Analysis on Security and Privacy-preserving in Federated Learning

Jipeng Li [1, *, †],Xinyi Li [2, †],Chenjing Zhang [3, †]

[1] School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110870, China;

[2.]Computer Science and Technology, Shandong University of Technology, Zibo 255022, China;

[3.]Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China.

* Corresponding Author Email: 1764300010@e.gzhu.edu.cn

[†]These authors contributed equally.

Abstract. Data privacy breaches during the training and implementation of the model are the main challenges that impede the development of artificial intelligence technologies today. Federated Learning has been an effective tool for the protection of privacy. Federated Learning is a distributive machine learning method that trains a non-destructive learning module based on a local training and passage of parameters from participants, with no required direct access to data source. Federated Learning still holds many pitfalls. This paper first introduces the types of federated learning, including horizontal federated learning, vertical federated learning and federated transfer learning, and then analyses the existing security risks of poisoning attacks, adversarial attacks and privacy leaks, with privacy leaks becoming a security risk that cannot be ignored at this stage. This paper also summarizes the corresponding defence measures, from three aspects: Poison attack defence, Privacy Leak Defence, and Defence against attack, respectively. This paper introduces the defence measures taken against some threats faced by federated learning, and finally gives some future research directions.

Keywords: Federated Learning, Poisoning Attack, Adversarial Attack, Privacy Leakage, Privacy Policy.

1. Introduction

Originally launched in 2016 by Google [1], Federated Learning (FL) aims to establish a shared pattern among the mobile devices and sensors to efficiently utilize these resources of data in the background of massive data and to ensure user confidentiality. Most of this scattered data, nevertheless, is highly heterogeneous and unbalanced. Jakub et al. [2] proposed a convenient and effective optimization algorithm for processing data distribution. Subsequently, some research has been already carried out to further optimize the federated learning model, such as the team of Mcmahan H B et al. proposing two ways to reduce communication consumption. They achieved a more efficient training process [3]. The team of Mohri M et al. solved the problem that shared models in the previous federated learning mechanism may be biased towards some participants, ensuring fairness among participants [4]. The team of Yurochkin M et al. proposed a single-sample/low-sample exploratory learning method to solve communication problems in compressive federated learning [5].

Once federal apprenticeship was initiated., it received widespread attention. Key technological and financial leaders have also begun to build open source projects, such as WeBank's FATE, Tensor Flow Federated (TFF) released by Google, and Uber's Horovod open source. Federated learning has been widespread in mobile communications and advanced computing [6], intelligent financing [7], intelligent health care [8], protection of the environment [9] and other areas, and it should change the business model of the new age and have an additional impact on building intelligent cities in the future.

However, there are still enormous impacts on safety in federated learning, such as the participants' low security level and the susceptibility to malicious threats, which affect the safety of the entire pattern. This paper analyses the security problems that may arise from federated learning, focuses on

the security threats of poisoning attacks, countermeasures and privacy leaks, and summarizes the defence measures in a targeted manner, in order to reduce the security risks of federated learning and to facilitate its future expansion and extension.

2. Style palette

2.1. Description of Federated Learning

FL is a distributed machine learning method. The stakeholders learn local database and then transfer the updating of parameters to the machine server, which in turn aggregates them to obtain the overall parameters. Federated learning improves learning efficiency compared to traditional machine learning techniques, solves the problem of data silos, and protects local data privacy [10].

2.2. Types of Federated Learning

For different datasets, federation learning is divided into three types, as presented in Figure 1.

