Information Processing and Management 60 (2023) 103162

Contents lists available at ScienceDirect

Information Processing and Management

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

Seeing is believing: Towards interactive visual exploration of data privacy in federated learning

Yeting Guo [a], Fang Liu [b,*], Tongqing Zhou [a,*], Zhiping Cai [a], Nong Xiao [a,*]

[a] College of Computer, National University of Defense Technology, Changsha, Hunan, 410073, China
[b] School of Design, Hunan University, Changsha, Hunan, 410082, China

ARTICLE INFO

Keywords:
Federated learning
Visualization
Privacy protection

ABSTRACT

Federated learning (FL), as a popular distributed machine learning paradigm, has driven the integration of knowledge in ubiquitous data owners under one roof. Although designed for privacy-preservation by nature, the supposed well-sanitized parameters still convey sensitive information (e.g., reconstruction attack), while existing technical countermeasures provide weak explainability for privacy understanding and protection practices of general users. This work investigates these privacy concerns with an exploratory study and elaborates on data owners' expectations in FL. Based on the analysis, we design the first interactive visualization system for FL privacy that supports intelligible privacy inspection and adjustment for data owners. Specifically, our proposal facilitates sample recommendation for joint privacy–performance training at cold start. Then it provides visual interpretation and attention rendering of privacy risks in view of multiple attacking channels and a holistic view. Further it supports interactive privacy enhancement involving both user initiative and differential privacy technique, and iterative trade-off with real-time inference accuracy estimation. We evaluate the effectiveness of the system and collect qualitative feedbacks from users. The results demonstrate that 96.7% of users acknowledge the benefits to privacy inspection and adjustment and 90.3% are willing to use our system. More importantly, 87.1% increase the willingness of contributing data for FL.

1. Introduction

Pervasive smart devices have facilitated the continually sensing and generation of data in our daily life. However, exposing these data compromises user privacy. To collaboratively and safely discover the knowledge in these ubiquitous data owners, federated learning (FL) comes to the spotlight recently as a distributed machine learning paradigm (Guo et al., 2021; Wu, Deng, & Li, 2022). As shown in Fig. 1, FL allows data owners (a.k.a., clients) to keep their sensitive data at local and share their trained machine learning models to a centralized server for aggregating a global model. It is thus supposed to break the data silos in a privacy-preserving way. We have already seen some pilot applications of FL, such as COVID-19 diagnosis (Dayan et al., 2021) and autonomous vehicles (Lim et al., 2021).

Interestingly, although proposed for privacy by nature, FL is still criticized for serious privacy vulnerabilities. Given that the reported parameters are representation for the features of local data, an adversary can infer the property (e.g., are there images for specific disease in a client) and even conditionally reconstruct the trained images (Zhu, Liu, & Han, 2019) from the periodically

* Corresponding authors.
E-mail addresses: guoyeting13@nudt.edu.cn (Y. Guo), fangl@hnu.edu.cn (F. Liu), zhoutongqing@nudt.edu.cn (T. Zhou), zpcai@nudt.edu.cn (Z. Cai), nongxiao@nudt.edu.cn (N. Xiao).

https://doi.org/10.1016/j.ipm.2022.103162
Received 28 June 2022; Received in revised form 31 October 2022; Accepted 7 November 2022
Available online 24 November 2022

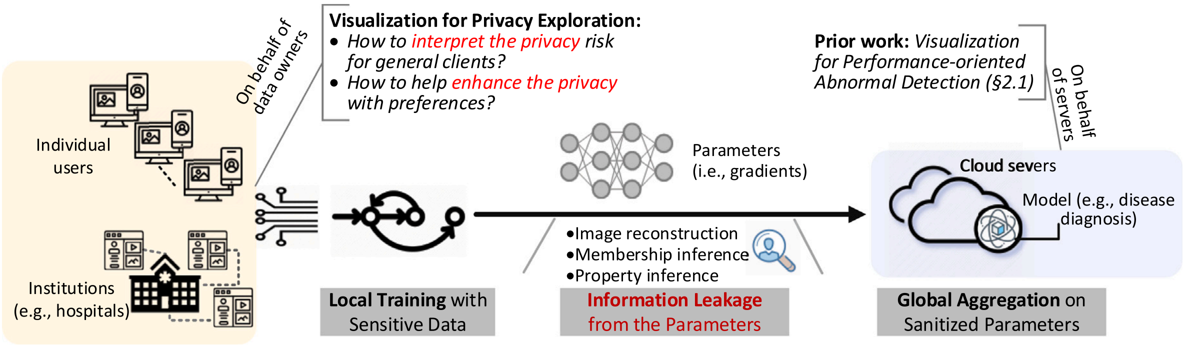

Visualization for Privacy Exploration:
• How to interpret the privacy risk for general clients?
• How to help enhance the privacy with preferences?
On behalf of data owners
Prior work: Visualization for Performance-oriented Abnormal Detection (§2.1)
On behalf of servers
Individual users
Institutions (e.g., hospitals)
Parameters (i.e., gradients)
Cloud severs
Model (e.g., disease diagnosis)
•Image reconstruction
•Membership inference
•Property inference
Local Training with Sensitive Data
Information Leakage from the Parameters
Global Aggregation on Sanitized Parameters

Fig. 1. A typical FL process with data owners (also called clients, either individuals or institutions) performing local training on their data and servers aggregating the distributed knowledge (i.e., parameters). Wherein, the supposed sanitized parameters still convey sensitive information (i.e. reconstruction attack, membership and property inference attack). This work explores the benefit of interactive visualization to mitigate these concerns by better interpreting and adjusting the privacy situation for general data owners.

released parameters. Reconstructed face images may be used for malicious face change, whereas the leaked personal medical information from institution would lead to the harassment of patients by drug advertisements (Verizone, 2021). Yet, **since FL itself is a 'black-box' for general data owners, directly implementing privacy enhancement techniques** (**e.g., differential privacy** Abadi et al., 2016) **can yield weak explainability and limited confidence on their situations**. In particular, different training samples will usually exhibit and experience different privacy leakage potentials and risks, then how can the data owners interpret the privacy contexts from the complex parameters? The sensitivity on privacy of different samples the utility expectation (e.g., incentives, model performance) vary widely among data owners, then how to guide them to accommodate the privacy adjustment on their own preferences? Thus, **there is an urgent need for an interactive method for data owners to explore privacy**.

In order to mitigate these concerns, we propose to leverage interactive visualization for intelligible privacy exploration on FL. We note that there are efforts devoted to the technical combination of visualization and FL, but they are designed from the perspective of cloud servers to identify abnormal local parameters for performance considerations (Li, Wei et al., 2021; Wei et al., 2019). As depicted in Fig. 1, we attempt to investigate the expectation and benefit of visually privacy interpretation and enhancement on behalf of the general FL clients, making it the first work of this line.

Since there lacks the investigation on people's perception of FL privacy in practice, we first conduct an exploratory study to figure out the major concerns and expectations. Specifically, we use qualitative survey to understand the awareness of general clients on FL parameter privacy, their practice routine of training data selection, their sensitivity on privacy threats in FL, the general countermeasures for potential threats, and ways to handle privacy and accuracy trade-off. Given the collected results, we elaborate on the following highlighted findings in terms of practical challenges:

- **Coarse-grained training data selection:** Different data samples usually own different (sometimes overlapped) contributions and privacy sensitivity for training, which makes it difficult for general clients to select proper samples from the cold training start. A random selection can hardly guarantee desired performance and privacy.
- **Limited knowledge for risk inspection:** Even aware of and worried about possible information disclosure from attacks, FL clients generally have no idea on how to manually build the attack models from the scratch and assess the complex metrics.
- **'All or none' protection intervention:** FL usually requires clients to upload local updates within a given time, so FL clients have limited time to make privacy-preserving countermeasures against attacks. This leads clients to tend towards strict privacy interventions for all samples, or even outright refusal to participate. The performance degradation of such choices is seldom investigated and evaluated for them to make proper trade-off.

Based on the above findings, we first propose a sample recommendation mechanism for joint privacy–performance training at cold start (Challenge #1). The system shuffles the local samples and performs individual-wise gradient contribution in terms of the downloaded global model to highlight a small (for privacy) yet proper training subset. Further, the system builds three main attacks on FL privacy at the back-end and dynamically simulates the attack processes on the local training results. Given the attacking simulation results, it provides a visual interface for the worth-attention images and regions, risk statistics, and holistic risk assessment of different attacks (Challenge #2). Finally, in view of the parsed risk contexts, dedicated privacy intervention suggestions on both sample removal and differential noise injection are then explicitly indicated to the users for prompt response. Meanwhile, we propose a model unlearning mechanism in avoidance to laboriously retrain the model after performing privacy interventions. It provides real-time accuracy influence feedbacks for each intervention behavior to interactively and seamlessly guide the user towards their preferred privacy and performance trade-off (Challenge #3).

Our contributions are summarized as follows:

- We provide an exploratory study on general data owners' privacy concerns, expectations, and requirements in FL practice. We believe the findings can provide insights to attain privacy explainability in FL, as well as other machine learning paradigms.

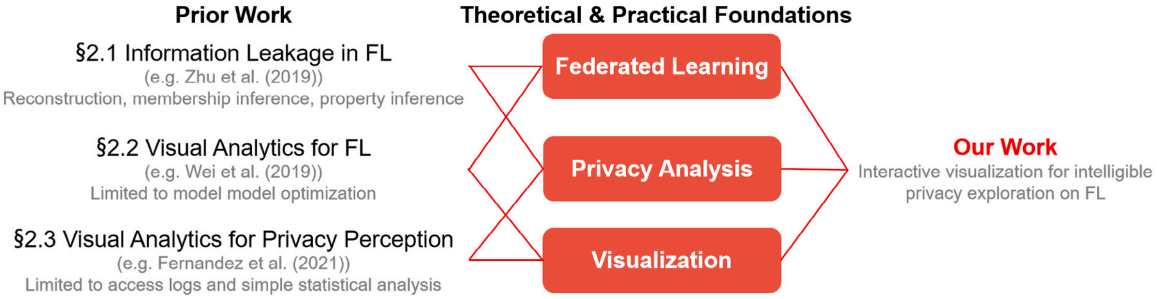

Prior Work
§2.1 Information Leakage in FL
(e.g. Zhu et al. (2019))
Reconstruction, membership inference, property inference
§2.2 Visual Analytics for FL
(e.g. Wei et al. (2019))
Limited to model model optimization
§2.3 Visual Analytics for Privacy Perception
(e.g. Fernandez et al. (2021))
Limited to access logs and simple statistical analysis
Theoretical & Practical Foundations
Federated Learning
Privacy Analysis
Visualization
Our Work
Interactive visualization for intelligible
privacy exploration on FL

Fig. 2. An overview of the related work.

