage can the overall direction of privacy protection methods for FL be found. As the current protection methods of VFL and FTL are similar to HFL, the privacy protection technologies investigated below are HFL scenarios unless otherwise specified.

3 ROOTS OF PRIVACY THREATS IN FEDERATED LEARNING

FL reduces the risk of leakage of centrally stored data by processing and training data on edge devices but also poses new privacy protection challenges. In horizontal federated learning, participants need to upload gradient parameters to the server for aggregation, which may leak private information about the local data. In vertical federated learning, a malicious adversary may steal sensitive information through gradient swapping due to overlapping data features but not shared labels. Federated transfer learning is also prone to exposing private information in the process of data backpropagation. This subsection lists the major privacy threats in FL.

3.1 Unauthorized Access and Extraction by Malicious Participants

In FL, model gradients and parameter updates are shared among multiple participants even though the data never leaves the local device. This sharing mechanism provides potential attack opportunities for malicious participants. For example, one participant may try to infer the characteristics of other participants' data by analysing the uploaded gradient. This phenomenon is known as a gradient leakage attack (Yang, Ge, & Xue, 2023). In addition, if a participant is infected with malware, it may unknowingly leak data or intentionally upload gradients containing malicious code to disrupt the FL process or steal data from other participants. Research has now evaluated gradient inversion attacks and their defense mechanisms, pointing out that a malicious participant can partially recover the client's private data by analysing the gradient (Huang, Gupta, & Song, 2021).

3.2 Privacy Breaches and Attacks on the Central Server

The central server plays the role of aggregating the gradients of each participant in FL, thus it becomes a potential privacy leakage point. If the server is under the control of an attacker or if there is an internal malicious actor, then the uploaded gradient information may be intercepted and analysed to reveal sensitive data (Sharma, & Marchang, 2024). Furthermore, software and hardware vulnerabilities of servers can also be exploited to steal or tamper with data. To protect privacy, server security needs to be ensured. Including the use of encryption to protect

data transmission and the implementation of strict access control and monitoring mechanisms.

3.3 Malicious Multi-Party Conspiracy to Steal Privacy

In FL, the cooperation of multiple participants is the key to improving model performance. If multiple participants collude, however, they may work together to analyse gradient information to extrapolate data from other participants (Luo, Li, & Qin, 2024). This collusive attack is more difficult to defend against than an attack by a single participant because it involves the concerted action of multiple seemingly legitimate participants. To combat this threat, mechanisms need to be devised that can detect and deter collusive behaviour. For example, by randomising the gradient or using differential privacy techniques to make it harder for attackers to steal information.

4 PRIVACY PROTECTION TECHNOLOGY IN FEDERATED LEARNING

4.1 Secure Multi-Party Computation (SMPC)

SMPC allows multiple participants to collaboratively compute an agreed function without mutual trust and a trusted third party. It also ensures that each participant cannot extrapolate the raw data of the other participants from the data interacted during the calculation process, except for the results of the calculation. This means that privacy and security are guaranteed in multi-party data fusion calculations, where each participant has absolute control over the data it owns, guaranteeing that basic data and information will not be leaked. However, the implementation of SMPC is relatively complex, which can lead to performance bottlenecks and increased computational and communication overheads (Gamiz, Regueiro, & Lage, 2025). To solve this issue, researchers have investigated the problem of latency in SMPC. They have proposed a lazy sharing approach with a view to reducing communication overhead and computational burden (Li, Zhang, & Lin, 2024).

SMPC allows multiple participants to collaborate on computations without revealing their private data, so it is well suited for collaboration between multiple organisations. This ability to collaborate with multiple parties makes SMPC useful in scenarios that require joint computation by multiple independent

data providers, such as financial collaboration or analysis of healthcare data. In response to the central server and multiple participant collusion, researchers have proposed a jointly secure multi-party deep learning protocol that incorporates additive homomorphic encryption and differential privacy, which can demonstrate better security, accuracy, and efficiency in supporting FL for large-scale user scenarios (Hao, Li, & Luo, 2019). Experimental results on the MNIST dataset show that the protocol achieves an average accuracy of 90.8% at $\epsilon = 0.5, \delta = 10^{-5}$ and 97.5% at $\epsilon = 2, \delta = 10^{-5}$, which is a good indication of the protocol's ability to maintain high accuracy and efficiency while protecting privacy (Hao, Li, & Luo, 2019).

4.2 Differential Privacy (DP)

The core idea of DP is to add noise to the computation process to ensure that information about individual data points cannot be extrapolated backward from the results of the analysis. In FL, DP is mainly applied in the aggregation phase of model parameters. Instead of directly aggregating the local model parameters from each participant upon receipt, the central server first trims the parameters to limit their range and then adds noise to obscure their exact values. In this way, even if the attacker intercepts the noisy model parameters, the original data of any participant cannot be recovered from them. Current research has explored the use of DP in data analysis, pointing out that adding noise to model updates can effectively prevent attackers from inferring information about individual training samples from the model (Subramanian, 2023). The researchers present a knowledge transfer method called PrivateKT that aims to enable effective and privacy-preserving knowledge transfer in FL (Qi, Wu, & Wu, 2023). Experimental results show that on the MNIST dataset, PrivateKT's accuracy is only 2.5% lower than that of centralised learning under $\epsilon = 2$, which proves that PrivateKT can effectively transfer high-quality knowledge while protecting privacy (Qi, Wu, & Wu, 2023).