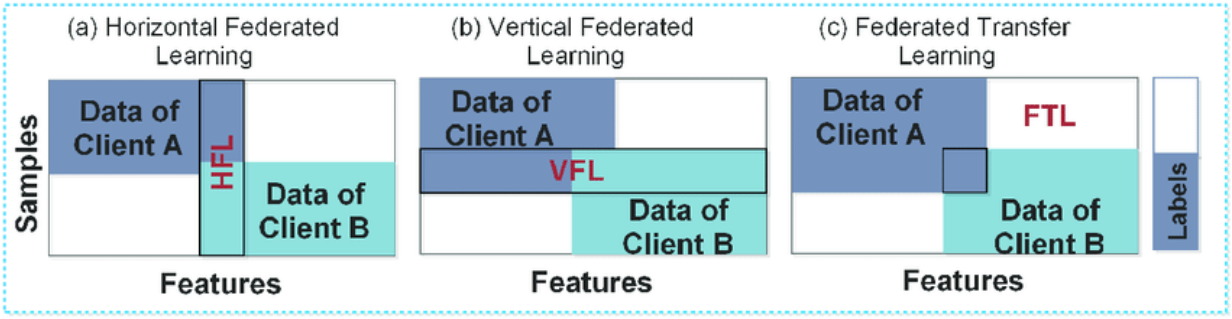

Figure 1 Classification of federated learning[11]

2.2.1. Longitudinal Federation Learning

Longitudinal federated learning refers to slicing the dataset according to the data characteristic aspects and taking out the portions that exercising with the same-user but different data features, where there is a large overlap of users but a small overlap of data features between different datasets.

2.2.2. Horizontal Federated Learning

Horizontal federation learning refers to slicing the dataset by user dimensions and taking out a portion of the data that has the same data features, but the users trained are different when there is more overlap in data features between different datasets and less overlap in users.

2.2.3. Federated Transfer Learning

FTL make mention of the use of transfer learning to train multiple datasets with a smaller amount of segistration between users and data lineaments, do not split the data. In place of splitting the data, FTL [12] is used to triumph over the dearth of data or labels.

3. Security issues in federated learning

Although FL is raised to better protect data privacy and its design and evolution are in keeping with the times, it can leak sensitive user information. Researchers have discovered a variety of privacy security problems in recent years, as seen in Figure 2. In FL, it is common for a malicious agent to exploit vulnerability [13] to control one or more actors to ultimately manipulate the global model. In this case, an attacker targets different clients who want to change and launch an attack in the global model by gaining access to local quiescent data, training processes, hyper-parameters, or updated metrics [14] in transit. The three most common of the attacks are poisoning attacks, adversarial attacks, and privacy leakage defence. In this section, we describe these three kinds of security threats in detail.

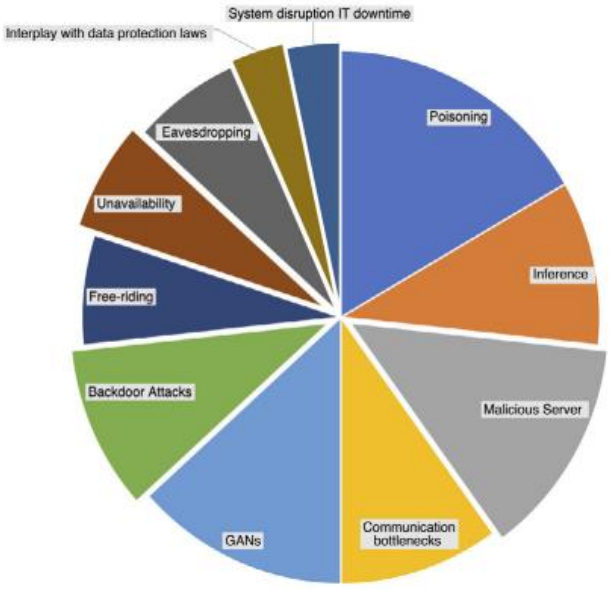

Figure 2 Severity of threats in FL. [15]

3.1. Poisoning Attacks

Poisoning attacks are more typical during the development and training of AI systems. The attacker disturbs the learning process such that the system accepts the attacker's chosen inputs and builds a backdoor through which he can control the output even in the future. In FL, the probability and intensity of poisoning attacks from client-side training set are both high. This section briefly describes two kinds of them, as shown in Figure 3.

