

## Article

# Threshold Filtering for Detecting Label Inference Attacks in Vertical Federated Learning

Liansheng Ding <sup>1</sup>, Haibin Bao <sup>1</sup>, Qingzhe Lv <sup>2</sup>, Feng Zhang <sup>3,\*</sup>, Zhouyang Zhang <sup>4</sup>, Jianliang Han <sup>5</sup> and Shuang Ding <sup>6</sup>

<sup>1</sup> Guoneng Zhejiang Beilun No. 1 Power Generation Co., Ltd., Hangzhou 315800, China; 16124054@ceic.com (L.D.); baohaibin@126.com (H.B.)

<sup>2</sup> Hangzhou Post Quantum Cryptography Technology Co., Ltd., Hangzhou 311122, China; lvqingzhe2021@163.com

<sup>3</sup> School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

<sup>4</sup> Zhejiang Laboratory, Hangzhou 311121, China; zhangzhouyang\_2021@163.com

<sup>5</sup> Shaoxing Big Data Assurance Centre, Shaoxing 312000, China; sxhj12013@163.com

<sup>6</sup> Shaoxing Data Co., Ltd., Shaoxing 312099, China; sxdc@sx-gzyy.com

\* Correspondence: f\_zhang@nuaa.edu.cn or zhangfeng\_202109@163.com

**Abstract:** Federated learning, as an emerging machine-learning method, has received widespread attention because it allows users to train locally during the training process and uses relevant cryptographic knowledge to safeguard the privacy of data during model aggregation. However, existing federated learning is also susceptible to privacy breaches, e.g., label inference attacks against vertical federated learning scenarios, where an adversary is able to reason about the labels of other participants based on the trained model, leading to serious privacy breaches. In this paper, we design a detection method for label inference attacks in vertical federated learning scenarios, which is able to detect the attacks based on the principles of the attacks. We design a threshold-filtering detection method based on the principle of attack to determine that the model is under attack when the threshold value is greater than a set parameter. Furthermore, we have created six threat model classifications based on different a priori conditions of the adversary to comprehensively analyze the adversary's attacks. In addition to the detection method of attacks, the extent of attacks on the model and the effectiveness of the defense can also be evaluated. The evaluation module will experimentally measure the changes in the relevant metrics such as the accuracy of the attack, the F1 score, and the change in the accuracy after the defense method. For example, detection in the full connected neural network model assesses the attack and defense effectiveness of the model with an attack accuracy of 86.72% in the breast cancer Wisconsin dataset and an F1 score of 0.743, which is reduced to 36.36% after dispersed training. This ensures that users have an overall grasp of the extent to which the training model is under attack before deploying the model.

**Keywords:** federated learning; label inference attacks; detection; security and privacy



**Citation:** Ding, L.; Bao, H.; Lv, Q.; Zhang, F.; Zhang, Z.; Han, J.; Ding, S. Threshold Filtering for Detecting Label Inference Attacks in Vertical Federated Learning. *Electronics* **2024**, *13*, 4376. <https://doi.org/10.3390/electronics13224376>

Academic Editor: Hung-Yu Chien

Received: 18 September 2024

Revised: 20 October 2024

Accepted: 23 October 2024

Published: 8 November 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Due to the great results of machine learning and federated learning in the past, a growing number of people are using machine learning models for predictive analytics data. The scope of federated learning applications has become broader; taking the medical scene scenario as an example, the use of federated learning can predict rare case onsets so that the relevant power units are better able to prepare for the development of the patient's condition in advance, ensuring better treatment.

However, if a vulnerable machine-learning model is deployed, it not only leaks the raw data but also results in privacy breaches for other participating parties [1–4]. Inference attacks are able to exploit intermediate results, such as gradients in the training process of a model, to cause serious privacy breaches. There are fewer existing tests for

vertical federated learning label inference attacks; although there are partial studies on both attacks and defense issues of vertical federated learning, they are all specific to one kind of research, all in isolation. Thus, there is a lack of a holistic grasp of the security issues of vertical federated learning, e.g., scenarios they are applied to respectively, such as whether label inference attacks conflict with other attacks and how to identify the label inference attacks—they are all needed [5,6].

In this paper, a detection method for label inference attacks in vertical federated learning is designed that enables the detection of label inference attacks, in addition to extending the detection functionality to enable the identification of other attacks such as membership inference attacks (determining whether or not a particular piece of data is within the training dataset). For models subjected to label inference attacks and membership inference attacks, they are also evaluated on the basis of the attacks they have been subjected to, including metrics such as the accuracy of the attacks, F1 scores, etc., which facilitate the user's overall understanding of the model. In addition to this, the defense attacks designed in the previous sections are integrated to this module, which enables a defense against label inference attacks. The design of such a detection function module fills the gap of overall risk assessment for label inference attacks.

As depicted in Figure 1, this paper designs a detection module for vertical federated learning. For the trained vertical federated-learning model, it is able to realize the detection of label inference attacks and membership inference attacks in vertical federated learning. For the different conditions possessed by the adversary, six threat models are analyzed and designed, and the attack effect is evaluated for each threat model. For models that have been attacked, a thorough evaluation of the attack is performed to assess the extent of the attack suffered, and a graph of the change in accuracy, F1 score, etc., is output for the effect of the attack. In addition, in the defense module, a dispersed training implementation is introduced to mitigate the risk of label inference attacks, and the defense effect is analyzed experimentally. With the detection module of vertical federated learning, users are able to use this detection module to make a comprehensive assessment of the potential security and privacy risks associated with the model prior to its deployment.

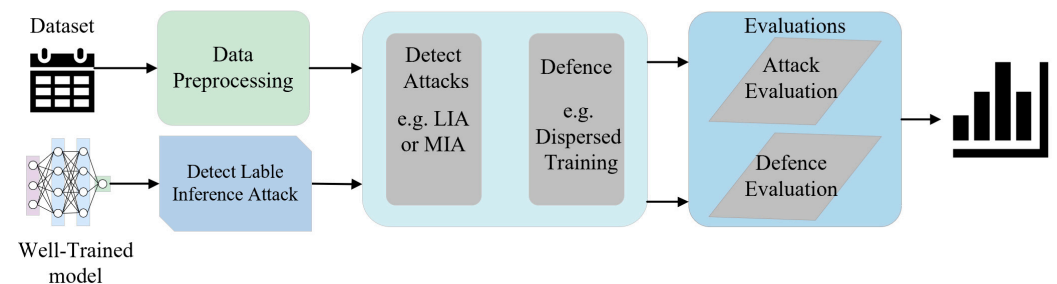


**Figure 1.** Framework of the detection method.

Overall, this paper contributes in the following ways:

- (1) We design a detection module for label inference attacks in vertical federated learning and classify threat models based on adversaries possessing different prior knowledge. Compared to the work of others, we refine the threat model classification to make the attacks more relevant to real attack scenarios. This refined classification helps us to understand more about the attack principles of the adversary. For label inference attacks, we classify based on whether the adversary uses additional labels in training, and for membership inference attacks, we classify based on the adversary's use of additional datasets and use of shaded datasets. The classification of various threat models facilitates the study of adversaries.
- (2) We propose a detection module that can detect label inference attacks in addition to membership inference attacks in vertical federated learning. In addition, in order to facilitate users to better understand their model situation, the detection module can also detect the degree of attack on the attacked model, which facilitates users' ability

to have an overall understanding of the trained model. We evaluate the BCW dataset for attacks on four models, FCNN, Alexnet, RetNet, and DenseNet, and the accuracies of the label inference attacks are 86.72%, 86.45%, 78.42%, and 77.69%, respectively.

- (3) As compared to previous work, we have designed a detection method to combine attack detection and evaluation for the models under attack to be evaluated for attacks, in addition to accessing dispersed training methods for defending against attacks and evaluating the effectiveness of the defense. For example, in the BCW dataset, we evaluate the effectiveness of the defense after accessing the dispersed training, and the accuracy of the label inference attack is reduced to 36.36% in the BCW dataset after defense, and the accuracy of the attack in the other image datasets is also reduced to about 10% of the random guessing. This detection module allows users to perform a comprehensive assessment of the potential security and privacy risks of a model before deploying it.

## 2. Related Work

### 2.1. Development Label Inference Attack Models

The label inference attacks on vertical federated learning were presented by Fu et al. at USENIX 2022 [5], the first paper on the label leakage problem for vertical federated learning. In fact, the relevant security papers on vertical federated learning go back to the model-divided federated learning proposed by Vepakomma et al., who found that federated learning can be segmented into two groups: model-divided federated learning and federated learning without model-divided federated learning [7]. For federated learning with model partitioning, the data and labels of the participants are separated, and data security is ensured with this method, but later research proved that this method is also open to attack. Later, Zhao et al. proposed iDLG based on the principle of label leakage, which is a deep leakage of gradients; for data reconstruction attacks in horizontal federated learning, the first step in performing data reconstruction is to process the symbols of the gradient, and Fu et al. demonstrated the feasibility of direct label inference attacks based on this approach [8]. Zhu et al., on the basis of iDLG, discovered the profound leakage of gradients during the analysis of leaked gradients. This included the revelation that the notation of the gradient can recover information, and the gradient can also directly restore data information [9].

### 2.2. Defense and Detect Label Inference Attacks

Label inference attacks have attracted wide attention since they were put forward. Wang et al. [6] proposed a method of decentralized training to defend against such attacks by separating the link between the labels and data, and experiments have proved that this method is effective.

At present, most of the work is focused on studying the defense of label inference attacks and how label inference attacks are caused, and there is no research on the detection method of label inference attacks.

## 3. Threat Models

Table 1 shows all the threat models of inference attacks against adversaries with different prior knowledge, and for label inference attacks, the attack principle is that the adversary exploits the “memorability” of the underlying model to execute the attack. The only thing to discuss categorically is whether the adversary uses additional labeled data when fine-tuning the underlying model, and we analyze this condition experimentally as a control group. The reason for doing so is based on the reasoning attack principle. Common member reasoning attacks and label reasoning attacks are based on a small amount of prior knowledge without knowing the original dataset. Therefore, the first distinguishing point of our threat model design is whether to use a small number of labeled datasets [4]. Based on the information possessed by the adversary, we classify it into two threat models; the adversary using a small amount of labeled data is set as Threat Model 1, and the adversary

not using labeled data is set as Threat Model 2. Labeled data are critical to the adversary, and by fine-tuning data with labels for the application of semi-supervised learning, the adversary is able to use this labeled data to more accurately learn more relevant information about the labels and features, which makes the adversary's ability to attack stronger [10–13].

**Table 1.** Classification of threat models.

| Threat Model   | Conditions of the Threat Model                                    |
|----------------|-------------------------------------------------------------------|
| Threat Model 1 | use of some additional data and a small amount of labeled data    |
| Threat Model 2 | use of some additional data and no use of additional labeled data |
| Threat Model 3 | use of some additional data and has black box access              |
| Threat Model 4 | use of some additional data and has black box access              |
| Threat Model 5 | uses the shadow dataset and has black box access                  |
| Threat Model 6 | uses the shadow dataset and has black box access                  |

Regarding attacks aimed at inferring membership, the types of adversary access are categorized into full access and restricted access. White-box access means that the adversary possesses complete information regarding the target model, encompassing both the model's parameters and its structure. Black-box access means that the adversary does not know any information, parameters or structure of the target model, and the adversary is only able to access it tentatively through the API. MIA by exploiting the overfitting of models that have “memory” of trained data. Because, in the scenario of MIA, the adversary possesses limited access, it is impossible to obtain information about the model parameters and structure of the target model. Therefore, the adversary is limited to utilizing an auxiliary dataset or leveraging known statistical information to generate a surrogate model used to train the attack model. Hence, when designing the threat model, we should also take this into consideration [14].

In summary, label inference attacks are categorized into two threat models depending on the prior conditions possessed by the adversary. Membership inference attacks are divided into a total of four threat models in two scenarios (black-box scenario and white-box scenario) depending on the prior conditions of the adversary.

#### 4. Overall Architecture

##### 4.1. Detecting Label Inference Attack Models

Label inference attacks are based on the principle that an adversary utilizes a participant's own model to carry out an attack on other participants. In the vertical federated learning scenario, labels are held by only one party, which means that the other participants only have some data features. The attacker employs a scant portion of the labeled data at their disposal to fine-tune their participating model through semi-supervised learning training. This enables the adversary's model to better learn the relationship between the labels and features, transforming the model into an attack model for targeted label attacks [15–18]. Taking the principle of label inference attack as an example formula (1), we judge whether the label is leaked according to the symbol of the gradient, because  $\sum_j e^{y_j^a} \in (0, 1)$ ; so, when the discriminating label is a real label ( $i = c$ ),  $g_i^a < 0$ . When the discriminating label is not a real label, ( $i \neq c$ ),  $g_i^a > 0$  because you can directly attack the label according to the symbol of  $g_i^a$ . This feature is used to obtain the gradient correlation information of federation learning without model partitioning, so as to detect the label inference attack on the model. Our detection method is to use the different principles of label inference attack symbols to detect the model.

$$g_i^a = \frac{\partial \text{loss}(x, c)}{\partial y_i^a} = - \frac{\partial \log e^{y_c^a} - \partial \log \sum_j e^{y_j^a}}{\partial y_i^a} = \begin{cases} -1 + \frac{e^{y_i^a}}{\sum_j e^{y_j^a}} & \text{if } i = c \\ \frac{e^{y_i^a}}{\sum_j e^{y_j^a}} & \text{if } i \neq c \end{cases} \quad (1)$$

#### 4.2. Detecting Membership Inference Attack Models

Membership inference attacks have received extensive research and attention as one of the most common privacy attacks. Shokri et al. [10] proved that membership inference attacks are due to overfitting through assumptions such as shadow models, use of multiple attack models, etc. Using this principle, we can detect and evaluate membership inference attacks. When data come for training, the attack model makes a judgment about the behavior of the data in the target model and uses the difference in behavior to ascertain whether it denotes a member within the model or a non-member. A piece of data that is used as training for a model will be “remembered” by the ML model, and the posterior probability of the data will be higher, whereas for a model that has not been trained, the posterior probability of its output will be lower than the posterior probability of the output of the trained data. Using this feature, we are able to make judgments based on the posterior probability of the output [19].

#### 4.3. Experimental Evaluation Module for Attack and Defense

For models that are vulnerable to attacks, the models that are attacked are evaluated as a whole, in addition to detecting the vulnerability of the models to that type of inference attack. For models subject to membership inference attacks, experimental data are presented and analyzed for two threat models (whether the attacker utilizes a limited quantity of labeled data). For models subjected to membership inference attacks, experiments are conducted for four threat models, namely, the white-box scenario where the adversary uses part of the original dataset; the shadow dataset, where the adversary simulates the real dataset through statistical information; and the black-box scenario, where the adversary uses part of the original dataset and uses the shadow dataset, and experiments are performed to assess and analyze the accuracy of the attacks of the four threat models. In the later experiments, we focus on analyzing the threat model in black-box scenarios, which are closer to realistic attack scenarios for adversaries [20–22]. Figure 2 shows the entire process of detecting the framework, starting with determining what attack the model is under and the member inference or tag inference, and then evaluating each attack and defense.

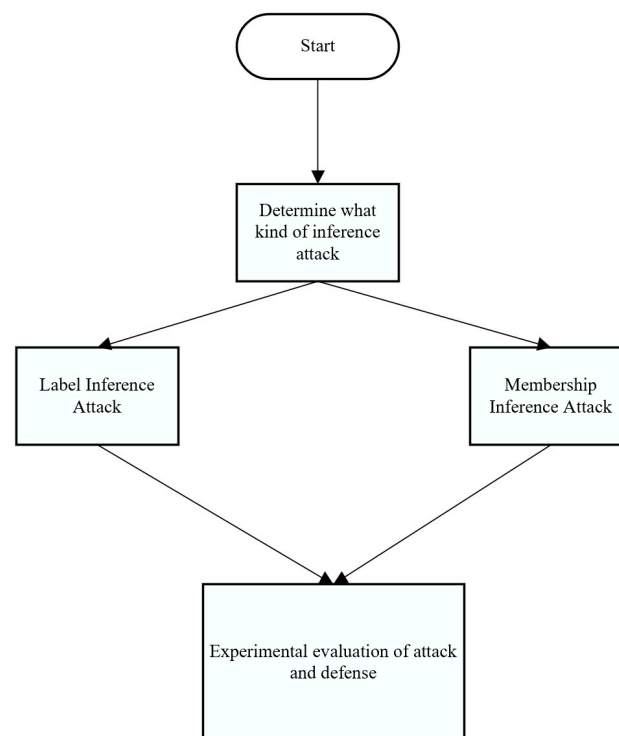


**Figure 2.** Steps to detect the model.

## 5. Experiment

### 5.1. Dataset Setup

For better validation of the experiments, we chose a common dataset, the same one used by Fu et al [5], and added a new dataset, FMNIST, for the overall evaluation of the model attacks and defenses. Table 2 shows the various Settings for the dataset.

**Table 2.** Threat model under different conditions of membership inference attacks.

| Dataset       | Number of Training Datasets | Number of Test Datasets |
|---------------|-----------------------------|-------------------------|
| BCW           | 426                         | 143                     |
| CIFAR-10      | 50,000                      | 10,000                  |
| CINIC-10      | 180,000                     | 90,000                  |
| Fashion-MNIST | 60,000                      | 10,000                  |

For all four datasets, their type was picture type; for the vertical federated learning, it was possible to divide the pictures according to the quality of participants. For example, vertical federated learning had two participants for the model's training, which enabled a dataset to split the image of a dataset into two parts. Table 3 provides detailed information on the four datasets.

**Table 3.** Experimental datasets and corresponding model structures.

| Dataset       | Bottom Model Structure | Top Model Structure |
|---------------|------------------------|---------------------|
| BCW           | FCNN-3                 | FCNN-3              |
| CIFAR-10      | RetNet-18              | FCNN-4              |
| CINIC-10      | RetNet-18              | FCNN-4              |
| Fashion-MNIST | RetNet-18              | FCNN-4              |

### 5.2. Model Structure Setup

To guarantee the comprehensiveness of both the defense and attack, we selected the federated learning model used in Table 4. For BCW dataset, the top model used a fully connected neural network with three fully connected layers (FCNN-3), and the bottom model also used FCNN-3. A FCNN-4 was selected for the CIFAR-10, CINIC-10, and Fashion-MNIST top-layer models, and a residual neural network with 18 layers was selected for the bottom layer model.

**Table 4.** RetNet-18 network parameters.

| Residual Block 1            | Residual Block 2             | Residual Block 3             | Residual Block 4             |
|-----------------------------|------------------------------|------------------------------|------------------------------|
| Conv $3 \times 3$ , 64 ReLU | Conv $3 \times 3$ , 128 ReLU | Conv $3 \times 3$ , 256 ReLU | Conv $3 \times 3$ , 512 ReLU |
| Conv $3 \times 3$ , 64 ReLU | Conv $3 \times 3$ , 128 ReLU | Conv $3 \times 3$ , 256 ReLU | Conv $3 \times 3$ , 512 ReLU |

Table 5 illustrates the structural framework diagram of the selected RetNet-18. For the fully connected neural network, the designed model incorporated a regularization module to prevent overfitting. For the underlying model of the other three datasets to be capable of ensuring the proper transmission of the gradient, the RetNet-18 model was chosen, which allowed for the solution of the gradient problem in neural networks through residual connectivity. The size of the convolution kernel used by RetNet-18 was  $3 \times 3$ .

**Table 5.** Indicates the output of the detection module.

| Output                  | Label Inference Attacks | Membership Inference Attacks | Both Attacks |
|-------------------------|-------------------------|------------------------------|--------------|
| Detection module output | 1                       | 2                            | 1, 2         |

### 5.3. Experimental Evaluation

In the detection function of label inference attacks and membership inference attacks, the attack accuracy of the model subject to both attacks, alongside the accuracy of the label inference attacks model after decentralized training, were experimentally verified and evaluated. The inspection module allows the user to have an overview of the model before deploying it.

#### 5.3.1. Evaluation of Detection Modules

For vertical federated learning in the detection function, it is necessary that the advance detection model of the user be easy to deploy in order to perform the risk prediction. For the detection module of vertical federated learning, we designed our approach to detect two kinds of attacks: label inference attacks and membership inference attacks.

As shown in Figure 3, the detection model detected both attacks, and the detection module output 1 for the models that were susceptible to label inference attacks, 2 for models that were susceptible to membership inference attacks, and 0 for models that were not susceptible to either attack. Membership inference attacks rely on the principle of overfitting, and the usual models were susceptible to membership inference attacks; for models that were susceptible to both attacks, the detection module output 1 and 2.

Detection result:0

Detection result:1

Detection result:2

Detection result:1,2

**Figure 3.** Detection module output results.

#### 5.3.2. Attack Evaluation

An experimental analysis of the attacks was performed for the threat models subjected to label inference attacks and membership inference attacks. We analyzed the accuracy of the attacks for each of the two cases (two threat models) of the label inference attacks for each of the four datasets, as well as F1 scores of the membership inference attacks for the different datasets and different threat models.

Figure 4 shows the accuracy of the attacking model formed by the adversary using a limited quantity of labeled data for pre-training and the attack accuracy of the model formed by the attack without utilizing a limited quantity of labeled data for fine-tuning in the two threat model conditions. The experiments here were for the accuracy of the active label inference attacks, and they showed that the attack performed by the adversary with a limited quantity of labeled data was better under Threat Model 1, which was the attack performed by the adversary with a limited quantity of labeled data. The accuracy of the attack was as high as 89% on the BCW dataset, which was a big threat to the user's privacy. Similarly, there was a high accuracy on several other relatively complex datasets, with 86.52%, 84.11%, and 80.82% for attacks on the CIFAR-10, CINIC-10, and FMNIST datasets, respectively. For Threat Model 2, the attack scenarios were executed by the adversary using a limited quantity of labeled data. Relative to Threat Model 1, the accuracy of the attack decreased slightly, with 78% on the BCW dataset and less than 80% on the other datasets and even 69% on the FMNIST dataset. This is because of the different information possessed by the enemy; an enemy that has a limited quantity of labeled data acquires more information about the features and labels when performing semi-supervised learning to fine-tune the attack model, resulting in a better attack, which is the principle of label inference attacks.

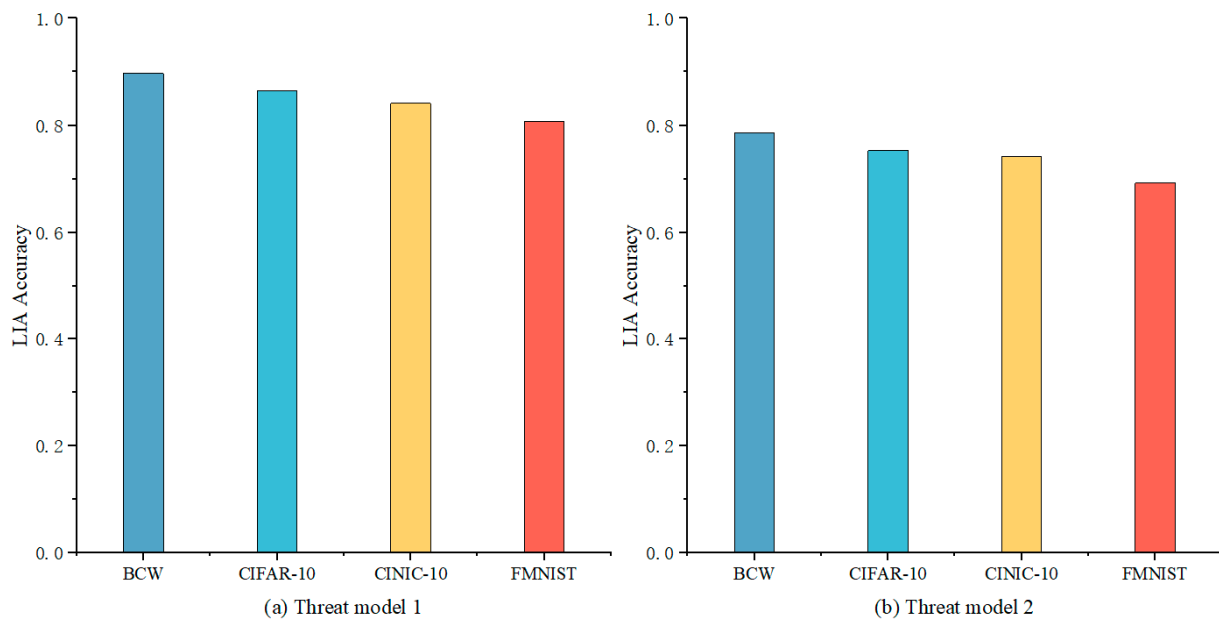


**Figure 4.** Attack accuracy of the two threat models for LIA.

As shown in Figure 5, the accuracy of the membership inference attacks was also analyzed experimentally. For the RetNet models in machine learning, we focused on two threat models in black-box scenarios. Because the black-box scenario is more representative of a realistic attack situation as compared to the white-box, for a realistic scenario, the adversary typically lacks the specific structure and specific parameters of the model, so the black-box scenario is more representative of a realistic attack situation; that is, the adversary utilizes the API to access the model. As Figure 5 shows, two threat models for membership inference attacks were analyzed experimentally on four datasets.

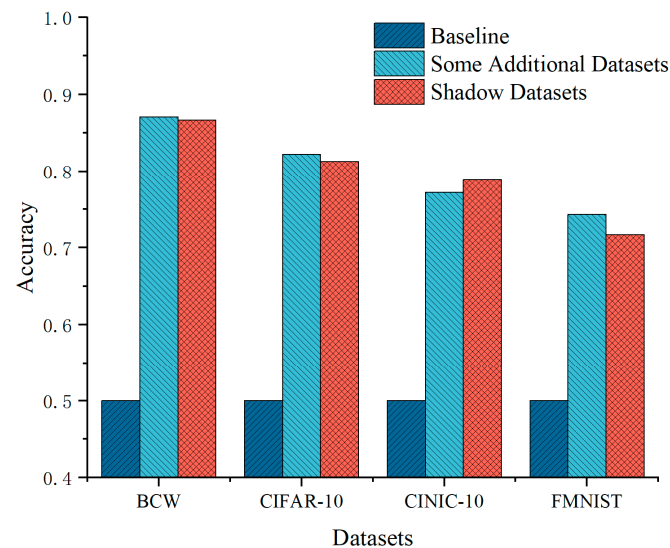


**Figure 5.** MIA accuracy assessment.

As shown in Figure 6, on the BCW dataset, the effectiveness of membership inference attacks under Threat Model 3 was 87.1% for adversaries using additional datasets, while the accuracy of attacks under Threat Model 5 was 86.7% for adversaries using shadow models. However, in the CINIC-10 dataset, the adversary's attack accuracy using some of the additional datasets was slightly lower than the attack accuracy using the shadow dataset, and the reason for this result may be that, when modeling the shadow dataset, the

adversary may have gained more information about the real dataset, which resulted in a better-trained shadow dataset. Overall, the adversary's attack on the model using some of the extra real datasets was better than the adversary's use of the shadow datasets, and the same attack worked for the CIFAR-10 and FMNIST.

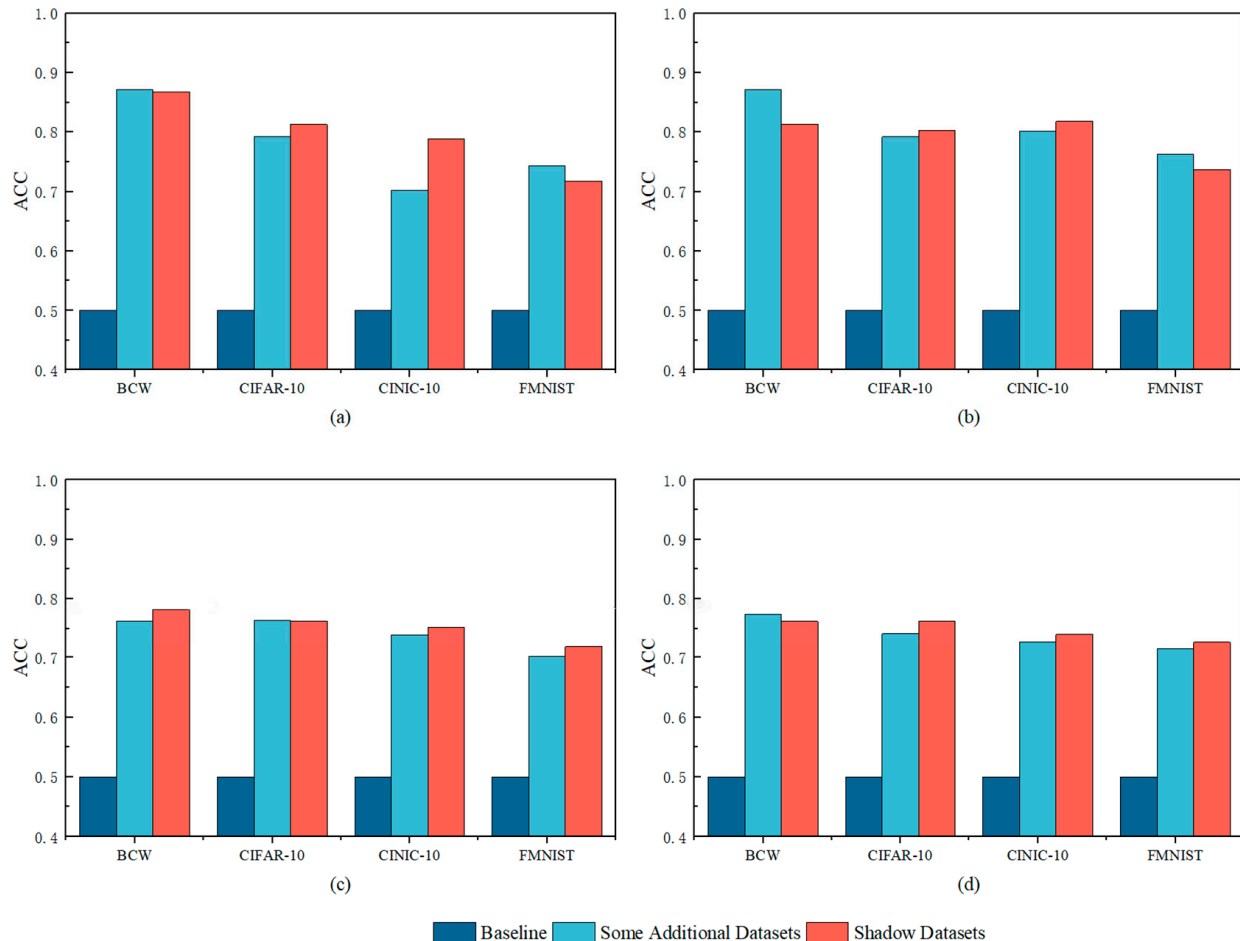


**Figure 6.** (a–d) Evaluation of the MIA accuracy in the four models.

Figure 7 shows the variation in the accuracy of attacks on the two threat models under different machine models. Figures 6 and 7a–d represents different threat models. We chose FCNN, Alexnet, RetNet, and DenseNet as the target models for experimental analysis. It is observed that, for Threat Model 3, the adversary had a real dataset and performed the membership inference attacks with a little better accuracy. The F1 scores of the attacks under the four models were also evaluated experimentally and again compared to Threat Model 5 (where the adversary utilized a shadow dataset); the accuracy of Threat Model 3 attacks was a bit better.

### 5.3.3. Defense Assessment

The analysis of the defense method was also performed experimentally for the defense method in vertical federated learning, and the defense method was added to the model using the defense method of label inference attacks based on the decentralized training in Chapter Four.

As Figure 8 shows, the accuracy of the attack and the accuracy of the defense were validated experimentally for the four datasets BCW, CIFAR-10, CINIC-10, and FMNIST trained in vertical federated learning. In the four datasets, the original accuracy of more than 80% was reduced to about 10% because in the CIFAR-10, CINIC-10, and FMNIST dataset, a total of ten categories, 10%, represents the results of the guessing, proving

the efficacy of the designed defense method. At the same time in the detection module, adding the label inference attack defense function was able to make the detection module more complete.

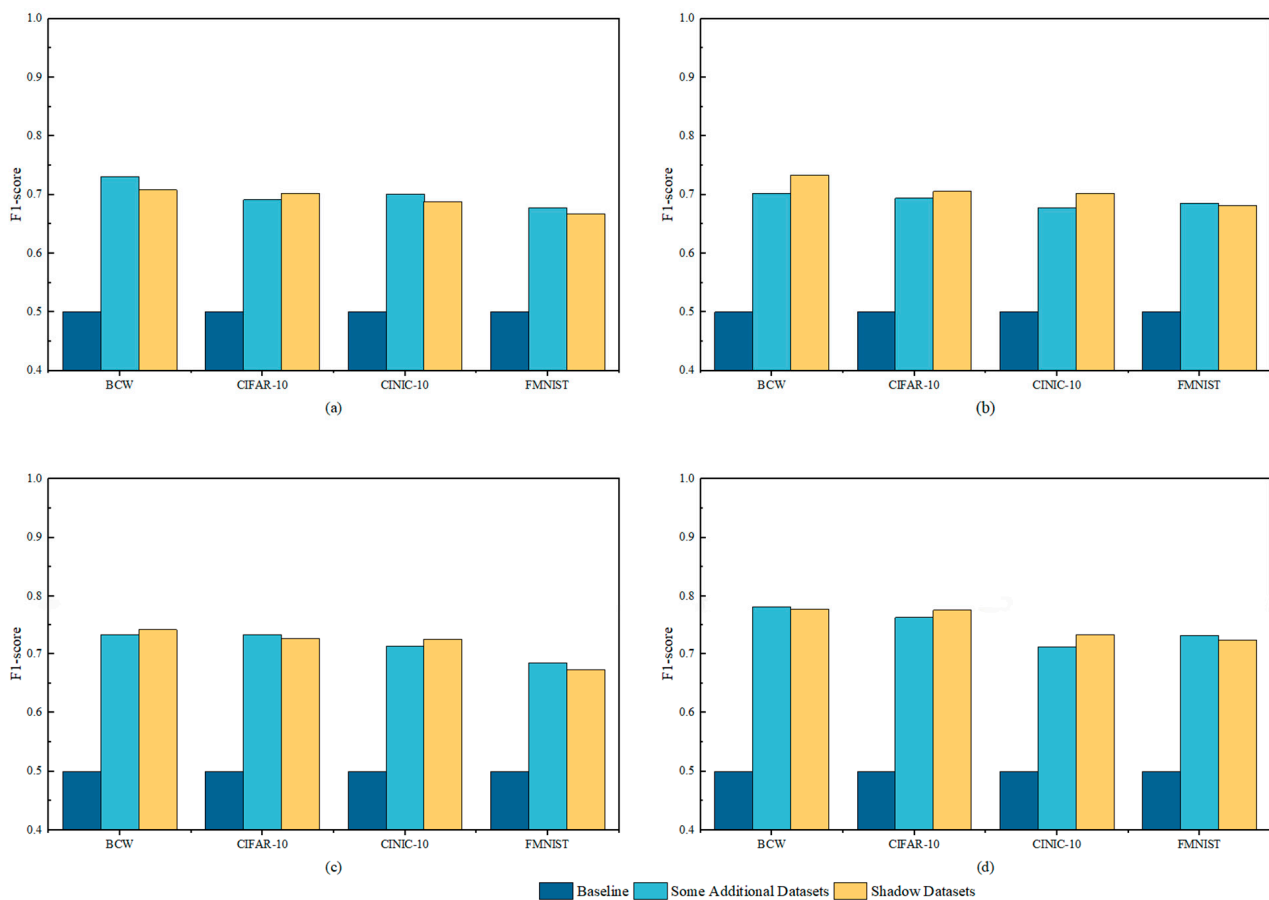


**Figure 7.** (a–d) Evaluation of the MIA defense in the four models.

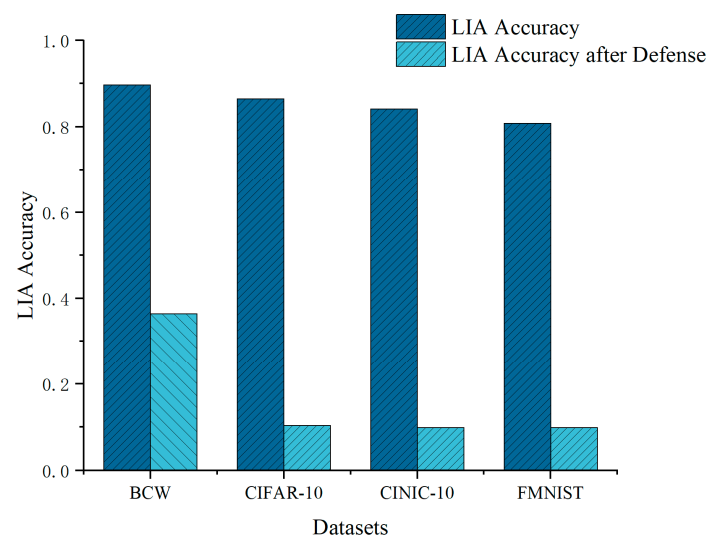


**Figure 8.** MIA defense assessment.

## 6. Conclusions

This paper presents the initial thorough analysis of the privacy vulnerabilities posed by inference attacks targeting machine learning models. In view of the different reasoning

conditions of the adversary, six threat models are set up to evaluate the experiment as a whole. We perform a thorough measurement study involving distinct constructions and datasets, focusing on both the attack and defense strategies. For users who want to deploy the models, it is possible to detect possible attacks on the models in advance, and for models that have been attacked, by using the detection module, it is possible to assess the extent to which the model has been attacked, as well as to assess the effectiveness of the defense. While the detection functionality is tailored for inference attacks, we intend to incorporate tools designed to assess risks aimed at compromising the model's functionality, e.g., adversarial examples. This detection capability can achieve AI security and trustworthiness.

**Author Contributions:** Methodology, L.D.; Software, H.B.; Validation, F.Z. and S.D.; Investigation, Q.L.; Resources, Z.Z.; Data curation, J.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This study is supported by the National Natural Science Foundation of China (Grant No. 62332004).

**Data Availability Statement:** All data can be requested from the second author.

**Acknowledgments:** The authors wish to express sincere appreciation to the reviewers for their valuable comments, which significantly improved this paper.

**Conflicts of Interest:** Authors Liansheng Ding and Haibin Baowas employed by the company Guoneng Zhejiang Beilun No. 1 Power Generation Co., Ltd., Author Qingzhe Lv employed by the company Hangzhou Post Quantum Cryptography Technology Co., Ltd., Author Shuang Ding employed by the company Shaoxing Data Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References

1. Yang, Q.; Liu, Y.; Chen, T.; Tong, Y. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol. (TIST)* **2019**, *10*, 1–19. [[CrossRef](#)]
2. Cao, M.; Zhang, L.; Cao, B. Toward on-device federated learning: A direct acyclic graph-based blockchain approach. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *34*, 2028–2042. [[CrossRef](#)] [[PubMed](#)]
3. Nguyen, D.C.; Pham, Q.V.; Pathirana, P.N.; Ding, M.; Seneviratne, A.; Lin, Z.; Dobre, O.; Hwang, W.-J. Federated learning for smart healthcare: A survey. *ACM Comput. Surv. (CSUR)* **2022**, *55*, 1–37. [[CrossRef](#)]
4. Zhang, W.; Li, Z.; Chen, X. Quality-aware user recruitment based on federated learning in mobile crowd sensing. *Tsinghua Sci. Technol.* **2021**, *26*, 869–877. [[CrossRef](#)]
5. Fu, C.; Zhang, X.; Ji, S.; Chen, J.; Wu, J.; Guo, S.; Zhou, J.; Liu, A.X.; Wang, T. Label inference attacks against vertical federated learning. In Proceedings of the 31st USENIX Security Symposium (USENIX Security 22), Boston, MA, USA, 10–12 August 2022; pp. 1397–1414.
6. Wang, Y.; Lv, Q.; Zhang, H.; Zhao, M.; Sun, Y.; Ran, L.; Li, T. Beyond model splitting: Preventing label inference attacks in vertical federated learning with dispersed training. *World Wide Web.* **2023**, *26*, 2691–2707. [[CrossRef](#)]
7. Vepakomma, P.; Gupta, O.; Swedish, T.; Raskar, R. Split learning for health: Distributed deep learning without sharing raw patient data. *arXiv* **2018**, arXiv:1812.00564.
8. Zhao, B.; Mopuri, K.R.; Bilen, H. iDLG: Improved deep leakage from gradients. *arXiv* **2020**, arXiv:2001.02610.
9. Zhu, L.; Liu, Z.; Han, S. Deep leakage from gradients. In Proceedings of the Advances in Neural Information Processing Systems 32 (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019.
10. Shokri, R.; Stronati, M.; Song, C.; Shmatikov, V. Membership inference attacks against machine learning models. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–24 May 2017; pp. 3–18.
11. Shokri, R.; Shmatikov, V. Privacy-preserving deep learning. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, 12–16 October 2015; pp. 1310–1321.
12. Fang, M.; Cao, X.; Jia, J.; Gong, N. Local model poisoning attacks to {Byzantine-Robust} federated learning. In Proceedings of the 29th USENIX Security Symposium (USENIX Security 20), Online, 12–14 August 2020; pp. 1605–1622.
13. Tramèr, F.; Zhang, F.; Juels, A.; Reiter, M.K.; Ristenpart, T. Stealing machine learning models via prediction {APIs}. In Proceedings of the 25th USENIX Security Symposium (USENIX Security 16), Austin, TX, USA, 10–12 August 2016; pp. 601–618.
14. Ressi, D.; Romanello, R.; Piazza, C.; Rossi, S. AI-enhanced blockchain technology: A review of advancements and opportunities. *J. Netw. Comput. Appl.* **2024**, *225*, 103858. [[CrossRef](#)]

15. Li, Z.; Zhang, Y. Membership leakage in label-only exposures. In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, 15–19 November 2021; pp. 880–895.
16. Liu, Y.; Zhao, Z.; Backes, M.; Zhang, Y. Membership inference attacks by exploiting loss trajectory. In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, Los Angeles, CA, USA, 7–11 November 2022; pp. 2085–2098.
17. Ye, J.; Maddi, A.; Murakonda, S.K.; Bindschaedler, V.; Shokri, R. Enhanced membership inference attacks against machine learning models. In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, Los Angeles, CA, USA, 7–11 November 2022; pp. 3093–3106.
18. Rezaei, S.; Liu, X. On the difficulty of membership inference attacks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7892–7900.
19. Liu, Y.; Wen, R.; He, X.; Salem, A.; Zhang, Z.; Backes, M.; De Cristofaro, E.; Fritz, M.; Zhang, Y. {ML-Doctor}: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models. In Proceedings of the 31st USENIX Security Symposium (USENIX Security 22), Boston, MA, USA, 10–12 August 2022; pp. 4525–4542.
20. Melis, L.; Song, C.; De Cristofaro, E.; Shmatikov, V. Exploiting unintended feature leakage in collaborative learning. In Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 23 May 2019; pp. 691–706.
21. Nasr, M.; Shokri, R.; Houmansadr, A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 23 May 2019; pp. 739–753.
22. Ling, X.; Ji, S.; Zou, J.; Wang, J.; Wu, C.; Li, B.; Wang, T. Deepsec: A uniform platform for security analysis of deep learning model. In Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 23 May 2019; pp. 673–690.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.