- We develop a client-oriented interactive visualization system to intelligibly explore FL privacy. It facilitates joint privacy–performance training from the cold start, visually and comprehensively interpretation on privacy contexts, and fine-grained privacy enhancement with interactively accuracy calibration according to clients' own preferences.
- We evaluate the running performance of key components and qualitative property from real-world survey. Our findings show the great potential of endowing client-centric privacy explainability in helping FL practicers understand, gain trust, and increase participation in such paradigms.

The remaining part of the paper proceeds as follows. We review the background and related works in Section 2 and introduce our exploratory study and analysis in Section 3. Then we give an overview of our visualization system in Section 4, followed by the detailed description of its back-end engine and front-end visualization in Sections 5 and 6, respectively. Finally, we evaluate and discuss the system in Sections 7 and 8, and conclude in Section 9.

2. Related work

In this section, we review prior works on relevant areas of privacy issues on FL, visual analytics for FL, and visual analytics for privacy perception. The theoretical & practical foundations mainly relate to FL, privacy analysis and visualization techniques. Fig. 2 provides an overview of the related work and illustrates our relation to them.

2.1. Information leakage in FL

As Gartner predicted, in the cycle from 2021 to 2025, FL will play a mainstream role and guide the tide of commercialization in privacy computation (Willemsen, 2021). Although FL does not directly expose user data to third parties, it is still feasible for attackers to infer some information through the analysis of the model parameters trained by user data (Banabilah, Aloqaily, Alsayed, Malik, & Jararweh, 2022). We divide the main attack on FL into the following three categories:

2.1.1. Reconstruction

Zhu et al. (2019) argue that the sharing of the gradients can leak private training data. They introduce their attack named DLG and validate its effectiveness on reconstructing different types of training data (*i.e.* images and texts). Followed by this work, Zhao, Mopuri, and Bilen (2020) propose an improved attack called iDLG to get rid of the slow model convergence and inaccurate label inference in DLG. Geiping, Bauermeister, Dröge, and Moeller (2020) propose a general optimization-based attack based on cosine similarity of gradients, and demonstrate that their attack could reconstruct training images even in a more realistic context, that is, multiple images at high resolution are trained in a complex machine learning model. Ren, Deng, and Xie (2022) improve the image reconstruction attack with generative regression neural network. The reconstruction shows better stability, stronger robustness, and higher accuracy. Dahlgaard, rgensen, Fuglsang, and Nassar (2022) demonstrate that the attack configurations significantly impact the quality of medical image reconstruction. And they argue that the optimal configuration largely depends on the target image distribution and the complexity of the model architecture.

2.1.2. Membership inference

Bu, Wang, and Tang (2021) and Erlich and Narayanan (2014) report on the dangers of Membership Inference Attack (MIA) in human genetic data. From the group's shared summary statistics, they accurately infer whether the target human subject belongs to a specific genetic case group. Parameters in machine learning models also leak membership information. Miao et al. perform MIA on automatic speech recognition models to audit the use of users' audio data (Miao et al., 2021). Zhang, Ren et al. (2021) discuss that MIA can infer the sensitive information of individuals in recommendation systems, such as shopping preference and geographical positions. Attackers eavesdrop on the model parameters exchanged between the FL clients and the global FL server to conduct MIA. Chen et al. (2022) devise such attack against collaborative inference in industrial internet of things. Zhang, Zhang, Chen, and Yu (2020) enhance the accuracy of MIA in FL with the generative adversarial network. Gu, Bai, and Xu (2022) analyze different tendencies of prediction confidence between training and testing data in FL, and design a severe MIA based on predication confidence series.

2.1.3. Property inference

Different with membership inference which is usually sample-level, property inference aims to inferring whether the training dataset contains samples with the target property (Melis, Song, Cristofaro, & Shmatikov, 2019). Ganju, Wang, Yang, Gunter, and Borisov (2018) point that the data owner does not want the trained model to infer other properties than the expected property originally designed by the model. They take a classifier for recognizing smiling faces as an example and found that the classifier would leak the attractiveness information of the training clients. Zhang, Tople and Ohrimenko (2021) prove the effectiveness of their attack on population-level properties in different datasets and different machine learning models. Shen et al. (2021) design an active property inference attack. It accelerates the selection process for target FL clients and improves the attack accuracy in large-scale FL settings. Yang, Wang, and Li (2022) exploit the data leakage in the deep feature space and target at making property inference decisions on the level of individuals.

These attacks on FL are constantly evolving with higher success rate of the attack and stronger robustness. Researchers have carried out some privacy protection algorithms to mitigate the above attacks. Secure multi-party computation enables multiple FL clients to jointly calculate an aggregation function without disclosing their model parameters (Li, Zhou et al., 2021). It could greatly decrease the possibility of information leakage, but the complex computation protocols make it relatively inefficient. Some studies propose to compress the model or squeeze the trained parameters by controlling local training iterations (Wei et al., 2020). And some propose to perturb the trained parameters or the predictions of the local model with noise generated by differential privacy theory (Sun & Lyu, 2021; Sun, Qian, & Chen, 2021). Zhang, Chen, Hong, Wu, and Yi (2022) conduct the convergence analysis of FL algorithms subject to differential privacy and optimize these algorithms with refined clipping operations. It usually requires professional knowledge of privacy protection mechanisms to identify privacy risks and avoid the risk in FL.

2.2. Visual analytics for FL

Visualization leverages graphical methods to convey and communicate information clearly, quickly and effectively (Lim, Hibschman, Zhang, & O'Rourke, 2018; Shao, Yang, Juneja, & Seetharam, 2022). Previous studies have reported on some visualization tools for FL. We divide these tools into two categories. One is designed for the global FL server, that is, the cloud server who globally manages the machine learning model. The other is for FL clients, that is, users and institutions who own the data. We would introduce the two, respectively.

2.2.1. For the global FL server

Wei et al. (2019) simulate FL in a video game scene to observe multi-agent coordination. They develop three main visualizations to demonstrate the local view and local performance of each player, the overview of the federated aggregation model and an AI view to observe the AI-controlled cars, respectively. Wu et al. (2022) discuss the importance of anomaly detection in FL and propose FL-MGVN, which uses mixed Gaussian variational self-encoding network for anomaly detection. Li, Wei et al. (2021) provide comparative visual interpretation at the level of overview, communication rounds and client instance levels, so as to support FL experts to analyze the relevance of clients' information and identify potential anomalies. Meng et al. (2021) also focus on leveraging visualization methods to help experts diagnose abnormal clients and improve FL models. Sun et al. (2022) support the global FL server to visually monitor the training state and interactively adjust the model selection. The interaction helps the FL server to reach their expectations on the model training efficiency and model performance. Chen et al. (2020) propose a privacy-preserving mechanism for distributed visualization. Each client encodes the visual features obtained by their local data and sends these features to the server. The server would decode these distributed visual features and aggregate them to form the global visualization. These studies are mainly for the global FL server to inspect the training state of distributed clients and optimize the federated aggregation model.

2.2.2. For FL clients

Cloudera Fast Forward Labs proposed an interactive FL prototype, Turbofan Tycoon (Mike, 2018). It presents the advantages of FL to users (i.e. factory owners) by providing an intuitive view that they can most accurately predict the maintaining of turbofan engines when choosing FL strategies. For better inspection of FL process, FATE, the world's first industrial level technology framework for FL, launched their visualization tool FATEBoard (Tao, 2019). It provides users with rich visual panels to monitor all running job in the progress, track the whole process and log details of each algorithm component. The data heterogeneity among different FL clients hinders model training. Wang et al. (2022) proposed HetVis, a visual analytics tool to explore data heterogeneity. It identifies the data heterogeneity based on the different prediction behaviors of the global model and the local model. Then it generates a context-aware clustering of the inconsistent data samples as the summary of data heterogeneity. These studies visualize the training process to attract users to understand, participate and explore FL. However, they ignore that the privacy issue during this process also greatly affects the clients' attitude.

2.3. Visual analytics for privacy perception

Privacy is a subjective concept, and its perception is critical. Velykoivanenko et al. (2022) discuss the privacy perception in fitness trackers. Fitness trackers always collect various physiological data, from which machine learning algorithms can infer some non-physiological sensitive information. Although users concerned about information leakage, they lack its corresponding perception. Visualization is a powerful tool to help individuals perceive privacy leakage and fill the gap between privacy perception and protection actions (Muchagata, Vieira-Marques, & Ferreira, 2019). Fernandez, Nurmi, and Hui (2021) investigate users' privacy perception in smart device ecosystems and proposed to visualize the privacy assistance to promote users to understand the information leakage. Soumelidou and Tsohou (2020) leverage visual features instead of textual features in the privacy policy of Instagram. Dang, Dang, and Küng (2020) explore an interactive visual model for users to control the information leakage in online social network environment. Wilkinson et al. (2020) describe smartphone apps' data sharing activities with different visualization features. Users can easily glance which data their apps access and when and how frequent, and engage more in privacy management. Karegar and Fischer-Hübner (2021) create and analyze metaphors for differential privacy mechanism to users who should decide about sharing their data in the context of differential private data analysis. Nanayakkara, Bater, He, Hullman, and Rogers (2022) design an interactive visualization system to understand the privacy–utility trade-offs when applying differential privacy in data releases. We highlight these studies because they offer some important insights into the design of security explainability and the design of privacy-preserving strategies. But these studies mainly interpret the access log and some simple statistics of privacy data. The analysis of private data is evolving as machine learning comes into play. The visualization for privacy perception in machine learning field has been a largely under explored domain.

As far as we know, our proposal is the first piece of work that supports data owners to visually analyze privacy risks in FL practice. For the joint research of FL and privacy analysis, existing studies have revealed the potential information leakage in FL with various attacks. General users often lack the relevant expertise for privacy inspection, not to mention the targeted remediation according to their own needs. It motivates us to leverage visualization technique to facilitate users to explore the data privacy in FL. For the joint research of FL and visualization, existing studies have applied visualization in FL. They are mainly focused on improving model performance while generally ignoring the privacy issues in FL. For the joint research of visualization and privacy analysis, we note that some studies have also explored the visualization of data privacy. Nevertheless, they are mainly targeted for explaining the textual privacy policy and presenting the access information of private data and not applicable to the novel FL paradigm. In view of the technique gap in applying visualization to FL privacy, we first analyze the challenges and opportunities of doing so and then develop our interactive visualization system to intelligibly explore FL privacy.