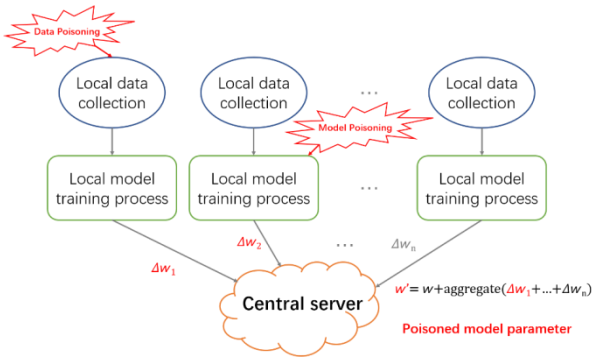

Figure 3 Data poisoning and model poisoning

3.1.1. Data Poisoning

In FL, data poisoning is when an assailant taints the samples in the training set by adding false labels or biased data, generating dirty samples, and falsifying model parameters. These parameters will bepostedto the server and ultimately impact the global pattern and destroy its usability or integrity. This is one of the most direct ways to corrupt a model.

G. Sun et al. [16] introduced a new systems-aware optimization approach (AT2FL) for rapidly deriving the hidden slopes of poisoning data and achieving first-rankambush methods in FL. Gorka Abad et al. [17] employed a combination of attacks to create a client-specific targeted backdoor with 100% ambushrate of success and 0% target label accuracy. H. Liu et al. [18] proposed a method for creating "poisonous-label" images, which increases the classification error dramatically.

3.1.2. Model Poisoning

Model poisoning is different from data poisoning in that the intruder does not directly operate the training data. They send wrong arguments to interrupt the learning process during aggregation [19], such as controlling some participated transmit to the tomcat. Therefore, it affects the changing

direction of the arguments of the entire learning pattern, slows down the convergence rate of the pattern, and destroys the accuracy of the whole pattern, which deeply affects pattern's performance. Bhagoji' research [20] only assumed a malicious agent (participant) to achieve a covert attack on the overall model, making the target model unable to correctly classify a certain type of data.

3.2. Privacy breaches

Participants can conduct data training locally using the federated learning method, and each participant is independent. There is no direct access to local data by any other individual entity, however, it guarantees a certain level of privacy. This security is not entirely secure and the risk of privacy breaches still exists. For example, a malicious party (curious or dishonest service providers) can infer sensitive information from other parties from the shared parameters.

Two forms of attack that can damage the private security of all parties are model extraction attacks and model reversal attacks [21]. The model exploitation attack attempts to undermine the secrecy of the model by stealing its arguments and hyper-arguments. For instance, a malicious actor can use the shared model to create a prediction query and then extract the trained model. The team of Tramer F et al. performed attacks on Amazon Machine Learning and BigML online services extracted a nearly identical pattern and demonstrated that the same attacks work with multiple machine learning methods [22]. Through a model, reverse attack, attempts to obtain statistics from the trained model of the training dataset to obtain the user's private information. The team of Ateniese G et al. implements an attack that infers the traffic types to be used in the model construction process [23]. Based on information inferred by the model from the trained set, the reverse attacks can be either members contained in the trained set or a number of synthetic features of the trained setting. Based on both training set information, inverted attacks can be split into limb inference attack and attacks by inference. This constitutes a grave threat to the privacy of the various participants in Federated Learning.

The server is trustworthy, however, in practice this is not the case. If the server is malicious, it identifies the source of the updated parameters, and further infer the participant's data set information through the parameters that the participant has repeatedly fed back, which may cause the participant's privacy leak. The team of Jiangcheng Qin et al. proposed a Privacy Recommendation System Framework (PPRSF) that uses federated learning to train and infer without concentrating user privacy data [24]. The risk of privacy leakage can be reduced, and the various recommendation algorithms can also be applied. Every member can build the module at the local level based on their own environment and summarise the local model parameters on the central server, and then return the resulting global model to each member using a single central server and multiple terminal devices. Figure 4 illustrates the procedure.

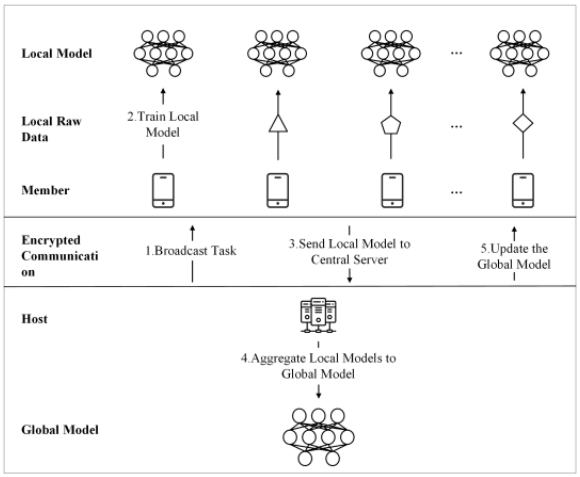

Figure 4 Centralized Cross-device Federated Learning Architecture [24]

3.3. Adversarial attacks

Attack algorithms that misclassify classifiers by adding small perturbations to the input are generally most common for networks used for deep learning, in scenarios such as the currently hot CV and NLP directions, for example, misclassifying classifiers by adding carefully prepared perturbation noise to images, or misclassifying sentiment by substituting synonyms for certain words in a sentence. There are several types of attacks, and they can be classed as black-box, white-box, or grey-box attacks depending on the attack context:

Black box attack: The attacker has no knowledge of the model's internal structure, training parameters, defence methods, or anything else, and can only interact with it through the output.

White-box attacks: In contrast to black-box models, the attacker has complete control over the model. White-box attacks make up the majority of current attack algorithms.

Grey-box attacks: Between black-box and white-box attacks, these are attacks in which only a portion of the model is known. (For example, just the model's output probabilities are given, or only the model's structure is known but not the parameters).

4. privacy leakage defence

In order to address the security issues mentioned in section 3, this section discusses the latest defence methods of each kind of attacks.

4.1. Poison attack defence

4.1.1. Data Poisoning Defence

Data protection should be the first line of defence against data poisoning. For one thing, the legitimacy and reliability of the origin of the data should be confirmed. On the other hand, before using data with unguaranteed security, corresponding tests should be carried out to ensure the data integrity.

Multiple defense methods have been proposed to resist data poisoning attacks. Xingyu Li et al. [25] proposed LoMar, a two-phase defence algorithm that improves target label accuracy testing under label rollover strike on the Amazon dataset from 96.0 percent to 98.8 percent and total average precision from 90.1 percent to 97.0 percent, in comparison to FG+Krum. Yuchen Tian et al. [26] proposed a defence against data poisoning assaults in FL situations that detects and suppresses potential outliers (DSPO), which outperformed existing defences in numerous cases. V. Tolpegin et al. [27] introduced a FL system aggregator that performs gradient clustering before the summary arguments are updated per round. effectively limiting their involvement in mobility training. Such clustering model gradients imply a robust defence since it is not necessary to obtain access to any public authentication dataset.

4.1.2. Model Poisoning Defence

Assuming the server is trusted, defenses focus on detection of incorrectly updated parameters. There are two usual methods of detecting anomalous update parameters [26]. One is by precision testing. The server uses the parameters U_i returned by the participant δ_i to calculate $w'_{G1} = w_G + f(\delta_i)$, and uses the parameters returned by other participants to calculate $w'_{G2} = w_G + f(\Delta)$, where $\Delta = \{\delta_j \mid j = 1,2,...,n, j \neq i\}$. w'_{G1} and w'_{G2} as the weight parameters of the model respectively, and the accuracies of the two models are compared on the validation set. If the precision of the pattern using w'_{G1} is significantly lower than that of the pattern using w'_{G2}. It is presumed that δ_i is abnormal. The other approach is to simply compare the digital statistical variance between the update parameters δ_1, δ_2, ... δ_n presented by each attendee. When there is a large statistical difference between the

update parameters, δ_i is fed back by a participant and those of other participants. It is inferred that δ_i is abnormal.

4.2. Privacy Leak Defence

The protection of privacy in federated learning is primarily ensured from the point of view of two large entities: the participant and the server. At the same time, the trained model should also prevent model extraction attacks and model reverse attacks.

4.2.1. Differential privacy

Think about malevolent parts in compared to honest servers. As any participant may derive overall parameters of the training process, the federated learning approach is susceptible to incremental attacks [28]. When analysing the shared model, the confidentiality of data from other honest parts is at stake. In that case, different privacy protection techniques are frequently employed.

The team of Geyer R C et al. proposes a federated optimization algorithm for participant differential privacy protection, the differential privacy stochastic gradient descent algorithm, which aims to hide participants' updated settings during the model training phase, and balance the loss of privacy with model performance [28]. This technology randomly splits the data sample into small portions, and adds Gaussian noise to the aggregation process to obtain a different privacy protection, while maintaining the high performance of the model.

4.2.2. Secret Sharing Mechanisms

Consider the situation of honest parties versus malicious servers. Servers play an important role in federated learning, where it can take parameters for feedback from individual clearly identified participants and infer sensitive information about participants, which threatens participants' privacy and can be prevented using a secret sharing mechanism.

The team of Bonawitz K et al. based on Shamir's secret sharing, a practical security aggregation scheme is designed that guarantees the security of update parameters in an authentic and inquisitive server context, that is, to assure data privacy for each individual participant whilst keeping protocol complexities in check in order to keep computational and communications expenses low in large datasets, for cooperative learning in federated training [29]. However, this agreement could not prevent complicity in the attack.

4.2.3. Homomorphic Encryption

Consider the situation of honest parties versus malicious servers. The use of encrypted data transmission to ensure privacy is an effective defence. Homomorphic encryption is a commonly used means of defence.

A new deep learning system is proposed based on honest and curious cloud servers, using isomorphic cryptography schemes to effect aggregations of the gradient on both honest and inquisitive servers, and to assure that the resulting system achieves the same level of accuracy as the in-depth learning system formed in the common data set from all parties [30]. CryptoDL was developed to train convolutional neural networks with approximate polynomials instead of the original activation function, which has been shown to be accurate at 99.52% in the MNIST dataset and can make nearly 164 000 predictions per hour, providing an efficient and accurate privacy protection scheme [31].

4.3. Defence against attack

A large number of adversarial attack defence mechanisms have been proposed in ML, which are also applicable to the adversarial defence of FL. This section discusses four effective methods.

4.3.1. Defensive distillation

When confronted with adversarial example attacks, modifying the network topology and retraining a more sophisticated model can sometimes protect against them, albeit at a considerable cost. More

computational resources and time are required to retrain a more complicated network. As a result, without modifying the network structure, the ideal option is to discover a new defence method that has the least impact on the model's accuracy. At this time a defensive method called defensive distillation [32] was proposed. The idea of this defence method is very simple. First, using the original training sample X and the label Y, a deep neural network is trained with an accident to produce the probability distribution F (X). The output result of the first stage, F (X), is used to train a retort system with the same structure and retort temperature, and a new distribution of probabilities, with sample X serving as a new label. Finally, the network as a whole is employed for classification and prediction, allowing it to effectively defend against adversarial examples.

4.3.2. Adversarial training

Another common defence is to develop the final model using adversarial training, which entails merging actual and antagonistic specimens as the train set. This method may be used for a variety of supervised tasks [33] and allows the module to improve the properties of the model by learning the properties of antagonistic specimens during the process of training. However, such a model is only capable of withstanding adversarial samples in the training set and is unprepared to fight against unexpected attacks.

Adversarial training is a special implementation of data augmentation. During training, since it is unrealistic to exhaust all adversarial examples, randomizing all of the original data can improve the model's generalisation ability. JianFei Zhang et al. [34] proposed a data augmentation method called FedDA, which is proved to efficiently utilize nonoverlap samples to enhance the effect of the data augmentation in a series of experiments.

4.3.3. Data processing

Unlike data augmentation, data processing denoises samples to reduce the interference of adversarial samples. The team of Zantedeschi Vintroduces et al. introduces two classical methods, which is scalar quantization and smooth spatial filtering, to reduce the influence of noise [35]. Using the image entropy as a metric, the adaptation of noise mitigation for diverse pictures is realized. Comparedwith the classification results of a given sample, this denoising method can effectively detect and reject adversarial samples.

4.3.4. GAN-based defence

GAN is a variety of generative adversarial network pattern. The formal statement is to obtain training samples and train a model, which can be generated according to the target data distribution we define data. This is the Nash equilibrium of game theory, which is the main idea of GAN. The two portions in this theory are set as a generator and a discriminator, respectively. There are many generative methods' drawbacks can be solved by the GAN, mainly including parallel generation of samples; few restrictions on generative functions, such as no need for suitable Markov sampling data distribution (Boltzmann machines), generative functions do not need to be invertible; no need to use Markov chain methods (Boltzmann machines, GSNs); compared to VAE methods, no variational bound is required; GANs generally perform better than other methods.

5. Future Directions In Fl Security And Privacy

As artificial intelligence technology expands and extends, people feel the convenience brought by technology,and gradually increase the demand for privacy protection.

However, at present, many safety issues remain in federal learning, and this article mainly focuses on the three types of security issues of poisoning attack, anti-attack and privacy leakage, and summarizes the targeted security and privacy protection defense measures. However, this is not a simple task, and the existing defense methods can only improve the robustness of the model under certain conditions and within a certain range. Among the security issues of federated learning, there are still some issues that remain to be resolved as presented as below.

5.1. Data quality issues

Since the dataset is stored locally, the server cannot access the data source. It is difficult to ensure that the labeling of the data is correct. Moreover, the degree of heterogeneity of the data between the participants is not known. If the data scale is not large enough, it is easy to cause frequent confrontation attacks due to the rare samples. The difficulty of confrontation and defense is also increased. Zero-knowledge proof and commitment protocols can be considered to ensure data quality by enabling verifiability of malicious user data.

5.2. Communication efficiency issues

Current federated learning is mostly synchronous. The servers can interact with data from a multitude of participants in one iteration. If the user wants to use a variety of defenses to ensure the security of models and sensitive information, it is bound to increase the communication burden on the server, and even cause denial of service attacks or single point failures. If the user considers multiple servers, the security of the interaction between servers is also a topic worth exploring in depth. Therefore, the methods have been proposed to achieve efficient privacy protection and to reduce the number of times used when it is necessary to use public key passwords to protect user privacy [36-37].

5.3. Model accountable issues

Federated learning methods further complicate models, and lack of interpretability can lead to potential threats in federated learning applications. Interpretability is the ability to point out human interpretation or to present understandable terms [38]. Improving the interpretability and transparency of federated learning models helps eliminate inherent security risks and further improves the reliability and security of models. Because of the inherent nature of federated learning, a focus may be placed on interpretability methods in the future.

6. Conclusion

FL is a relatively new market framework that requires further research to determine the enhanced overflow that suits the styles of different FL environments. Federated learning is a very promising field of research, which has attracted many scholars to conduct research in the related fields and has also achieved a series of important research results. This paper mainly describes the security issues and defense measures in federated learning. This paper discusses that security threats refer to vulnerabilities that may be exploited by malicious attackers to affect system security and violate their privacy policies. In this paper, the security issues include poison attacks, counter-attacks, privacy leaks and the attack model. The advantages at this stage are summarized. For defensive measures, this paper mainly describes the latest solutions to defend against the above attacks. However, the development of federal learning technology is still in its infancy, and there are still many problems to be solved. In the future work, it will be necessary to continue to study security issues in the field of federal learning, in order to accelerate the research and development of the related security and privacy protection technologies, and to promote the further development of federal learning.

References

[1] H. B. Mcmahan, E. Moore, D. Ramage, et al.Federated learning of deep networks using model averaging[J]. arXiv preprint, arXiv: 1602.05629,2016.

[2] J. Konečný, H. B. Mcmahan, D. Ramage,et al.Federated optimization:distributed machine learning for ondevice intelligence[J]. arXiv preprint, arXiv: 1610.02527, 2016.

[3] J. Konečný, H. B. Mcmahan, YU F. X.,et al.Federated learning: Strategies for improving communication efficiency[J]. arXiv preprint, arXiv: 1610.05492,2016.

[4] G. Sivek, M. Mohri, A. T. Suresh.Agnostic federated learning[J]. arXiv preprint, arXiv: 1902.00146,2019.

[5] M. Yurochkin, M. Agarwal, S. Ghosh, et al. Bayesian nonparametric federated learning of neural networks[J]. arXiv preprint, arXiv: 1905.12022, 2019.

[6] S. Niknam, H. S. Dhillon, J. H. Reed. Federated learning for wireless communications:Motivation, opportunities and challenges[J]. arXiv preprint, arXiv: 1908.06847, 2019.

[7] G. A. Rena, M. J. Sheller, B. Edwards, et al. Multi-institutional deep learning modeling without sharingpatient data:A feasibility study on brain tumor segmentation[C]// International MICCAI Brainlesion Workshop. Springer, Cham.

[8] Y. Chen, X. Qin, J. Wang, C. Yu and W. Gao, "FedHealth: A Federated Transfer Learning Framework for Wearable Healthcare" in IEEE Intelligent Systems, vol. 35, no. 04, pp. 83-93, 2020.doi: 10.1109/MIS.2020.2988604

[9] B. Hu, Design and implementation of air quality monitoring system based on federated learning[D]. Beijing: Beijing University of Posts and Telecommunications, 2019.

[10] Custers, B., Sears, A. M., Dechesne, F., Georgieva, I., Tani, T., & Van der Hof, S. (2019). EU personal data protection in policy and practice. TMC Asser Press.

[11] Asad, Muhammad & Moustafa, Ahmed & Yu, Chao. (2020). A Critical Evaluation of Privacy and Security Threats in Federated Learning. Sensors. 20. 7182. 10.3390/s20247182.

[12] Pan S. J., Yang Q.. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 22(10): 1345 − 1359.

[13] X. Zhou, J. Men, G. Xu, Z. Han, Z. Sun, W. Lian, X. Cheng, Finding sands in the eyes: vulnerabilities discovery in IoT with EUFuzzer on human machine interface, IEEE Access 7 (2019) 103751–103759.

[14] Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2020, June). How to backdoor federated learning. In International Conference on Artificial Intelligence and Statistics (pp. 2938-2948). PMLR.

[15] Mothukuri, V., Parizi, R. M., Pouriyeh, S., Huang, Y., Dehghantanha, A., & Srivastava, G. (2021). A survey on security and privacy of federated learning. Future Generation Computer Systems, 115, 619-640.

[16] G. Sun, Y. Cong, J. Dong, Q. Wang, L. Lyu and J. Liu, "Data Poisoning Attacks on Federated Machine Learning," in IEEE Internet of Things Journal, doi: 10.1109/JIOT.2021.3128646.

[17] Abad, G., Paguada, S., Picek, S., Ramírez-Durán, V. J., & Urbieta, A. (2022). Client-Wise Targeted Backdoor in Federated Learning. arXiv preprint arXiv:2203.08689.

[18] H., Liu, D., Li & Y., Li Poisonous Label Attack: Black-Box Data Poisoning Attack with Enhanced Conditional DCGAN. Neural Process Lett 53, 4117–4142 (2021).

[19] Lim, W. Y. B., Luong, N. C., Hoang, D. T., Jiao, Y., Liang, Y. C., Yang, Q., ... & Miao, C. (2020). Federated learning in mobile edge networks: A comprehensive survey. IEEE Communications Surveys & Tutorials, 22(3), 2031-2063.

[20] Bhagoji, A. N., Chakraborty, S., Mittal, P., & Calo, S. (2019, May). Analyzing federated learning through an adversarial lens. In International Conference on Machine Learning (pp. 634-643). PMLR.

[21] He, Y. , Hu, X. , He, J. , Meng, G. , & Chen, K. . (2019). Privacy and security issues in machine learning systems: a survey. Journal of Computer Research and Development.

[22] Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing Machine Learning Models via Prediction {APIs}. In 25th USENIX security symposium (USENIX Security 16) (pp. 601-618).

[23] L. V. Mancini, G. Ateniese, G. Felici, et al.Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers[J]. arX[1]iv preprint, arXiv: 1306.4447, 2013.

[24] Qin, J., Liu, B., & Qian, J. (2021, January). A Novel Privacy-Preserved Recommender System Framework based on Federated Learning. In 2021 The 4th International Conference on Software Engineering and Information Management (pp. 82-88).

[25] Li, X., Qu, Z., Zhao, S., Tang, B., Lu, Z., & Liu, Y. (2021). LoMar: A Local Defense Against Poisoning Attack on Federated Learning. IEEE Transactions on Dependable and Secure Computing.

[26] Tian, Y., Zhang, W., Simpson, A., Liu, Y., & Jiang, Z. L. (2021). Defending Against Data Poisoning Attacks: From Distributed Learning to Federated Learning. The Computer Journal.

[27] Tolpegin, V., Truex, S., Gursoy, M. E., & Liu, L. (2020, September). Data poisoning attacks against federated learning systems. In European Symposium on Research in Computer Security (pp. 480-501). Springer, Cham.

[28] Geyer, R. C., Klein, T., & Nabi, M. (2017). Differentially private federated learning: A client level perspective. arXiv preprint arXiv:1712.07557.

[29] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., ... & Seth, K. (2017, October). Practical secure aggregation for privacy-preserving machine learning. In proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 1175-1191).

[30] Aono, Y., Hayashi, T., Wang, L., & Moriai, S. (2017). Privacy-preserving deep learning via additively homomorphic encryption. IEEE Transactions on Information Forensics and Security, 13(5), 1333-1345.

[31] E. Hesamifard, H. Takabi, M. Ghasemi. Cryptodl: Deep neural networks over encrypted data[J]. arXiv preprint, arXiv: 1711.05189, 2017

[32] N. Papernot, P. McDaniel, X. Wu, S. Jha and A. Swami, "Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks," 2016 IEEE Symposium on Security and Privacy (SP), 2016, pp. 582-597, doi: 10.1109/SP.2016.41.

[33] T. Miyato, S. Aeda, M. Koyama, et al. Virtual adversarial training: a regularization method for supervised and semi-supervised learning[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(8): 1979 − 1993.

[34] Zhang, J., & Jiang, Y. (2022). A Data Augmentation Method for Vertical Federated Learning. Wireless Communications and Mobile Computing, 2022.

[35] V. Zantedeschi, M. I. Nicolae, A. Rawat. Efficient defenses against adversarial attacks[C]//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security.Dallas,Texas, USA:ACM, 2017: 39 - 49.

[36] J. Zhou,X. Dong,Z. Cao,Research Progress on Privacy Protection of Recommendation System[J]. Journal of Computer Research and Development, 2019, 56(10): 2033 − 2048.

[37] J. Zhou,X. Dong,Z. Cao, et al. Research progress on big data security and privacy protection[J]. Journal of Computer Research and Development, 2016, 53(10): 2137 −2151.

[38] S. Ji, J. Li, T. Du , et al. A Review of Interpretable Methods, Applications and Safety Research on Machine Learning Models[J]. Journal of Computer Research and Development, 2019, 56(10): 2071 – 2096