Noise introduced by DP may lead to degradation of model performance. It has been shown that adding noise can provide better privacy protection, but may harm the accuracy of the model (Zhang, Mao, & Tu, 2023). Some researchers have successfully deployed differential privacy within a FL framework for analysing health-related data, but experiments have demonstrated that DP may result in large function loss values (Choudhury, Gkoulalas-Divanis, & Saloniadis, 2019). Therefore, the balance between DP and model accuracy needs to be carefully studied and adjusted.

4.3 Homomorphic Encryption (HE)

HE is a special type of encryption. It allows calculations to be performed on encrypted data and decrypted when the result is obtained and the result is the same as if the same calculation had been performed on the original data. The core feature of this technology is that the data can be counted invisible, which means that the data can be processed and analysed without decryption. The absence of transmission of both the data itself and the underlying models ensures that the other party is unable to make any deductions. Consequently, the probability of leakage at the raw data level is minimal (Hong, 2025).

HE allows computation on encrypted data, meaning that clients can train local models without decrypting the data, thus protecting the privacy of client data. However, HE is usually computationally intensive and can increase the computational burden on the client, especially on resource-constrained devices. HE can be a performance bottleneck in FL systems due to the high computational overhead. Researchers have now proposed FedML-HE, an efficient privacy-preserving FL system based on homomorphic encryption, which employs selective parameter encryption and significantly reduces communication and computation overheads, making FL based on HE more efficient in real scenarios (Jin, Yao, & Han, 2023). Experiments have demonstrated that FedML-HE significantly reduces overheads, by a factor of about 10 for the ResNet-50 model in terms of both computation time and communication file size, and by a factor of about 40 for the BERT model (Jin, Yao, & Han, 2023).

5 CHALLENGES AND FUTURE DIRECTIONS

Due to its unique architecture and training process, FL encounters various privacy invasion methods and urgent privacy protection requirements. To predict the future direction of FL privacy protection, it is important to first identify and understand the problems encountered in the balance, cost-effectiveness, and practical application of current privacy protection technologies.

5.1 Privacy-Preserving Approaches or Techniques for Vertical Federated Learning and Federated Transfer Learning

The primary focus of current research endeavours pertains to the realm of privacy protection and

security defence in the context of horizontal federated learning. Relatively little research has been done on longitudinal federated learning and federated transfer learning (Liu, Lv, & Guo, 2024). These two models require deeper levels of data interaction and cooperation, so the development of reliable security protocols or hybrid strategies integrating multiple privacy-preserving technologies to achieve optimal privacy protection at each step is the key to future research.

5.2 Balancing the Contradictions of Privacy Protection, Model Accuracy, and Algorithm Efficiency

Building efficient and highly accurate privacy-preserving security algorithms is the main problem that needs to be addressed by current FL. Existing privacy protection methods generally enhance privacy protection at the expense of efficiency or model accuracy (Zhang, Kang, & Chen, 2023). If the encryption degree is too weak, it will increase the risk of privacy leakage, while if the encryption degree is too strong, it will cause large computational overhead and may affect the global model performance. Computational and communication overheads need to be addressed for privacy-preserving technologies, especially encryption technologies. Future research can develop privacy protection schemes that integrate multiple technologies, reasonably select suitable privacy protection schemes for different application scenarios, reduce scheme complexity, and optimise model accuracy and training efficiency. In this way, privacy can be protected while avoiding large performance loss and data encryption protection and communication security can be complementary guarantees.

5.3 Establish the Metrics of Privacy Leakage and Privacy Protection Degree

FL currently lacks a unified privacy metric, making it difficult for researchers to accurately assess the effectiveness of privacy-protecting programmes and for users to know exactly how well they are protecting their privacy (Jagarlamudi, Yazdinejad, & Parizi, 2024). Attempts to systematically measure the degree of user privacy protection and the amount of protection provided by different technologies can help refine evaluation metrics and facilitate iterative research on privacy attacks and protection schemes. At the same time, the privacy leakage risk assessment system of each link in the FL system is not perfect.

For example, the aggregation results observed by the server in the secure aggregation mechanism may cause privacy leakage, so it is necessary to further study and evaluate the risk of exposure of intermediate parameters. Overall, the construction of a unified and perfect privacy protection metric is crucial for privacy protection in FL systems, which can provide evaluation metrics and promote the development and optimisation of privacy protection techniques.

6 CONCLUSION

In recent years, with the rapid development of artificial intelligence, there have been numerous reports on the Internet about artificial intelligence leaking personal privacy, and people have begun to pay more attention to data privacy. The emergence of FL has brought new hope and methods to researchers to a certain extent. However, with the deepening of research, FL also faces the risk of privacy leakage different from other machine learning methods.

The present paper conducts an exhaustive investigation and in-depth analysis of the latest research on privacy leakage risks and protection technologies of FL. The architecture and classification of FL are introduced, the root causes of privacy risks are analysed and a list of three privacy protection technologies is provided: secure multi-party computation, differential privacy and homomorphic encryption. This paper first briefly introduces these technologies, and then comprehensively analyses their communication efficiency and privacy protection effects when combined with FL. Among them, secure multi-party computation is suitable for protecting multi-agency collaboration scenarios, differential privacy protects data parameters by adding noise, and homomorphic encryption focuses on protecting the original data. Finally, according to the shortcomings of the existing research, the future research direction was discussed. In the process of realising FL applications, there are still some unresolved challenges. In particular, the three major issues of developing privacy-preserving solutions applicable to different types of FL, balancing the contradiction between accuracy and efficiency, and establishing a unified metric deserve more in-depth research.

REFERENCES

- Choudhury, O., Gkoulalas-Divanis, A., Salonidis, T., Sylla, I., Park, Y., Hsu, G., & Das, A. 2019. Differential privacy-enabled federated learning for sensitive health data. *arXiv preprint arXiv:1910.02578*.
- Gamiz, I., Regueiro, C., Lage, O., Jacob, E., & Astorga, J. 2025. Challenges and future research directions in secure multi-party computation for resource-constrained devices and large-scale computations. *International Journal of Information Security*, 24(1), 1-29.
- Goddard, M. 2017. The EU general data protection regulation (GDPR): European regulation that has a global impact. *International Journal of Market Research*, 59(6), 703-705.
- Hao, M., Li, H., Luo, X., Xu, G., Yang, H., & Liu, S. 2019. Efficient and privacy-enhanced federated learning for industrial artificial intelligence. *IEEE Transactions on Industrial Informatics*, 16(10), 6532-6542.
- Hong, C. 2025. Recent advances of privacy-preserving machine learning based on (Fully) Homomorphic Encryption. *Security and Safety*, 4, 2024012.
- Huang, Y., Gupta, S., Song, Z., Li, K., & Arora, S. 2021. Evaluating gradient inversion attacks and defenses in federated learning. *Advances in neural information processing systems*, 34, 7232-7241.
- Jagarlamudi, G. K., Yazdinejad, A., Parizi, R. M., & Pouriyeh, S. 2024. Exploring privacy measurement in federated learning. *The Journal of Supercomputing*, 80(8), 10511-10551.
- Jin, W., Yao, Y., Han, S., Gu, J., Joe-Wong, C., Ravi, S., ... & He, C. 2023. FedML-HE: An efficient homomorphic-encryption-based privacy-preserving federated learning system. *arXiv preprint arXiv:2303.10837*.
- Li, S., Zhang, C., & Lin, D. 2024. Secure Multiparty Computation with Lazy Sharing. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security* (pp. 795-809).
- Liu, B., Lv, N., Guo, Y., & Li, Y. 2024. Recent advances on federated learning: A systematic survey. *Neurocomputing*, 128019.
- Luo, Y., Li, Y., Qin, S., Fu, Q., & Liu, J. 2024. Copyright protection framework for federated learning models against collusion attacks. *Information Sciences*, 680, 121161.
- McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics* (pp. 1273-1282). PMLR.
- National People's Congress of the People's Republic of China. (2021). Data Security Law of the People's Republic of China. *Communique of the Standing Committee of the National People's Congress of the People's Republic of China*, 5(5), 951-956.
- Qi, T., Wu, F., Wu, C., He, L., Huang, Y., & Xie, X. 2023. Differentially private knowledge transfer for federated learning. *Nature Communications*, 14(1), 3785.
- Sharma, A., & Marchang, N. 2024. A review on client-server attacks and defenses in federated learning. *Computers & Security*, 103801.

- Subramanian, R. 2023. Have the cake and eat it too: Differential Privacy enables privacy and precise analytics. *Journal of Big Data*, 10(1), 117.
- Yang, H., Ge, M., Xue, D., Xiang, K., Li, H., & Lu, R. 2023. Gradient leakage attacks in federated learning: Research frontiers, taxonomy and future directions. *IEEE Network*.
- Yang, Q., Liu, Y., Chen, T., & Tong, Y. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2), 1-19.
- Zhang, B., Mao, Y., Tu, Z., He, X., Ping, P., & Wu, J. 2023. Optimizing Privacy-Accuracy Trade-off in DP-FL via Significant Gradient Perturbation. In *2023 19th International Conference on Mobility, Sensing and Networking (MSN)* (pp. 423-430). IEEE.
- Zhang, X., Kang, Y., Chen, K., Fan, L., & Yang, Q. (2023). Trading off privacy, utility, and efficiency in federated learning. *ACM Transactions on Intelligent Systems and Technology*, 14(6), 1-32.
- Zhu, L., & Chen, X. (2025). Privacy protection in federated learning: a study on the combined strategy of local and global differential privacy. *The Journal of Supercomputing*, 81(1), 1-29.