3. Exploratory study on privacy inspection

Since there is not much prior work investigating the client perception on FL privacy, we conducted a user study to understand the main concerns and elaborate on their expectations in inspecting data privacy. Its primary goal was to use the results to inform the design of a privacy visualization system for FL.

3.1. Participant recruitment

We recruited 38 participants (31 men and 7 women) by sharing our questionnaire in FL Developer, AI Security and Healthcare Information Management Agency forums, and through contacting with personal contacts who show interests to this study. These participants ranged in age from 20 to 45 years and are from different educational backgrounds (i.e., 3 undergraduates, 21 with master's degree, and 14 with PhD degree). For analyzing the effect of user experience in machine learning and FL, we divided participants into two categories: (1) experts who know the basics of machine learning and FL; (2) beginners who do not have relevant knowledge. This gave us 21 experts and 17 beginners.

3.2. Procedure

We designed a questionnaire to analyze the requirements to explore data privacy in FL. And we published our questionnaire in the Tencent Questionnaire platform, and participants could scan the QR code or click the link for answering. For beginners with little experience in machine learning and FL, we briefly introduced the background and basic workflow of FL, and highlighted some applications of FL to make it more understandable.

3.3. Design of survey

The questionnaire consists of 17 items spread over 6 sections. The first two sections are designed to grasp the basic information of the participants and their general attitude towards FL. And then we specifically understand the privacy issues of participants in the FL process. We decoupled the issues into four stages. They are privacy data selection before training, information leakage inspection, privacy protection enhancement and the trade-off between privacy and model performance during training. We provide the script of the questionnaire in Appendix A. Here we mainly introduce the design ideas of each section.

Sec.1. This section is concerned about the demographic questions. Specifically, it collects information on age, gender, experience in FL and machine learning. This information is used to describe the characteristics of participants, so as to help us analyze the views of participants with different characteristics.

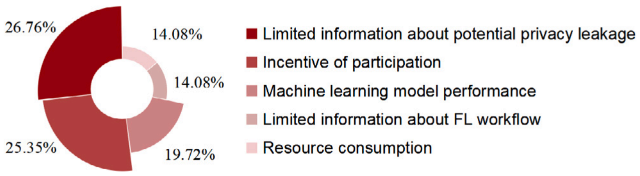


Fig. 3. The ranking of factors that affects FL clients' participation in FL.

Sec.2. This section discusses the awareness of general clients on FL parameters privacy. Since FL is known for protecting privacy through sharing model parameters, we investigate to what degree participants recognize its privacy protection, and what the main factors will affect their acceptance of FL. These factors are selected from the main bottlenecks restricting the development of FL proposed in the existing FL work.

Sec.3. This section discusses the practice routine of training data selection. When invited to participate in FL, FL clients need to select local data for training. Since the private data minimization rule has been highly expected to protect privacy, we observe the behaviors of participants to select training data, and analyze the bottlenecks in data selection.

Sec.4. This section discusses the clients' sensitivity to privacy threats in FL. We have previously introduced that attackers could infer various information from the trained model parameters. In this questionnaire design, we focus on the extent to which participants are worried about various types of information leakage and how they use their experience to estimate the probability of such information leakage. Each type of information leakage is introduced with a simple case description to make it more understandable to participants. We classify worry into four levels according to their severity, i.e. not worried, slightly worried, moderately worried and very worried. And we summarize several estimation methods through internal discussion with team members. These methods are provided to participants to select. This information would demonstrate the significance of our study.

Sec.5. This section discusses the general countermeasures for potential threats. We list several common protection methods in FL according to the discussion with team members. Then we ask participants what methods they might use in response to the leakage. Further, we investigate participants' views on the importance of various factors (e.g. diversity) about privacy enhancement through a set of 7-point Likert scale. This information is used to analyze preference and requirements when they engage in privacy enhancement.

Sec.6. This section discusses the trade-off between privacy protection and model performance in FL. It is known that taking privacy protection enhancement always sacrifices a certain model performance. There are various performance metrics, such as the inference accuracy and the loss value of the local model. We try to find which performance metrics participants care most and which kinds of interactive methods participants prefer when making a trade-off between privacy protection and model performance. The information would help us understand the preference and requirements in making the trade-off.

3.4. Findings

According to the collected questionnaire results, we summarize our findings as follows. They provide insights into the necessity of inspecting information leakage in FL and its requirement analysis.

3.4.1. The effect of privacy concerns on the acceptance of FL

When asked the attitude towards FL, 39.47% of participants acknowledged that FL promotes the privacy protection for them, 47.37% of them held the view that FL could only protect privacy in a certain degree, and 10.53% disagreed and 2.63% had no idea about it. We set a hypothetical scenario that the private data of clients would contribute to the diagnosis of a disease, and that they would be rewarded for contributing their data in the way of FL. We asked them to pick the three most important factors that affect them to participate. The results are indicated in Fig. 3. We found that the limit information about potential privacy leakage is the most important factor, followed by incentives of participation, machine learning model performance, resource consumption, limited information about FL workflow. In this paper, we focus on privacy issues in FL and aim to provide users with a visualization system to inspect information leakage and engage in privacy enhancement to increase their confidence and willingness in FL.

Remark 1. FL clients still have great concerns about the privacy protection of FL, which significantly affects their willingness to participate

3.4.2. The awareness of sensitive data selection

Different data contribute differently to the machine learning model. The incentive is determined based on the contribution. Although training with all local data can maximize the incentive, it will also increase the probability of information leakage. Since the contribution may be overlapped among data, majority of participants (89.47%) would select partial instead of all data samples in order to use the least data to obtain the maximum incentive. The remaining 10.53% of them train with all the data because the selection is time-consuming and the selection principle is complex for them to understand. We found that 52.63% of them do not know how to select the right data.

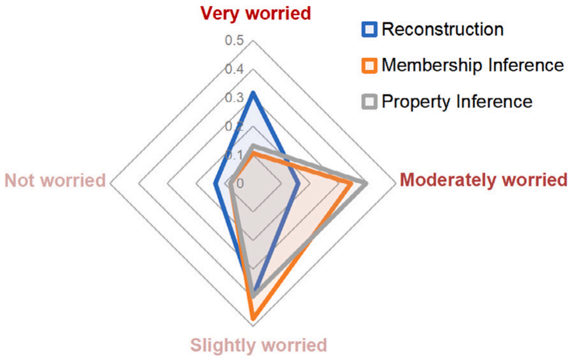


Fig. 4. The statistics on the level of worry about information leakage in FL.

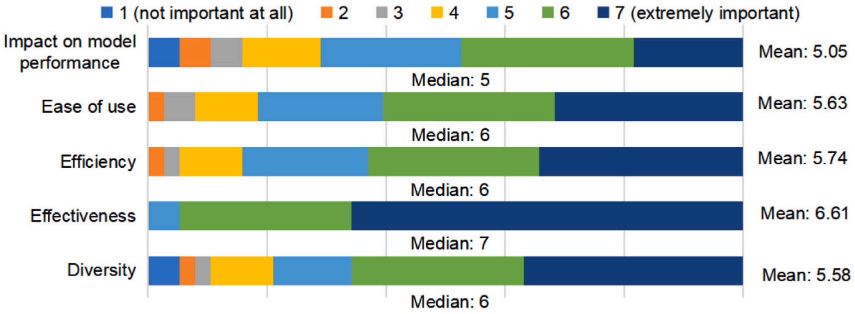


Fig. 5. Importance of various factors in privacy enhancement in FL.

Remark 2. Most data owners want to train models with the least amount of private data to get the most incentive, but they know little about how to make it

3.4.3. The concern about information leakage in FL

Fig. 4 demonstrates the extent to which our participants are concerned about information leakage in FL. It is found that participants were generally slightly or moderately worried about all these types of information leakage. The level of worry towards different types of information leakage is different. The most worrying attack is reconstruction, followed by property inference and membership inference. Besides, FL beginners are more worried about the information leakage than FL experts because they are not quite aware of the probability of leakage. When asked about how to measure such probability, 47.37% of participants would understand the principles of these attacks, and train their corresponding attack models to analyze the leakage probability. The measurement is accurate but very hard for FL beginners. 31.58% of them would study the related work, and refer to the most related work to estimate the probability. It is relatively inaccurate. And 21.05% of them have no idea about the measurement. It is necessary to provide FL clients with a simple and convenient way to browse the leaked information and the corresponding probability of leakage.

Remark 3. Attacks on FL have caused varying levels of concern. Lack of intuitive and reliable tools for FL clients to inspect data privacy in FL

3.4.4. The behavior analysis of privacy enhancement

We make statistics on privacy enhancement taken by participants. 36.84% of them would remove vulnerable data from the training dataset. 50.00% of them would apply some privacy preserving algorithms, like differential privacy mechanism. And 5.26% of them have no idea about how to protect and 7.89% of them even refuse to participate in FL for the sake of information leakage. The refuse may cause the model to miss learning some high-value data. Further, we investigated the importance of various factors in privacy enhancement in FL. The results are demonstrated in Fig. 5. Effectiveness is obviously the most important factor, followed by diversity, efficiency, ease of use and impact on model performance. As no previous work designed privacy enhancement interfaces for users in FL visualization systems, these results provide some insights to our design in helping users to enhance privacy protection.

Remark 4. FL clients tend to use privacy protection algorithms or remove vulnerable data to enhance protection Effectiveness, efficiency and ease of use are the main indicators to evaluate privacy protection methods.

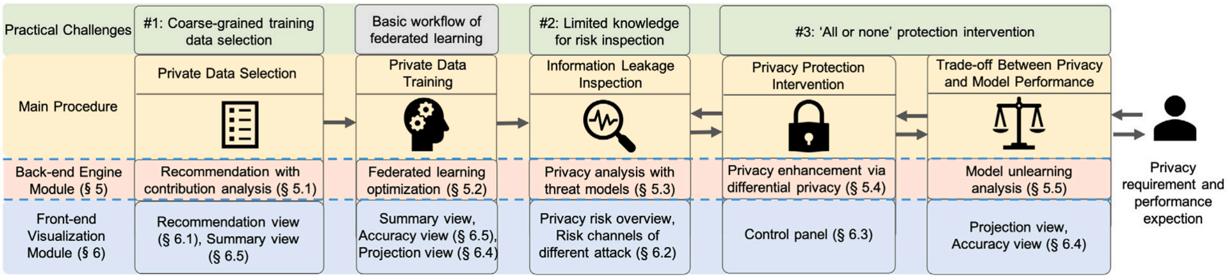

Practical Challenges
#1: Coarse-grained training data selection
Basic workflow of federated learning
#2: Limited knowledge for risk inspection
#3: 'All or none' protection intervention
Main Procedure
Private Data Selection
Private Data Training
Information Leakage Inspection
Privacy Protection Intervention
Trade-off Between Privacy and Model Performance
Back-end Engine Module (§ 5)
Recommendation with contribution analysis (§ 5.1)
Federated learning optimization (§ 5.2)
Privacy analysis with threat models (§ 5.3)
Privacy enhancement via differential privacy (§ 5.4)
Model unlearning analysis (§ 5.5)
Front-end Visualization Module (§ 6)
Recommendation view (§ 6.1), Summary view (§ 6.5)
Summary view, Accuracy view (§ 6.5), Projection view (§ 6.4)
Privacy risk overview, Risk channels of different attack (§ 6.2)
Control panel (§ 6.3)
Projection view, Accuracy view (§ 6.4)
Privacy requirement and performance expection

Fig. 6. An overview of our FL visualization system.

3.4.5. The behavior analysis when weighing privacy and model performance

Inference accuracy of model is the most concerned performance metric in the trade-off process. 36.84% of participants prefer to adjust the two through interaction and make the desired trade-offs by observing their changes. 31.58% of them want to output the probability of information leakage when the model performance reaches its default value. The remaining 31.58% prefer to output the model performance when the leakage probability reaches its upper limit.

Remark 5. FL clients prefer to weigh model performance and privacy in an interactive way, and the model performance metric they mainly pay attention to is model inference accuracy

In summary, we have observed user attitudes towards FL privacy, greedy learning intentions in FL, diverse concerns in practice, common privacy protection measures and human involvement expectations.

4. Approach overview

In view of the above requirement analysis, we designed a FL visualization system for FL clients to explore data privacy in FL. Fig. 6 illustrates the main pipeline of the system. The step division of the pipeline is basically consistent with the stage division in Section 3.3. Each stage can be reflected in a corresponding step in the pipeline. The step of private data training is inserted between private data selection and information leakage inspection to reflect the FL training process. Users often need to interact with the system several iterations to calibrate the protection strength and judge whether their personal privacy requirements and performance expectations are met. The system consists of a back-end engine module (cf. Section 5) and a front-end visualization module (cf. Section 6). The back-end engine module describes the algorithm designs of each step. The front-end visualization module presents the algorithms' results on the corresponding views.

During introducing the two modules, we use genetic disease diagnosis based on face images (Kumov & Samorodov, 2020) as a case study to make our system more understandable. Imagine that due to privacy and ethical considerations (Andalibi & Buss, 2020; Bethge et al., 2021), medical institutions[1] want to train their face images through FL to obtain a high-accuracy diagnosis model. And considering the attacks on FL, medical institutions need to inspect the privacy leakage of their disease diagnosis model and take actions to enhance privacy protection.

5. Back-end engine

In this section, we describe the design details of the components of the back-end engine module. Firstly, in view of the coarse-grained training data selection, we propose a recommendation mechanism (Section 5.1) to guide users to select a small (for privacy) yet proper training subset. Then the selected training data is trained to optimize the model (Section 5.2) according to FL paradigm. Since the trained model would leak sensitive information about the training data, we perform privacy analysis on the trained model with threat models (Section 5.3). Users could inspect privacy leakage from the analysis.

Finally, to meet the demands of privacy protection, we apply differential privacy mechanism to the trained model (Section 5.4) and support users to delete vulnerable training data. Strong privacy enhancement may result in poor model performance. Users may adjust the enhancement several times to obtain the desired trade-off. During this process, we propose a model unlearning analysis (Section 5.5) to avoid the model retraining after privacy enhancement.

[1] We use users and institutions interchangeably hereafter, as the latter makes the predominant user of the proposed tool.

5.1. Recommendation with contribution analysis

The lack of understanding of the contribution of local data to the machine learning model (i.e. the contribution of each face image to the genetic disease diagnosis model in our case study) always results in the blindness of selecting training data and the deviation from the private data minimization rule. In view of Remark 2 in Section 3.4.2, we introduce a recommendation mechanism to assist data selection.

We preliminarily estimate the contribution of each data to the model by Gradient Shapley value (Ghorbani & Zou, 2019). The estimation algorithm is described in Algorithm 1. In the process, we firstly initialize the contribution of each data to zero (line 1). Then we analyze the contribution of data to the model over several iterations (line 2–10). In each iteration, we shuffle the data randomly to avoid data order affecting the analysis (line 4). And we initialize the local model gradients W_i to global model gradients W_S and the accuracy v_0^t with W_S (line 5). For each individual data $d_{i,j}$, we analyze its impact on the local model gradients W_i by removing $d_{i,j}$ from the shuffled dataset π^t, then measure the model inference accuracy $v_{d_{i,j}}^t$ brought by the gradients, and finally obtain the contribution of data $r_{i,j}$ to the model according to the change of accuracy (line 6–10).

We rank the data from highest to lowest by contribution value and then recommend it to the user. Users could select training data from the recommendation list. For ease of description, the selected data of client i is denoted as D_i afterwards.

Algorithm 1: Estimation of data contribution

```
    Input: Loss function $\mathcal{L}$, learning rate $\eta$, local dataset $D_i = \{d_{i,1}, d_{i,2} \dots d_{i,|D_i|}\}$, accuracy evaluation function $V(w)$, parameters
           of the global model $W_S$
    Output: Data contribution $R_i = \{r_{i,1}, r_{i,2}, \dots r_{i,|D_i|}\}$
 1  Initialize $R_i = [0, 0, \dots 0]$, $t = 0$;
 2  while not reach convergence do
 3      $t = t + 1$;
 4      $\pi^t \leftarrow$ random arrangement of $D_i$;
 5      $v_0^t \leftarrow V(W_s)$, $v^{t'} \leftarrow v_0^t$, $W_i \leftarrow W_s$;
 6      for $d_{i,j} \in D_i$ do
 7          $W_i \leftarrow W_i - \eta \nabla_{W_i} \mathcal{L}_{W_i}(\pi^t[d_{i,j}])$;
 8          $v_{d_{i,j}}^t \leftarrow V(W_i)$;
 9          $r_{i,j} \leftarrow \frac{t-1}{t} r_{i,j} + \frac{1}{t}(v_{d_{i,j}}^t - v^{t'})$;
10          $v^{t'} \leftarrow v_{d_{i,j}}^t$;
```

5.2. FL optimization

The selected data D_i is used to optimize the machine learning model in the FL paradigm. The workflow of FL can be decoupled into the following steps. These steps are repeated until the model converges.

Step 1. The FL client i downloads the global model with parameters W_s from the FL server. And then the client i initializes the local model based on the global model. The local model parameters W_i is initialized to W_s.

Step 2. The FL client i selects the local dataset D_i as training data to fine tune W_i. W_i is iteratively updated according to gradient descent algorithm, as illustrated in Eq. (1). $\mathcal{L}$ is the loss function, and η is the learning rate. After a certain iterations, the client i sends W_i to the global FL server.

$$W_i = W_i - \eta \nabla_{W_i} \mathcal{L}_{W_i}(D_i) \tag{1}$$

Step 3. The global FL server aggregates distributed model parameters $\{W_1, W_2 \dots W_N\}$ to update the global W_s. N is the number of FL clients. We take FedAvg (Bonawitz et al., 2019) as an example of FL aggregation algorithm. It is one of the most classic aggregation algorithm and demonstrated in Eq. (2). $|.|$ is the size of dataset, and D is the collection of $\{D_1, D_2 \dots D_N\}$. After aggregation, the global FL server distributes the newest W_s to FL clients for further model optimization.

$$W_s = \sum_{i=1}^{N} \frac{|D_i|}{|D|} W_i \tag{2}$$

5.3. Privacy risk analysis

The local model parameters W_i can still leak various information. Considering Remark 3 in Section 3.4.3, before sending W_i to the FL server (step 3 in Section 5.2), we would analyze the probability of information leakage through some attack models in the back end. Here we also call such probability privacy risk score in this paper.

In the following, we would introduce the basic principles of each attack model corresponding to the attack types we introduced earlier. Note that our scheme is mainly reflected in the design logic and visualization, rather than the improvement of specific FL attack algorithms. Attack models are constantly evolving, and these attack models can be embedded into our system according to the requirements of FL clients with little effort.

5.3.1. Reconstruction

The attacker attempts to reconstruct the original local dataset D_i. We denote the reconstructed data as D_i'. In order to make the reconstructed D_i' as similar as the original D_i as possible, the basic idea of this attack is to minimize the gap between the original gradients $\bigtriangledown_{W_i}\mathcal{L}_{W_i}(D_i)$ and the gradients $\bigtriangledown_{W_i}\mathcal{L}_{W_i}(D_i')$ generated by D_i'. The optimization objective is shown as Eq. (3). We use a L-BFGS solver (Liu & Nocedal, 1989) to solve this optimization problem.

$$arg\min_{D_i'} \|\bigtriangledown_{W_i}\mathcal{L}_{W_i}(D_i) - \bigtriangledown_{W_i}\mathcal{L}_{W_i}(D_i')\| \tag{3}$$

5.3.2. Membership inference

The attacker observes that machine learning models often behave differently on the data they train than the data they have not trained (Chen et al., 2022). The attack model recognizes such differences and then infers whether the target data is the membership in the local dataset.

In this attack, we firstly train a shadow model for each category of labels (e.g. genetic disease Chronic Myeloid Leukemia (CML) and Burkittlymphoma (BUR)). The shadow models are similar to the target model (i.e. the disease diagnosis model in the medical institution in our case study) in input, output and model structure. We use the model-based synthesis method to generate the training set and test set of shadow model. Then we train the attack model with these shadow models. Specifically, the prediction of shadow models is the input of the attack model, and the information of whether the data is in the training set (e.g. Alice's face image *in* the set and Bob's image *out* the set) is the label of the data in the attack model. At last, we regard the original target model in the FL client as a black box, and obtain the prediction of the target model to our target data. The attack model can judge whether the target data has been trained or not based on its prediction.

5.3.3. Property inference

The attack principle of property inference is very close to that of membership inference. The only difference is reflected in the classification of attack model is no longer whether the target data is in the training set, but whether the training set includes the target property (e.g. images of CML *in* the set and images of BUR *out* the set).

5.4. Privacy enhancement based on differential privacy

As indicated in Remark 4 in Section 3.4.4, in response to the above attacks, we suggest providing users with interfaces to call the differential privacy mechanism or delete vulnerable data to strengthen privacy protection. Differential Privacy (DP) is a privacy protection algorithm that aims to maximize the availability of data analytics while minimizing the risk of identifying one single record from the dataset (Abadi et al., 2016). Specifically, it first defines the adjacent dataset, which refers to two data subsets that differ by only one data sample. And then it measures the probability of adjacent data sets obtaining the same output. In our case study, it refers to the probability of obtaining the same model parameters with and without a single patient's face image in the dataset. The greater this probability, the smaller the probability of identifying a single sample. In mathematical form, a function M, whose domain and range are $\mathbb{D}$ and $\mathbb{R}$, satisfies (ϵ, δ)-DP if it holds Eq. (4) for any pair of adjacent datasets $\langle d, d'\rangle$. s is a subset of $\mathbb{R}$. ϵ indicates the privacy loss of DP, also called privacy budget, and δ indicates the probability that original ϵ-DP is broken.

$$Pr\{M(d) \in s\} \leq e^{\epsilon} Pr\{M(d') \in s\} + \delta \tag{4}$$

Gaussian mechanism is one of the mechanisms to realize (ϵ, δ)-DP. It is proposed to build a new function G on the basis of the original function M. Function G is formulated in Eq. (5). S_f is called sensitivity, which refers to the maximum difference between query results in Function M on adjacent data sets. σ is the noise intensity. $N(0, S_f^2 \times \sigma^2)$ is a noise generated by a normalized distribution with mean of 0 and standard deviation of $S_f\sigma$.

$$G(d) = M(d) + N(0, S_f^2 \times \sigma^2) \tag{5}$$

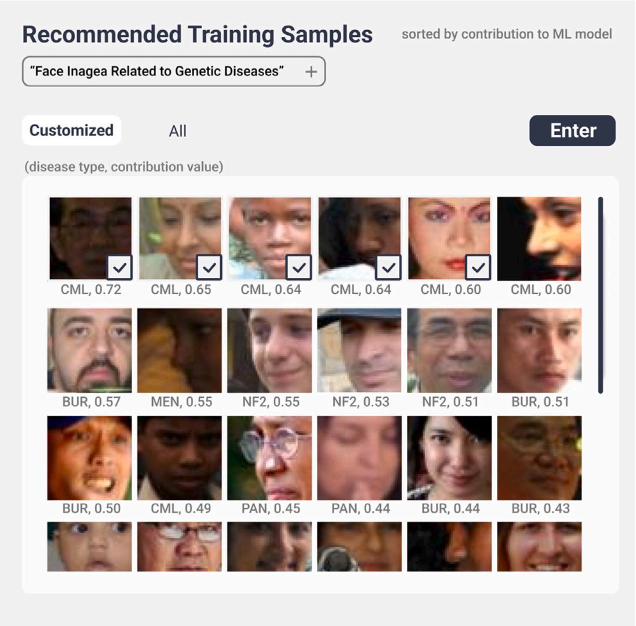
We apply Gaussian mechanism to interfere with local model parameters W_i in the system, so that the attacker can obtain very limited information from the interfered parameters, and reduce the sacrifice of the model performance as much as possible. In particular, we calculate the L2-normalization of each W_i and then scale it to ζ. In the way, ζ meets the above definition of sensitivity. Then we build a new interfered parameters W_i' as follows. The FL clients would send W_i' to the FL server instead of W_i.

$$W_i' = W_i + N(0, \zeta^2 \times \sigma^2) \tag{6}$$

5.5. Model unlearning analysis

We support users to delete vulnerable data for privacy protection, such as face images that are easily reconstructed. It needs to remove the memory of these deleted data from the trained machine learning model. The most straightforward way is to retrain a model based on the re selected data. However, it is very time-consuming, especially when the training data can be interactively modified many times (cf. Remark 5 in Section 3.4.5). Since efficiency is important for users to take privacy protection, we adopt a model unlearning method to avoid retraining (Bourtoule et al., 2021). It makes some modifications to Step 2 in Section 5.2.

In particular, we divide the local dataset D_i into several shards, and train a model for each shard, aggregate these models as the local model, rather than directly train the local model with all D_i at once. During training each shard, we further divide each



"Face Inagea Related to Genetic Diseases" +

Fig. 7. Recommendation view for data selection.

shard into several slides and perform incremental training in these slides. When data needs to be unlearned, we just need to remove the model containing the data from the aggregated local model and retrain the removed model. And in the retraining, we can start from the slide containing the data to be unlearned rather than a fresh start. The amount of calculation of retraining can be greatly reduced when the training dataset changes.

6. Front-end visualization

In this section, we present the views in the front-end visualization module and the supported interaction behaviors.[2] These views are assigned into two pages. The first page is the recommendation view (Fig. 7), which would recommend data selection to the data owner to help them realize private data minimization. The second page is the main page, which supports to interactively inspect and control private information leakage.

6.1. Recommendation view for data selection

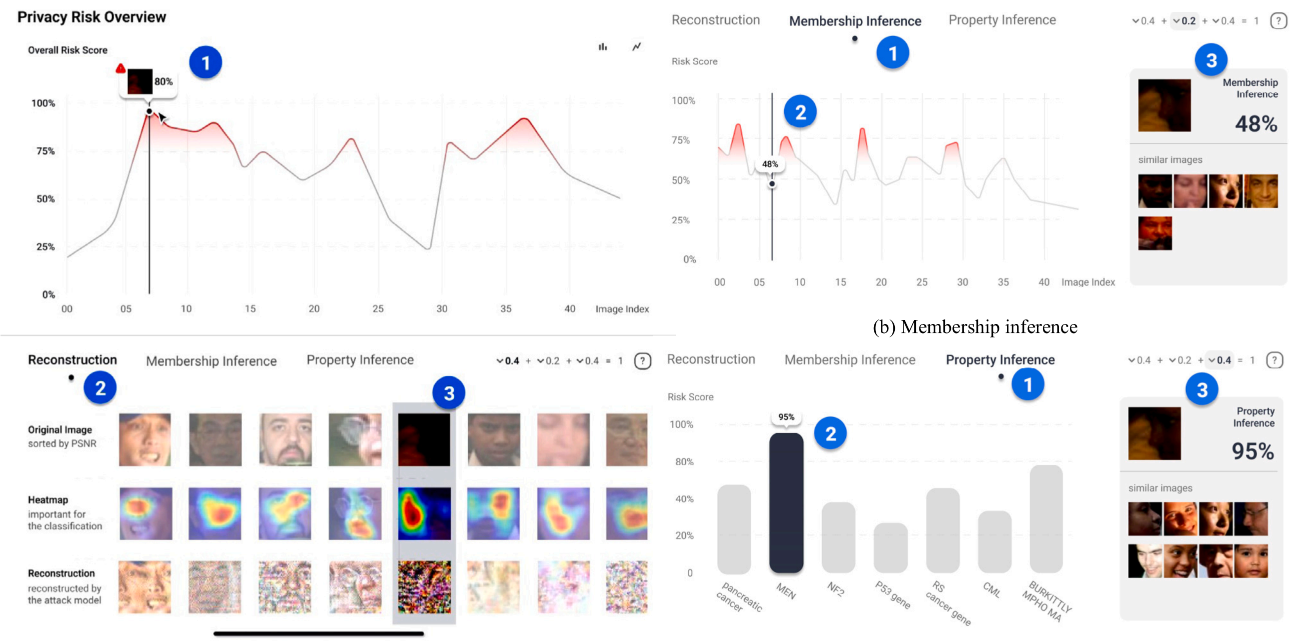
We use the method introduced in Section 5.1 to evaluate the contribution of each face image and rank them from high to low. The sequence below each image represents the genetic disease of the patient, e.g., CML and BUR, and its individual contribution value to model training. This information could help users roughly understand which type of data may be needed by the model. They can click the items in the recommendation list to customize their own training dataset, or click the 'all' button to select all images for training. Then they can click the 'Enter' to enter the next page.

6.2. Privacy risk views for inspecting information leakage

The selected data are fed to train the FL model (cf. Section 5.2). Some sensitive information could be inferred from the trained model. Considering the limited knowledge for risk inspection (cf. Remark 3 in Section 3.4.3), we provide visually interpretation and attention rendering of privacy risks in view of multiple attacking channels (Fig. 8). These channels are folded in the main page (Fig. 9(E)). Besides, we provide a holistic risk assessment of these attacks in Fig. 9(D).

'Reconstruction' channel (Fig. 8(a)): Our purpose is to exhibit the information leakage in reconstruction attack. This channel has three lines of images, showing the original image, the heat map of the image and the reconstructed image respectively. The

[2] Since the real face samples shall not be disclosed considering the patient's privacy, we use some public face images as toy examples in the following illustration.

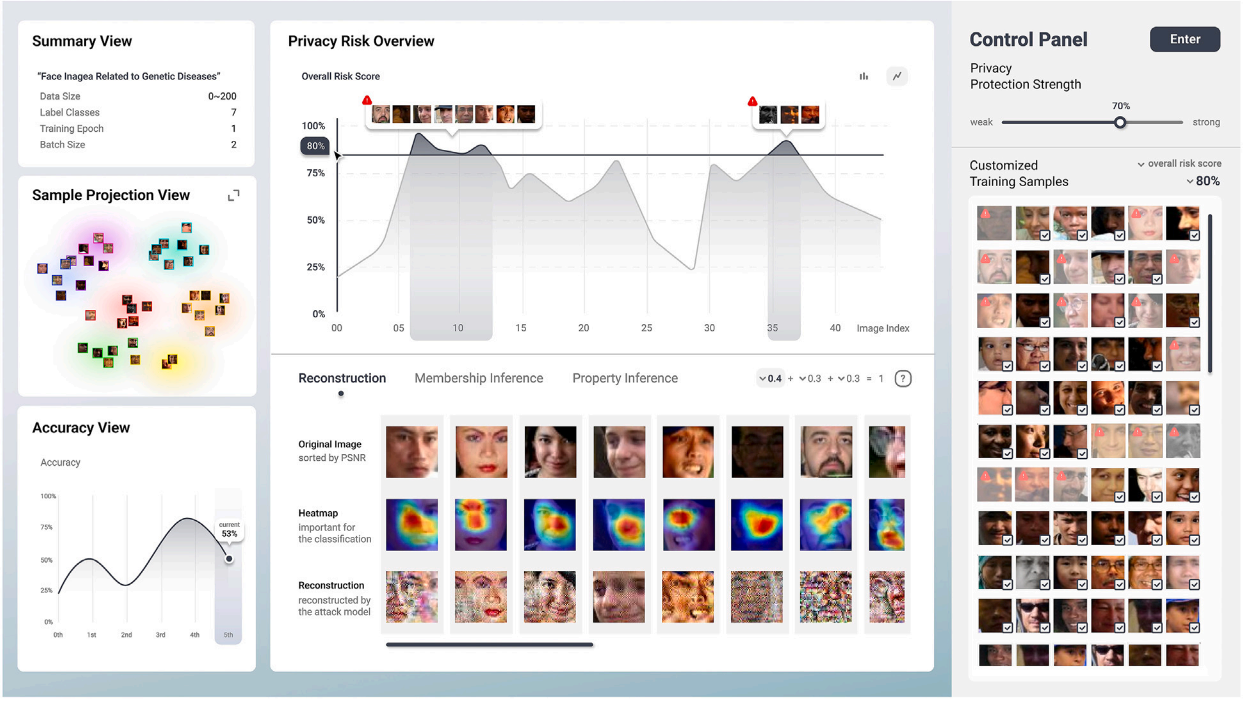


(b) Membership inference

(a) Reconstruction

(c) Property inference

Fig. 8. Privacy risk channels for each kind of attack (As shown in (a), users first (1) click a specific point in the risk overview. The corresponding image and score are shown. Then they (2) click an alternative risk channel. Details of information leakage about the image under this attack are shown in (3). As shown in (b) and (c), users could also first (1) click a channel, and then (2) click a point or pillar. The details of the corresponding image would be shown in (3).).



Summary View

"Face Inagea Related to Genetic Diseases"

Data Size	0~200
Label Classes	7
Training Epoch	1
Batch Size	2

Control Panel

Privacy Protection Strength: 70%

Customized Training Samples: overall risk score 80%

Fig. 9. (A): The summary view shows the basic information about training data and training model. (B): The sample projection view reflects the current classification of this model. Different colors represent different data labels. The distance between the images indicates their similarity. (C): It presents the accuracy to participate in FL in each round. (D): The privacy risk overview demonstrates the holistic risk assessment of three attacks for each data. (E): This view is composed of three channels that correspond to the three attacks described above. These channels demonstrate the information leakage situations faced by each attack mode. (F): This control panel supports to modify the training samples and disturb model parameters with differential privacy. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

heat map is generated by Grad-CAM (Gildenblat & contributors, 2021). It can highlight the important areas in the image for target prediction and classification in genetic diseases, and also imply the areas in the image that are easier to be reconstructed in this attack. The reconstructed image is obtained by the reconstruction attack model in Section 5.3.1. In order to facilitate users' browsing, we calculate the Peak Signal to Noise Ratio (PSNR) between the reconstructed image and the original image, and rank these images based on PSNR. PSNR is a performance metric to evaluate the similarity between images, which also represents the risk score of data reconstruction.

'Membership inference' channel (Fig. 8(b)): We present the risk score of each face image in membership inference attack. Here the risk score refers to the probability that the attacker correctly infers whether the patient's image is used for training diagnostic model of genetic diseases. We use a line chart to reflect this score. When the user clicks any point on the polygonal, the corresponding image and its value would appear on the right in the channel. Besides, the images whose risk scores are the same as the clicked image could also appear to provide a warning to the user. The user can also click a value (e.g. 75%) on the vertical axis. This operation is also demonstrated in Fig. 9(D). In that way, the images with the risk score higher than the value would be displayed. When the user clicks one of them, the image information would be shown on the right.

'Property inference' channel (Fig. 8(c)): The histogram shows the risk score of each property in property inference attack, which indicates the probability that the attacker identifies whether the target property is contained in the training dataset. When clicking one pillar of the histogram (e.g. MEN), images with the corresponding property would be presented. One of the images is represented at the top and the rest is placed below.

Privacy risk overview (Fig. 9(D)): While these attacks have been studied in depth, they are analyzed in isolation. It lacks a comprehensive view of these privacy risks. In view of this, we fill this gap by presenting a privacy risk overview in our visualization, as shown in Fig. 9(D). It illustrates the overall risk score of each image, which is a weighted sum of the risk scores in the above three attacks. As we investigated before, users worry differently towards different attacks. The user could set the weights of each attack $\{w_r, w_m, w_p\}$ based on their own privacy requirements. The sum of the weights is equal to 1. For example, in Fig. 9(E), they are set to 0.4, 0.2 and 0.2, respectively. Assuming the risk scores of a face image facing these three kinds of attacks are s_r, s_m and s_p respectively, the overall risk score of the image is $w_r s_r + w_m s_m + w_p s_p$. We show the score in a line chart. The user can click a point in the polygonal line or in the longitudinal axis, which is similar to that of the 'membership inference' channel. When a click happens, the corresponding images would appear above to indicate high risks. When a specific image is clicked, each risk channel will synchronously locate the image, as shown in Fig. 8.

6.3. Interactive control panel for privacy-preserving actions

Given the rendered privacy risks, corresponding protection measures are further required for the users to perform necessary intervention (cf. Remark 4 in Section 3.4.4). Considering the subjectivity of privacy, this function is accomplished by combining human preference with automated privacy protection techniques to agilely mitigate the privacy issues, instead of rigidly imposing the same stringent privacy intervention on all images. Specifically, we design an interactive control panel with two types of interfaces, as shown in Fig. 9(F).

Adjustable protection strength: Since general users may lack background knowledge of DP or other privacy-preserving algorithms, we discard the use of technical terms about DP and instead use easy-to-understand adjectives (i.e. weak and strong) to describe the DP noise. The user could intuitively adjust the sliding key to call the DP-based algorithm in Section 5.3 and control the level of DP noise.

Customized training samples: By loading the user selected samples from the cached local face images, a dedicated model is trained and its privacy risk is analyzed. Based on the perceived privacy leakage statistics, a user can set a threshold to filter vulnerable images. For example, in Fig. 9(F), the images whose overall risk score is higher than 80% could be blurred, and are marked with red warnings during sample customization. Meanwhile, the user can also increase or decrease training samples according to their preference/experience in this phase. For example, even the picture in the upper left corner of the control panel is rendered with a warning, the user may find it a scarce sample for predicting specific disease, s/he can still reserve it as a training sample regardless of the warning.

Iterative protection calibration: Privacy protection control is difficult to put in place in one step. Too strict privacy protection deteriorates the effectiveness of the model, while weak privacy control cannot meet the privacy needs. Users often need to interact with the system several iterations to calibrate the protection strength and judge the performance feedbacks are satisfactory, towards a personally ideal trade-off between privacy protection and model performance.

6.4. Projection view and accuracy view for privacy–accuracy trade-off

To help find the trade-off (cf. Remark 5 in Section 3.4.5), we provide an intuitive view for FL clients to observe the fluctuations in model performance when taking extra privacy-preserving actions. In the prior investigation and analysis, we have found that the most concerned model performance metric when making the trade-off is inference accuracy. Thus, we provide the accuracy view shown in Fig. 9(C) to promote the trade-off process. In this view, the user could view the model inference accuracy obtained during each round of FL participation. And the last marked point symbolizes the current round, and its value represents the accuracy obtained under the current settings.

Besides, we provide the sample projection view in Fig. 9(B) to demonstrate the current training state. We adopt t-SNE (Dimitriadis, Neto, & Kampff, 2018) to project these face images from the high-dimensional machine learning parameters to two-dimensional plane coordinates. Face images with the same genetic disease are marked in the same color. The ideal training state is that images with the same color form a separate cluster. However, it also means that it is very likely to leak the membership and property information. This projection could help the user understand the conflict between model performance and privacy.

6.5. Auxiliary information and interactions among the views

The basic auxiliary information about the training data and the model is presented in Fig. 9(A). It contains the title, size and label information of the dataset, the epoch and batch size of the model.

During the participation in FL, the main objective is to obtain an accurate model while guaranteeing privacy. The visual exploration in our system mainly consists of visual privacy interpretation and enhancement. Specifically, users could achieve their FL objectives through the visual exploration as follows.

In the recommendation view, the user could select some data with high contribution value for training a good model in FL. The selected training data is checked by default in the control panel. Its size and other basic information are presented in the summary view. The back-end would use the data for training and perform privacy analysis on the trained model. The privacy risks faced by the data are shown in the privacy risk overview. And the current training state of the data is reflected in the sample projection view and the accuracy view.

If the above objective is not met, the users could take some actions in the control panel. If some vulnerable data is deleted, it would perform the model unlearning analysis in Section 5.5. If the dataset is supplemented, it just performs incremental learning on the basis of the original model. Privacy enhancement with DP does not require the additional training on the dataset, only disturbing the original model. These actions would trigger the main page refresh. The size of the new selected data would be synchronized in the summary view. Based on the new model, we would re analyze the privacy risk scores and inference accuracy, and synchronize these analytical results in the related views. In this way, the user could judge whether the privacy enhancement is appropriate to meet the FL objective, and adjust the enhancement if it does not.

7. Evaluation

In this section, we evaluate our systems in terms of both real-time and quality of interactions.

7.1. Real-time interaction analysis

To evaluate the running performance of the system, we measure the response time for utilizing different components in the system. Since the delay of rendering the results of the back-end engine module to the front-end visual page is almost negligible in our system, the response time is nearly equal to the running time of the back-end engine module.

7.1.1. Settings

We implement our back-end engine in a laptop with 4 Intel Core i7@1.5 GHz and 16 GB of RAM. Data is scattered across virtual institutions with only a small amount of data per institution and therefore FL is required. Specifically, we assume each institution has 40 face images and trains them on the global model of Resnet-32-10 (Lu, Jiang, & Kot, 2018).

7.1.2. Results

The time costs mainly are taken in leakage inspection, privacy enhancements, and the trade-off phase. Whereas the first part is generally performed at the back-end for offline analysis, the last two parts involve real-time interactions, thus are more sensitive to the users.

(1) To inspect the information leakage, the three types of attack models run concurrently in the back end. Wherein, the running time for membership inference attack and property inference attack is **7.53 s** and **8.21 s**, respectively; For reconstruction attack, we note that it is essentially an optimization problem and requires constant iterations to approach the optimal solution. Empirically, we set the upper time bound for solving this problem to be **15 min**,[3] in order to guarantee sufficient awareness of possible information leakage on reconstruction attacks and ensure the FL efficiency. Hence, the simulation and risk perception on reconstruction attacks constitute the bottleneck here. In practice, the institutions can adjust this setting according to the sensitivity about reconstruction attacks and its available computation resources.

(2) For privacy enhancement, it takes an average of **1.41 ms** to disturb the model parameters with DP noise.

(3) For the privacy–utility trade-off module, the major time occupation is devoted to model unlearning analysis, whose latency is affected by the number of images to be deleted and the distribution of these images. From the average testing performance, the unlearning time is around **4.0 s** when randomly deleting an image from the training data.

In summary, its interactive performance, reflected on privacy intervention and adjustment, is generally in real-time. As such, we claim that the response time of the system is basically acceptable to clients, especially data custodians like institutions, when exploring data privacy.

7.2. Qualitative analysis on interaction quality

In terms of the interaction quality, we conducted a survey to investigate client feedback on the system.

[3] This is the suggested setting for common end devices with relatively weak computation capacity. An institution can certainly use a much smaller time budget as owning stronger computation resources (e.g., data centers).

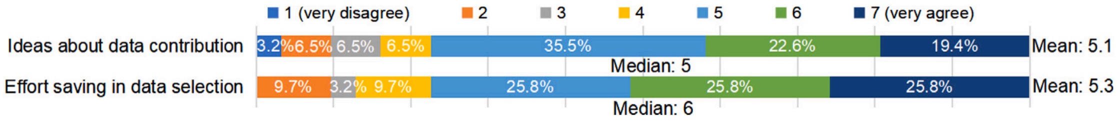


Fig. 10. Feedback on the comments about the recommendation.

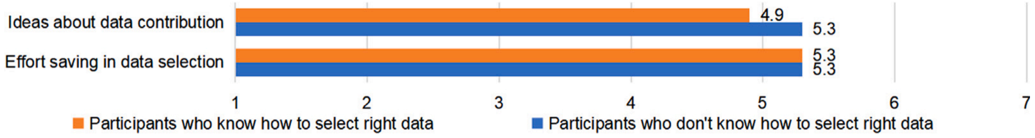


Fig. 11. Mean value for different groups in the comments about recommendation.

7.2.1. *Participant recruitment*

We asked the participants of the questionnaire on requirement analysis in Section 3 whether they would like to continue to assess whether our system meets their requirements. The majority of participants (31/38) expressed their interest in the evaluation. They consist of 5 women and 26 men. 17 participants are experts with relevant knowledge of FL and machine learning, and the remaining 14 are beginners.

7.2.2. *Study method and procedure*

In this study, we firstly presented our system design to the participants and described the main functions in the system. And then according to the stage division in Section 3, we collected and analyzed feedback during the four stages to assess our system. Specifically, we provided ten comments to this system, and used 7-point Likert scales to analyze the participants' acknowledgment of these comments. The script of comments is given in Appendix B.

7.2.3. *Analysis of the feedback*

We analyze the questionnaire data and organize our findings as following.

Recommendation for data selection. We investigate the participants' feedback on the comments about recommendation to see if the recommendation view could satisfy the demands of private data selection. As illustrated in Fig. 10, a majority of participants (77.5%) agree that the recommendation inspires them to think about which type of data contributes more to the diagnosis model. And 77.4% hold the positive view towards the effort saving of this recommendation in data selection. Then we further explore the attitudes of different groups of participants in combination with the previous survey of whether participants know how to select the right data. The results are shown in Fig. 11. For participants who know the selection, the mean values on the two comments are 4.9 and 5.3. And for participants who do not know the selection, the mean values are 5.3 and 5.3, respectively. The p-values of the two groups' feedback on the comments are 0.97 and 0.44. It demonstrates that both of them roughly agree our comments about the recommendation. And we found that participants who know the selection rate slightly higher than those who do not. More consideration may be needed for the latter.

Information leakage in FL. The main purpose of our system is to help users and institutions inspect the private information leakage in FL. We analyze whether our privacy risk views have achieved this target and provided some insights on information leakage and FL security issues. Fig. 12 demonstrates that most of participants affirmed the system in these aspects. In particular, 96.8% of participants agreed that this FL visualization system intuitively increases their perception of information leakage, and approximately 35.5% are highly agreed. By browsing and analyzing these privacy risks of data leakage, 87.1% think that this system prompts them to think about what kind of information is more vulnerable to leakage. And participants rated the comment highly in terms of the reflection on whether FL is privacy-preserving enough to meet their privacy demands.

According to the method of perceiving privacy risk we investigated before, we divided the participants into three groups, that is, participants perceiving by understanding the attack principle and building attack models by themselves, participants perceiving by studying the related literature and participants having no idea. The mean values of each group on our comments about information leakage are shown in Fig. 13. We found that all of them rated higher than 5.5 points on these comments. Participants building attack models by themselves rated the highest (higher than 6 points), followed by participants having no idea and participants perceiving by literature research. The p-values are higher than 0.73 when comparing the latter two groups, while the p-values range from 0.18 to 0.66 in the other group comparisons.

Interactive privacy control. We investigate the participants' feedback on the comments about the interactive privacy control to check whether the control can well assist users in enhancing privacy protection. The results are depicted in Fig. 14. For the familiarity of privacy control, 74.2% of participants agreed with their familiarity, 9.7% stayed neutral and the remaining 16.1% disagreed. Then we focus on the significance of our designed privacy control. Up to 96.7% believed that our system facilitates privacy analysis and enhancement when medical institutions use patients' face images to train the diagnosis model. We are very pleased to find that 87.1% of participants are more willing to contribute their sensitive data if our system is integrated in their participation in FL. 90.3% of them are willing to use our system to inspect and analyze the information leakage. It implies that

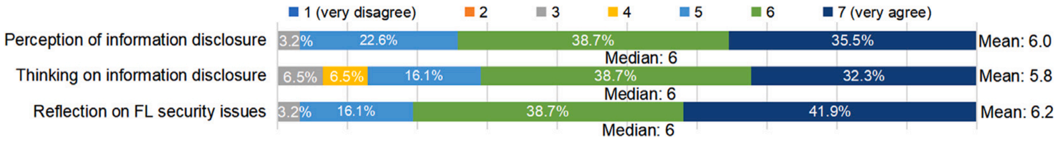


Fig. 12. Feedback on the comments about information leakage.

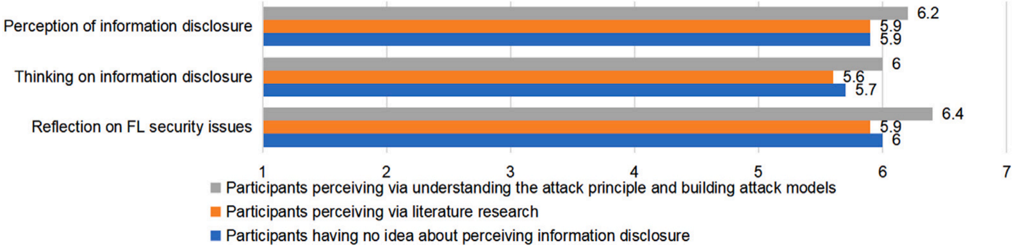


Fig. 13. Mean value for different groups in the comments about information leakage.

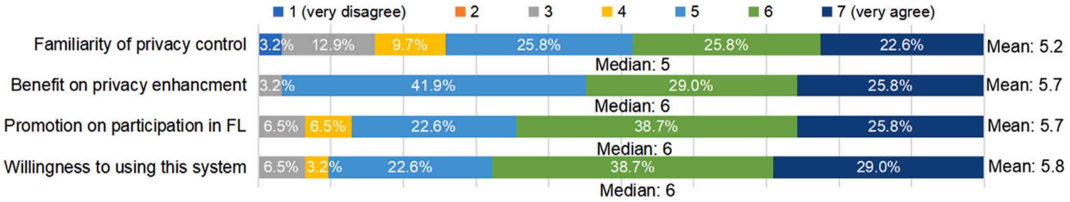


Fig. 14. Feedback on the comments about privacy enhancement.

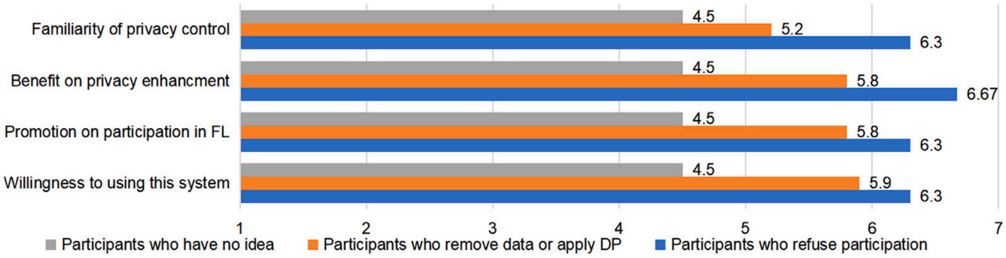


Fig. 15. Mean value for different groups in comments about privacy enhancement.

improving privacy perception and enhancing privacy control have great potential to encourage users and institutions to participate in FL.

We grouped participants based on their idiom of privacy enhancement methods and further discussed their feedback. The mean values for these groups are presented in Fig. 15. For participants having no idea to the enhancement, this control design seems to meet their demands very roughly. There still exists much room for improvement. For participants that remove sensitive data and apply DP algorithm, this control is basically in line with their preferences. Their mean values range from 5.2 to 5.9. For participants who refuse to participate in FL for protection, they understood the control design, rated the highest points (6.3–6.67) and thought they would change the previous refusal decision. The p-values among the three groups range from 0.06 to 0.53. From these p-values, we found that the latter two groups have relatively large differences in the promotion on participation in FL.

Trade-off between privacy protection and inference accuracy. As for the trade-off, we want to know whether these participants can make the desired trade-off more accurately and quickly through our system. Fig. 16 demonstrates their feedback on this comment about the trade-off. 93.5% of participants agreed that this system could benefit them to make the trade-off between privacy protection and inference accuracy, and the remaining 6.5% hold a neutral attitude. The median and mean of the rated values are 6 and 5.7, respectively.

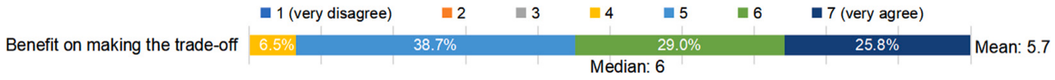


Fig. 16. Feedback on the comment about making the trade-off between privacy protection and inference accuracy.

In summary, our visualization system can generally meet the needs of users to explore data privacy in FL. Our recommendation view can generally help users think about which type of data is in need and saves their effort in selecting a small yet proper training subset. For privacy perception, our visualized privacy risks help users inspect information leakage and reflect on FL security issues. Especially, the users who are used to inspecting via understanding the attack theory themselves are very supportive of our privacy risk views. For privacy enhancement, most users are willing to use our interactive privacy control. And the control promotes their participation in FL, especially for users who refuse to participate due to privacy concerns. Almost all users agree that they could make a desired trade-off more accurately and quickly with our system.

8. Discussion

Privacy exploration based on the loop of perception and analysis: Our visualization system is designed for data owners to explore the data privacy in their FL practice. We first conducted a user study to perceive their privacy concerns and expectations. Then we designed our visualization system based on the perception. Finally, we collected their feedback on our system to perceive and analyze whether the system satisfies their demands. The findings suggest that the loop of perception and analysis helps us to design and improve our visual privacy exploration system.

Interactive privacy exploration: In practice, data/content privacy risks may vary with experts' domain knowledge or individuals' privacy sensitivity, so bringing human in the loop can effectively overcome the non-scalability of static privacy settings or rules. In fact, visualization systems have the nature advantage of embodying user interaction. Our results suggest that designing an interactive visualization system for privacy exploration has the potential to promote the recognition of new privacy-preserving computing paradigms.

Limitations and future work: We mainly focus on the horizontal FL scenario in this work. For example, multiple medical institutions collaboratively train the diagnosis model with their face images. The user feedback implies that our visualization system is effective in exploring data privacy in horizontal FL. We have also conducted a semi-structured interview with some participants to further explore the improving possibilities. In the review, some participants suggest exploring privacy in the vertical FL. The vertical FL also has many application scenarios, such as the recommendation via the cooperation between e-commerce and advertising platform providers. We are interested in extending our system for vertical FL scenarios.

9. Conclusion

This paper presented a study on general users' privacy concerns about FL and practical behaviors to protect their privacy. We found that due to the privacy concerns, the lack of capacity to understand and mitigate potential privacy risks had greatly hindered people's willingness of participation in FL. To bridge the gap, we proposed an interactive FL visualization system for users to efficiently explore the privacy information leakage properties in FL, including private data selection and training, information leakage inspection, iterative protection intervention towards a satisfactory trade-off between privacy and model performance. It is shown that our system can interact with clients within an acceptable response time range, and can provide them insights to understand and enhance privacy according to their preferences.

As a generalized FL privacy 'sensor' and 'protector', the proposed tool is compliance with different FL application scenarios (e.g., FL for driver drowsiness detection in intelligent driving, FL for individual health monitoring, and FL for masked face detection). It can bring convenience to the data release in these contexts and help to maintain a benign ecosystem for ubiquitous data sharing.

CRediT authorship contribution statement

Yeting Guo: Conceptualization, Investigation, Methodology, Writing – origin draft. **Fang Liu:** Resources, Supervision, Writing – review & editing. **Tongqing Zhou:** Conceptualization, Investigation, Methodology, Writing – review & editing. **Zhiping Cai:** Resources, Supervision, Writing – review & editing. **Nong Xiao:** Resources, Supervision.

Data availability

Data will be made available on request.

Acknowledgment

This work is supported by the National Key Research and Development Program of China (2020YFC2003404), National Natural Science Foundation of China (62172155, 61832020, 62072465, 62102425), Natural Science Foundation of Guangdong Province, China (2018B030312002), Science and Technology Innovation Program of Hunan Province, China (2021RC2071,2022RC3061), and National Natural Science Foundation of China-Guangdong Joint Fund (NSFCU1811461).

Appendix A. The script of the exploratory study on privacy inspection

Sec. 1. Demographics

Q1. With which gender do you identify?

○ Female ○ Male ○ Prefer not to answer

Q2. What is your current education?

○ Bachelor ○ Master ○ Doctor ○ Other

Q3. How old are you?

○ 0-15 ○ 15-30 ○ 30-45 ○ 45-60 ○ > 60

Q4. Do you have any expertise in machine learning and federated learning?

○ Experts who have prior ML experience (e.g., took at least one ML course) and FL experience (e.g., knows the workflow of FL)

○ Beginners who are from an Engineer background but not specific ML and FL experience

Sec. 2. Awareness of general clients on FL parameters privacy

Q5. Federated learning (FL) shares the local model parameters instead of their data. Do you agree that FL improves the privacy protection?

○ Yes ○ No ○ To a certain degree ○ Have no idea

Q6. Imagine that your sensitive data is helpful to model training in FL (for example, face images are helpful to build a diagnostic model for genetic diseases), and you will be paid for participation in training. What factors most affect your willingness to participate in FL? Please select the three most important factors.

○ Limited information about potential privacy leakage ○ Limited information about FL workflow

○ Resource consumption ○ Incentive of participation

○ Machine learning model performance

Sec. 3. Practice routine of training data selection

Q7. Different data may have different contributions. Will you choose partial local data to participate in the training in order to use the least data to get the maximum incentive?

○ Yes ○ No

Q8. [only shown if "No" is selected in Q7] Why are you not willing to select the data samples?

○ Training with all data ensures maximum incentive. I want the maximum incentive.

○ I don't know how to select the right data for training.

○ The selection is time-consuming.

○ Others [please specify] [__] (text)

Q9. Do you know how to select the right data?

○ Yes ○ No

Sec. 4. Sensitivity on privacy threats in FL

Q10. To what extent would you be worried if your data are accurately reconstructed from your local model parameters?

○ Not worried ○ Slightly worried ○ Moderately worried ○ Very worried

Q11. To what extent would you be worried if your participation is inferred? For example, you are inferred to train the model for a particular disease.

○ Not worried ○ Slightly worried ○ Moderately worried ○ Very worried

Q12. To what extent would you be worried if the attacker could accurately infer that the data contain some attributes? For example, the data include patient data from a particular disease.

○ Not worried ○ Slightly worried ○ Moderately worried ○ Very worried

Q13. From your current FL practices, how do you explore the data privacy?

○ Understand how these attacks work, and train corresponding attack models at local and then use them to evaluate data leakage

○ Search related work, find the work closest to my own scenarios to evaluate data leakage

○ Have no idea

○ Others [please specify] [__]

Sec. 5. General countermeasures for potential threats

Q14. How do you enhance your privacy protection? ○ Refuse participation in FL

○ Remove some sensitive data from the training dataset

○ Apply some privacy preserving algorithms, like differential privacy

○ Have no idea

○ Others [please specify] [__]

Q15. What do you think of the importance of the following factors in privacy intervention?

Raw options:

- Diversity of privacy protection methods - Effectiveness of privacy protection methods
- Efficiency of privacy protection methods - Ease of use of privacy protection methods
- Impact on model performance

Column options:

○ Very unimportant ○ Moderately unimportant ○ Slightly unimportant

○ Neither important nor unimportant ○ Slightly important ○ Moderately important

○ Very important

Sec. 6. Ways to handle privacy and utility trade-off

Q16. Privacy enhancement always sacrifices a certain model performance. What kind of performance metrics do you focus on?

○ Inference accuracy in each round ○ Loss value in each round ○ Incentive in each round

○ None of them ○ Others [please specify] [__]

Q17. What kind of methods would you prefer to making a desired trade-off between model performance and privacy protection?

○ Output the model performance when the probability of privacy leakage reaches its preset upper limit

○ Outputs the probability of privacy leakage when the model performance reaches its default value

○ Interactive adjustment method

Appendix B. The script of the comments on our visualization system

C1: This recommendation page can save effort when selecting face images for FL-based genetic disease diagnosis.

C2: The recommendation page can help me think about which kinds of face images contribute more to the genetic disease diagnostic model.

C3: This system can help me inspect the privacy information leakage in FL.

C4: The system can help me reflect on whether the training process of FL is safe enough.

C5: The system can help me think about what kind of face image data is easy to leak.

C6: I am familiar with the interactive privacy control.

C7: The system facilitates privacy analysis and privacy enhancement when hospital organizations use patient information in FL.

C8: I would be more willing to contribute my data if institutions such as hospitals use the system during their participation in FL.

C9: I would like to use this system to inspect information leakage and enhance privacy protection in federated learning.

C10: The system can help achieve the desired trade-off between privacy and accuracy.

Appendix C. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ipm.2022.103162.

References

Abadi, M., Chu, A., Goodfellow, I. J., McMahan, H. B., Mironov, I., Talwar, K., et al. (2016). Deep learning with differential privacy. In *Proc. of ACM SIGSAC conference on computer and communications security* (pp. 308–318). ACM.

Andalibi, N., & Buss, J. (2020). The human in emotion recognition on social media: Attitudes, outcomes, risks. In *Proc. of CHI* (pp. 1–16). ACM.

Banabilah, S., Aloqaily, M., Alsayed, E., Malik, N., & Jararweh, Y. (2022). Federated learning review: Fundamentals, enabling technologies, and future applications. *Information Processing & Management*, *59*(6), Article 103061.

Bethge, D., Kosch, T., Grosse-Puppendahl, T., Chuang, L. L., Kari, M., Jagaciak, A., et al. (2021). Vemotion: Using driving context for indirect emotion prediction in real-time. In *Proc. of UIST* (pp. 638–651). ACM.

Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., et al. (2019). Towards federated learning at scale: System design. CoRR, https://arxiv.org/abs/1902.01046.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., et al. (2021). Machine unlearning. In *Proc. of IEEE symposium on security and privacy (S&P)* (pp. 141–159). IEEE.

Bu, D., Wang, X., & Tang, H. (2021). Haplotype-based membership inference from summary genomic data. *Bioinformatics*, *37*(Supplement), 161–168.

Chen, H., Li, H., Dong, G., Hao, M., Xu, G., Huang, X., et al. (2022). Practical membership inference attack against collaborative inference in industrial IoT. *IEEE Transactions on Industrial Informatics*, *18*(1), 477–487.

Chen, W., Wei, Y., Wang, Z., Zhou, S., Lin, B., & Zhou, Z. (2020). Federated visualization: A privacy-preserving strategy for decentralized visualization. CoRR, https://arxiv.org/abs/2007.15227.

Dahlgaard, M. E., rgensen, M. W. J., Fuglsang, N. A., & Nassar, H. (2022). Analysing the influence of attack configurations on the reconstruction of medical images in federated learning. CoRR, https://arxiv.org/abs/2204.13808.

Dang, T. T., Dang, T. K., & Küng, J. (2020). Interaction and visualization design for user privacy interface on online social networks. *SN Computer Science*, *1*(5), 297.

Dayan, I., Roth, H. R., Zhong, A., Harouni, A., Gentili, A., