GBMIA: Gradient-based Membership Inference Attack in Federated Learning

Xiaodong Wang, Naiyu Wang, Longfei Wu, Zhitao Guan

June 7, 2024

Abstract

Membership inference attack (MIA) has been proved to pose a serious threat to federated learning (FL). However, most of the existing membership inference attacks against FL rely on the specific attack models built from the target model behaviors, which make the attacks costly and complicated. In addition, directly adopting the inference attacks that are originally designed for machine learning models into the federated scenarios can lead to poor performance. We propose GBMIA, an attack model-free membership inference method based on gradient. We take full advantage of the federated learning process by observing the target model's behaviors after gradient ascent tuning. And we combine prediction correctness and the gradient norm-based metric for membership inference. The proposed GBMIA can be conducted by both global and local attackers. We conduct experimental evaluations on three real-world datasets to demonstrate that GBMIA can achieve a high attack accuracy. We further apply the arbitration mechanism to increase the effectiveness of GBMIA which can lead to an attack accuracy close to 1 on all three datasets. We also conduct experiments to substantiate that clients going offline and the overlap of clients' training sets have great effect on the membership leakage in FL. **Index Terms**——Membership Inference Attack, Federated Learning, Membership Privacy, Privacy Leakage …

1 Introduce

In recent years, federated learning (FL) has received extensive attention from academia and industry. The secure [1] and efficient [2] communication technologies also support the further development of FL. In FL, participants (clients), who are reluctant to share the local data, are entitled to collaboratively train a deep learning model [3]. One of the commonly used federated learning paradigm is that under the coordination of a centralized parameter aggregator (server), each client uses local data to train a local model based on a global model. Clients contribute to the global model by uploading the updated parameters of the local models in every federated training iteration, and the server needs to aggregate

local updates to construct the new global model. As the result, clients can benefit from the final global model without revealing their local data [4], [5].

FL can avoid the leakage of user's privacy caused by the direct sharing of training data, but deep learning models may still disclose private information from the uploaded local parameters. Therefore, the deep learning model may leak the information of the training set during the model training and model inference stages. An attacker can obtain information about the training data by observing and analyzing the behavior of the target model. Shokri et al. [6] first proposed a membership inference attack (MIA) method against machine learning models, and since then the MIA has attracted a considerable attention from scholars. For a given target data, the attacker can verify whether the target data point exists in the training set of the target model through MIA [7]. We concentrate on the MIA against FL. The attack can be conducted by both the server (global attacker) and clients (local attackers).

The existing MIAs are mostly based on the respective attack models constructed from the target model behaviors [8], that is, the attacker trains an attack model which takes the target model's behaviors (e.g. prediction confidence, prediction label, hidden layer output, etc.) as inputs to make membership inference. Such attack model-based MIAs are costly and complex, as they require attacker to design, train, and optimize the attack model. For example, the lack of training data for the attack model is a common problem that the attack model-based MIAs face [3]. Zhang et al. [9] utilized generative adversarial networks to produce and enrich the attack model's training data, but it does not reduce the overhead and complexity of the attack model-based MIAs. Some existing works have proposed MIAs against independently trained models, which do not require the training of attack models [10], [11]. However, directly adopting these attack model-free MIAs against machine learning models into the federated scenarios leads to poor performance.

In this paper, we focus on the horizontal federated learning scenario and propose a gradient-based membership inference attack (GBMIA), which can be conducted by both global and local attackers. GBMIA is attack model-free so it can get rid of the trouble of constructing attack model and its training data. Our attack approach makes the full use of the federated learning process by observing how the target model behaves after gradient ascent tuning. We conduct experiments on three real-world datasets to demonstrate the effectiveness of GBMIA. Our experimental results reveal that GBMIA has an outstanding performance compared to other approaches. Our contributions can be summarized as follows:

- We present GBMIA, an attack model-free MIA approach, which poses a critical threat to FL. GBMIA performs membership inference by observing the predicted labels and gradient norm of the target model on the target data during the federated learning process.
- We conduct experiments on different datasets, and compare our approach with existing attack approaches. The experimental results show that GBMIA has better attack performance. By applying arbitration mechanism, GBMIA achieves the highest attack accuracy which is close to 1.

• We further investigate the factors that may influence the MIAs against FL. We simulate the scenario where clients go offline and there is an overlap between clients' training sets. Experiments show that when the target participant is offline, the accuracy of member inference attack decreases, and the existence of target data in the training sets of multiple clients will aggravate the membership privacy leakage.

The rest of this paper is organized as follows. In section , we perform a review of the related works. In section , we introduce the proposed attack method in detail. Section IV presents the experimental results. Section V concludes this paper.

2 Related Work

Membership inference can be used to attack different models and tasks [12]. Meanwhile, the data security issue has been widely concerned [13], [14]. However, when the training sets of the target model contains sensitive data such as genomic data [15], geographic location data [16], and medical data [17], MIAs will cause serious privacy leakage.

Shokri et al. [3] first proposed a membership inference attack method, which simulates the behavior of target models by building shadow models of target models and further uses attack models to analyze target models. As a follow-up study, Salem et al. [11] gradually realized the independence of attacker's knowledge about the target model and its training set through three adversaries. The above two methods are based on the prediction confidence vectors. Choquette-Choo et al. [18] proposed a label-only MIA that does not rely on confidence scores. Overfitting is generally considered to be an important factor affecting the degree of membership privacy leakage [3], [10], [11]. Leino et al. [19] provided a profound insight on the overfitting in deep neural networks and proposed an improved membership inference method leveraging the white-box information.

Different from the independently trained deep learning models, federated learning models may suffer from the MIA launched by the server or a clients. Nasr et al. [12] studied the leakage of membership privacy in independently trained models and federated learning models in white-box scenarios, and proposed the concepts of active and passive attacks. Zhang et al. [9] proposed a passive local MIA, which only makes use of the output of the target model. The attack process is divided into two stages. One is to expand the training set of attack model through data augmentation. The second is to train the attack model. Hu et al. [20] proposed a follow-up attack of MIA in the FL scenarios, called source inference attack. The purpose of this attack is to further determine which client's training set the target data comes from. But the source inference attacks can only be carried out by the server.

3 Gredient-based membership interference attack

In this section, we first introduce the threat model, then we present an overview of the attack framework, and finally we unfold the details of GBMIA in two separate phases, gradients ascent tunning and membership inference.

3.1 Threat model

Membership inference attack can be described as a binary classification task:

$$AX_{Target}, f_M, P \rightarrow 0, 1 \tag{1}$$

where X_{Target} denotes the target data of membership inference attack, f_M denotes the target model and P is the extra knowledge that attacker achieves. If A outputs 1, it means that the target data X_{Target} is member data which is in the training set of one or multiple clients. If A outputs 0, it means that the target data X_{Target} is non-member data which is not in the training set of any clients.

Similar to the existing membership inference attack against FL [12], we assume that the attacker can be either the server or a client. We assume that the local training set of each client has the same data distribution. The attacker can actively affect the training process of FL, like the server maliciously modifies the global model parameters before sending to clients, or a client maliciously modifies the local parameters before uploading to the server.

3.2 The attack framework

We present an overview of the attack architecture of the membership inference against FL. The overall framework of our proposed membership inference attack approach is shown in Fig.1. The serial numbers in the figure represent the specific steps in the single-round federated learning process. The attack framework includes a central server and multiple clients. The attacker can be the central server (global attacker) or a client (local attacker). The global attacker takes the global model updated with the aggregation results as the target model, while the local attacker takes the global model distributed by the server in each iteration as the target model. A malicious server can obtain more information about the private training sets of the clients by modifying the global parameters sent to each client. This is an active attack launched by the global attacker. A malicious client can obtain more information about the local training sets of other clients by modifying the local parameter updates uploaded to the server. This is an active attack launched by the local attacker.

3.3 Gradient ascent tunning

Each client in FL uses the Stochastic Gradient Descent (SGD) algorithm to train a local model. The SGD algorithm makes the model update the parameters in

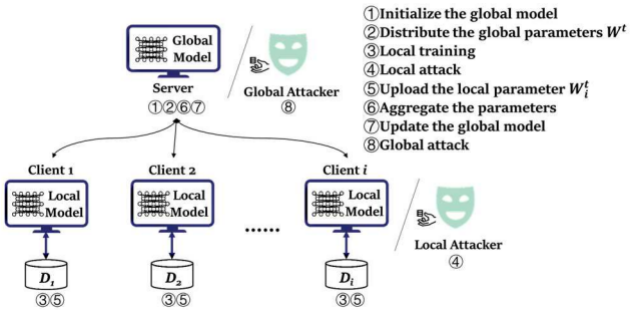


Figure 1: The attack architecture of the membership inference against FL.

the direction of gradient descent during the training process so as to decrease the gradient of the loss function on the training data. Nevertheless, gradient ascent is to update the parameters in the opposite direction of gradient descent, so as to increase the loss on the target data points, which is able to change the model prediction. If a certain client has the target data x Target in its training set, that client is denoted as a target client; otherwise, the client is denoted as a non-target client. The purpose of gradient ascent is to induce the target client to decrease the loss of its local model on the target data points through SGD, as the loss is maliciously increased by the attacker using gradient ascent. This can be detected by the attacker, and be used to infer membership. Gradient ascent is firstly used in MIA by Nasr et al. [12]. However, their approach needs to construct the training set to build an attack model which learns to distinguish member data from non-member data, which is costly and complicated.

Base on the approach of Nasr et al. [12], we propose an improved gradient ascent tunning method. Rather than directly adding the gradient to the model parameters, we firstly input the target data point X_{Target} into the target model, then observe the corresponding output. If the output matches the label of the target data point, we compute the gradient of the loss L on X_{Target} and add it to the parameters, leading the model to produce a wrong prediction on X_{Target} . We update the model parameters using gradient ascent tunning as follows:

$$w \leftarrow w + \gamma \frac{\partial L_{X_{Target}}}{\partial w} \tag{2}$$

where w denotes the model parameters and γ denotes the ascent rate. If the output of the target model does not match the label of X_{Target}, we do not modify the model parameters. If X_{Target} is the member data, the gradient of the loss on X_{Target} will be reduced due to target client's local SGD, which makes the target model prediction correct on X_{Target} . This is exactly what an attacker can use to make membership inference.

3.4 Membership inference

The attacker infers membership of the target data on the target model after gradient ascent tunning. The membership inference contains three steps: (1) querying the target model, (2) label matching, and (3) threshold comparison. Firstly, the attacker queries the target data X_{Target} to the target model, and obtains the prediction label y' of the target model on the target data. Then, the attacker compares the prediction label y' with X_{Target} 's ground-truth label y. If y' matches y, the attacker determines that X_{Target} is in the local training sets of one or multiple clients. If not, it goes to step three, where the attacker computes the gradient norm and compares it with a threshold τ . If the gradient norm is above τ , the attacker takes X_{Target} as the member data and vice versa. The necessity of step two is that label matching can decrease the complexity and calculation overhead of the attack. If y' matches y, it does not need to go to step three. The inspiration of gradient norm comparison is that the Euclidean norm of the gradients of the loss is greatly associated with membership leakage, as the gradient norms of the member data is apparently higher than those of non-member data [12].

Choosing the threshold τ . We provide a method for the attacker to choose the gradient norm threshold. First, the attacker generates samples from the distribution which is consistent with the distribution of clients' training data. This is not difficult for the attacker in the FL scenarios, as all parties participating in FL have a high probability of coming from the same area and have similar business. These samples can be hypothetically regarded as non-member data. Next, the attacker queries these samples to the target model and computes the gradient norms. Afterwards, the attacker takes the bottom-k percentile of these gradient norms as the threshold τ . In some situations, the server also have a part of testing set to estimate the performance of the global model, so both global and local attackers have access to testing set whose distribution is consistent with training set. Therefore, as for a lazy attacker, he can simply use some data points in his test set to compute the threshold, which makes it more convenient to perform the attack.

Global attack. The malicious server (global attacker) observes the clients' uploaded local model updates to make membership inference. Algorithm 1 shows the pseudocode of the global attack of GBMIA. In iteration t, the global attacker aggregates the local updates received from each client, and then chooses whether to conduct the attack according to the attack tag array Aattack. If the answer is yes, the global attacker updates the global model by gradient ascent tunning. Then, in iteration t+1, the global attacker receives the local model updates, and aggregates them and update the global model. If the attack tag of the previous iteration is yes, the attacker makes the membership inference. The global attacker takes the updated global model as the target model, queries the target data x Target to the target model, and compares the prediction label y' to x Target 's ground-truth label y. If y' does not match y, the attacker computes the threshold τ and the gradient norm, and compares them to make the inference.

Algorithm 1 Global attack of GBMIA

Input: Target data X_{Target} , the ground-truth label of target data y, current iteration t, the previous iteration's global model parameters W_g^{t-1} , the local model parameters W_t . **Output:** m

ClientsTrain
ServerAggregate
if $A_{attack}\,[t-1]$ is TRUE **then**
    $m \leftarrow$ **MembershipInference**()
end if
ClientsTrain
    Receive W_g^{t-1} and train the local model
    Upload W_l to the server
function SERVERAGGREGATE(A)ggregate local model updates using FedAvg
    if $A_{attack}\,[t]$ is TRUE **then**
        GradientAscent
    end if
end function

Local attack. The local attacker acts as a client participating in federated training with a malicious purpose of inferring private information about other clients' training set. Algorithm 2 shows the pseudocode of the local attack of GBMIA. In iteration t, the attacker receives the global model parameters from the server, and chooses whether to conduct the attack according to the attack tag array Aattack. If the answer is yes, the attacker updates the local model by gradient ascent tunning and uploads the parameter updates to the server for aggregation. Then, in iteration t+1, the attacker receives the global model. If the attack tag of the previous iteration is yes, the attacker makes the membership inference. Similarly, the attacker queries target data x Target to the target model and makes membership inference via label matching or threshold comparison.

4 Experiments

In this section, we evaluate the performance of our proposed membership inference attack GBMIA. We first introduce the datasets and experimental settings. Then we present the performance of GBMIA and compare it with other attack methods. Finally, we provide an analysis of how GBMIA performs under different scenarios, including clients going offline and clients' training sets having overlaps.

Algorithm 2 Local attack of GBMIA

Input: Target data X_{Target} , the ground-truth label of target data y, current iteration t, the previous iteration's global model parameters W_g^{t-1} , the local model parameters W_t . **Output:** m

ClientsTrain
if $A_{attack}\,[t-1]$ is TRUE **then**
    m = MembershipInference
end if
ServerAggregate
function ClientsTrain(R)eceive W_g^{t-1} and train the local model
    if $A_{attack}\,[t]$ is TRUE **then**
        GradientAscent
    end if
    Upload W_t to the server
end function
function ServerAggregate(A)ggregate local model updates using FedAvg
end function
function GradientAscent(C)alculate the gradient ascent on X_{Target}
    Update W_g^t using gradient ascent tuning
end function

4.1 Datasets and experimental setups

We implement the experiments on a PC equipped with Windows11 OS 16GB RAM Intel(R) Core(TM) i7-12700F CPU @ 2.10GHz NVIDIA GeForce RTX 3060 12GB GPU Python(3.8.13) torch(1.8.0). We ran the experiments on three real-world datasets, CIFAR10, Purchase100 and Texas100 [3]. CIFAR10 is an image dataset widely used in the field of image recognition. There are a total of 60,000 color images from 10 classes. Purchase100 consists of 197,324 purchase records of consumers, and each record contains 600 binary attributes. Each attribute represents a product and the value indicates whether the consumer has purchased the product. Texas100 contains 67,330 patients records from multiple medical institutions in Texas. These records are divided into 100 classes which represent different types of patients. Table presents how these three datasets are used for our experiments. There is no overlap between clients' local training sets. As for membership inference, the attacker holds part of clients' training data as test member data. The framework of FL in our experiments consists of a server and five clients. The five clients jointly train a global model under the coordination of the server, and each client has a local model. To decrease the impact of different types of target models, we use fully connected networks for the three datasets. The details of the model trainings are given in Table .

Table 1: TABLE I: SIZE OF DATASETS

Datasets	**Training**	**Testing**	**Non-Member**	**Attackers test**
CIFAR-10	6,000	6,000	1,000	1,000
Purchase100	19,700	19,700	1,000	1,000
Texas100	6,700	6,700	1,000	1,000

Table 2: TABLE II: HYPER-PARAMETERS OF MODELS

Datasets	**Optim.**	**Iter.**	**Batch size**	**Size of hidden layers**	**Notes**
CIFAR-10	Adam	-	32	1024, 512, 128	-
Purchase100	Adam	-	600	1024, 512, 256, 128	(lr=1e-3)
Texas100	Adam	-	500	1024, 512, 128	-

4.2 Performance of GBMIA

We use the membership prediction accuracy and precision metrics to assess GBMIA. We test the global and local membership inference attack respectively. The member and non-member data points in the attacker's test datasets are selected from clients' local datasets. We assume the attacker is a lazy attacker who uses his own testing set to choose the threshold τ with regard to the bottom-k percentile. We set the k as 10, and set the ascent rate γ as 0.5. To avoid of being detected, γ can be set to a smaller value. We compare the proposed GBMIA in Section (active attack) with the simplified variant of GBMIA without gradient ascent tunning (passive attack) to demonstrate the impact of gradient ascent tunning. The difference between the active and passive attack method is that the passive attack only contains membership inference phase, that is, the attacker does not maliciously modify the target model parameters. We launch the attack in ten iterations randomly selected from the last 100 iterations and record the best results. The experimental results are presented in Table . As

Table 3: TABLE III. PERFORMANCE OF GBMIA

Dataset	Passive Model			Active Model		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
CIFAR-10	0.92	0.91	0.94	0.85	0.83	0.87
Purchase100	0.93	0.92	0.94	0.86	0.84	0.88
Texas100	0.94	0.93	0.95	0.87	0.85	0.89

shown in Table , our MIA method GBMIA has over 90% accuracy for global attack and over 86% accuracy for local attack on CIFAR-10, Purchase100, and Texas100, which demonstrates that our attack method poses a great threat to FL. The active GBMIA achieves much higher accuracy than the passive attack which proves the advantage of the improved gradient ascent tunning. The attack accuracy of global GBMIA is higher than that of local GBMIA, which

shows that the global model leaks more membership privacy. We also compare GBMIA with three existing membership inference attack methods including ML-Leaks [11], loss-based attack [10], [21], and attack model-based method [12]. ML-Leaks (Adversary 3) and loss-based attack are both threshold-based attack. ML-Leaks is based on the prediction confidence vector, while loss-based attack is based on the model loss. ML-Leaks queries non-member samples to the target model to get the maximum of the confidence vector. The attacker takes the top t percentile of these maximums as the threshold. If the maximum of target data's confidence vector is higher than the threshold, the attacker assumes the target data as a member data and vice versa. As for the loss-based attack, the attacker queries non-member data points to the target model to get the loss and takes the minimum of these losses as the threshold. If the lo ss on the target data is smaller than threshold, the attacker assumes the target data as a member data. The attack model-based method takes use of the target model's layer outputs, loss and gradients. The attacker trains an attack model with multiple components to make membership inference. We can see from Fig. 2 that GBMIA outperforms all three other membership inference methods on all datasets. To further increase the attack accuracy of GBMIA, we introduce

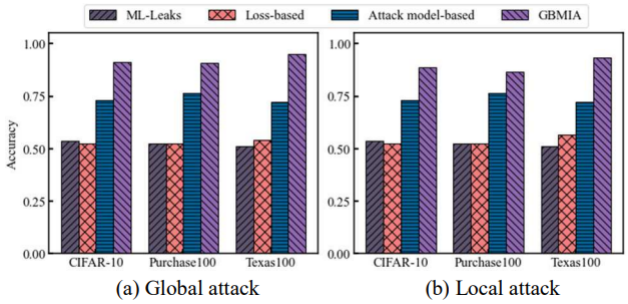
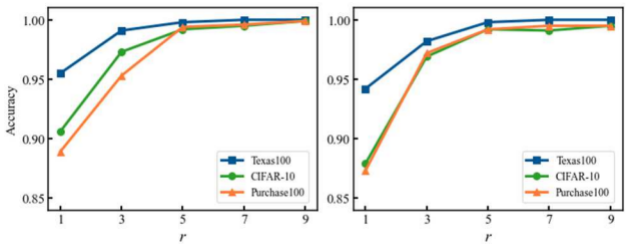


Figure 2: Comparison of attack accuracy.

the arbitration mechanism that allows the attacker to repeat the GBMIA in r different iterations and make a final decision on the membership according to the r results of the same target data. We evaluate the relation between r and the attacker's performance. As we can see in Fig. 3, setting r to 5 is a great choice, as it can achieve a high accuracy that is close to 1 for both global attack and local attack, and causes less overhead in terms of the number of repetitions.

4.3 GBMIA performance when clients go offline

In the real-world scenario, due to network delay or human factors, the client may go offline unintentionally or maliciously during the iterations of federated learning. We explore the impact of the client going offline on the membership leakage through simulation experiments. We assume that the client go offline

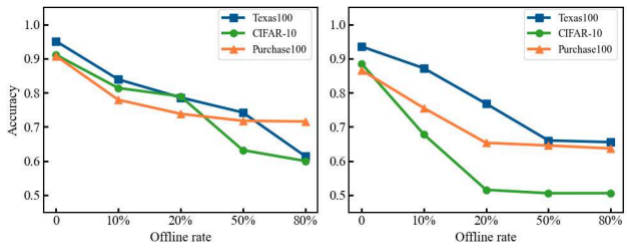


(a) Global attack (b) Local attack

Figure 3: Comparison of attack accuracy.

randomly in the iterations. The client randomly selects the offline iterations during the federated training process with a preset offline rate. The offline rate equals to the proportion of the offline iterations among the total number of training iterations.

Fig. 4 illustrates how the membership inference accuracy varies with the offline rate of the target client. We can see that the target client going offline has a negative effect on the accuracy. As the offline rate increases, the accuracy of GBMIA decreases on all datasets. The impact of the target client going offline on the global attack is smaller than its impact on the local attack. On the other hand, it is found that the non-target clients going offline has little effect on the membership inference accuracy. The reason is that the purpose of the active attack based on gradient ascent tunning is to make the target client reduce the loss of the model on the target data through SGD algorithm, and the non-target clients are not affected. This provides an inspiration for source inference attack [20] which distinguishes the target clients from the non-target clients. On the other side, in some works [22], [23], the server selects a subset of the clients participating in each iteration, aiming to reduce the communication overhead. According to experimental results, these approaches have a potential for membership protection.



(a) Global attack (b) Local attack

Figure 4: Comparison of attack accuracy.

4.4 GBMIA when clients' trainning sets overlap

The clients' training sets may overlap in FL. Take a real-world scenario as an example, where several hospitals jointly train a global model through FL. If a patient went to two hospitals for physical examinations, the two hospitals would have the same clinical data of the patient and their training sets overlap. We suppose that the target data exists in the training sets of N clients. We explore the relation between N and the accuracy of membership inference by adding the attacker's member test dataset to N clients' training sets, where N varies from 1 to 5 in global attack and from 1 to 4 in local attack. Fig.5 shows that the attack accuracy increases with N. The global attack and local attack both have an accuracy above 90% on all datasets when N 3. Therefore, it is suggested to remove the overlapped data points from the clients' training sets through private set intersection [24] to reduce the membership leakage.

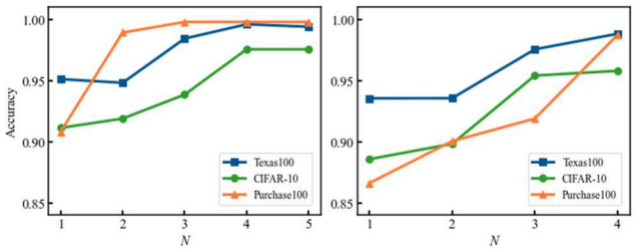


Figure 5: Comparison of attack accuracy.

5 Conclusion

In this paper, we proposed GBMIA, an attack model-free membership inference attack against FL based on gradient. We evaluated GBMIA on three real-world datasets. The experimental results showed that GBMIA achieves a higher attack accuracy than other membership inference methods. We also explored the factors that can affect the membership privacy leakage, and the findings can provide inspirations for the protections against MIA in FL. Our future work will concentrate on extending our attack approach to the privacy-preserving FL scenarios where the homomorphic encryption or differential privacy has been implemented.

6 References

[1] Y. Xiao, H. -H. Chen, X. Du and M. Guizani, "Stream-based cipher feedback mode in wireless error channel," IEEE Transactions on Wireless Communications, vol. 8, no. 2, pp. 622-626, Feb. 2009.

[2] D. Wang, B. Song, D. Chen and X. Du, "Intelligent Cognitive Radio in 5G: AI-Based Hierarchical Cognitive Cellular Networks," IEEE Wireless Communications, vol. 26, no. 3, pp. 54-61, June 2019.

[3] W. Y. B. Lim et al., "Federated Learning in Mob ile Edge Networks: A Comprehensive Survey," IEEE Communications Surveys and Tutorials, vol. 22, no. 3, pp. 2031-2063, thirdquarter 2020.

[4] N. Wang, W. Yang, Z. Guan, X. Du and M. Guizani, "BPFL: A Blockchain Based Privacy-Preserving Federated Learning Scheme," in 2021 IEEE Global Communications Conference (GLOBECOM), pp. 1-6, Madrid, Spain, 2021.

[5] M. Shen et al., "Exploiting Unintended Property Leakage in Blockchain-Assisted Federated Learning for Intelligent Edge Computing," IEEE Internet of Things Journal, vol. 8, no. 4, pp. 2265-2275, Feb. 2021.

[6] R. Shokri, M. Stronati, C. Song and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," in 2017 IEEE Symposium on Security and Privacy (SP), pp. 3-18, San Jose, CA, USA, 2017.

[7] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier and Hervé Jégou, "White-box vs black-box: Bayes optimal strategies for membership inference," in International Conference on Machine Learning (ICML), pp. 5558-5567, Long Beach, CA, USA, 2019.

[8] H. Hu, Z. Salcic, G. Dobbie and X. Zhang, "Membership inference attacks on machine learning: A survey," arXiv preprint, arXiv:2103.07853, 2021

[9] J. Zhang, J. Zhang, J. Chen and S. Yu, "GAN Enhanced Membership Inference: A Passive Local Attack in Federated Learning," in ICC 2020 - 2020 IEEE International Conference on Communications (ICC), pp. 1-6, Dublin, Ireland, 2020.

[10] S. Yeom, I. Giacomelli, M. Fredrikson and S. Jha, "Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting," in 2018 IEEE 31st Computer Security Foundations Symposium (CSF), pp. 268-282, Oxford, UK, 2018.

[11] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M.Backes, "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models," in Network and Distributed System Security Symposium (NDSS), San Diego, CA, USA, 2019.

[12] M. Nasr, R. Shokri and A. Houmansadr, "Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning," in 2019 IEEE Symposium on Security and Privacy (SP), pp. 739-753, San Francisco, CA, USA, 2019.

[13] S. Xuan, L. Zheng, I. Chung, W. Wang, D. Man, X. Du, W. Yang, M. Guizani, "An incentive mechanism for data sharing based on blockchain with smart contracts", Computers and Electrical Engineering, Vol. 83, May 2020.

[14] Y. Yu, L. Xue, Y. Li, X. Du, M. Guizani, B. Yang, "Assur ed data deletion with fine-grained access control for fog-based industrial applications", IEEE Transactions on Industrial Informatics, Vol. 14, Issue 1), Pages 4538-4547, May 2018.

[15] A. Mittos, B. Malin and E. De Cristofaro, "Systematizing genome privacy research: A privacy-e Membership Inference Attacks Against Recom-

mender Systems nhancing technologies perspective," in 19th Privacy Enhancing Technologies Symposium (PETs), pp. 87-107, Stockholm, Swe den, 2019.

[16] Apostolos Pyrgelis, Carmela Tr oncoso, Emiliano De Cristofaro, "Knock knock, who's there? membership inference on aggregate location data", in Network and Distributed System Security Symposium (NDSS), San Diego, CA, USA, 2017.

[17] Zhang Z, Yan C and Malin BA, "Membership inference attacks against synthetic health data," Journal of biomedical informatics, vol. 125, pp. 103977, 2022.

[18] Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini and Nicolas Papernot, "Label-only membership inference attacks," in International Conference on Machine Learning (ICML), pp. 1964-1974, virtual, 2021.

[19] K. Leino and M. Fredrikson, "Stolen memories: Leveraging model memorization for calibrated white-box membership inference," in 29th USENIX Security Symposium (USENIX Security '20), pp. 1605-1622, Boston, MA, USA, 2020.

[20] H. Hu, Z. Salcic, L. Sun, G. Dobbie and X. Zhang, "Source Inference Attacks in Federated Learning," in 2021 IEEE International Conference on Data Mining (ICDM), pp. 1102-1107, Auckland, New Zealand, 2021.

[21] Abhyuday Jagannatha, Bhanu Pratap Singh Rawat, and Hong Yu, "Membership inference attack susceptibility of clinical language models," arXiv preprint, arXiv:2104.08305, 2021.

[22] Y. Fraboni, R. Vidal, L. Kameni and M. Lorenzi, "Clustered sampling: Low-variance and improved representativity for clients selection in federated learning," in International Conference on Machine Learning (ICML), pp. 3407-3416, virtual, 2021.

[23] L. Nagalapatti and R. Narayanam, "Game of gradients: Mitigating irrelevant clients in federated learning," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 10, pp. 9046– 9054, 2021.

[24] C. Hazay and M. Venkitasubramaniam, "Scalable Multi-party Private Set-Intersection," Public-Key Cryptography, vol. 10174, pp. 175-203, 2017.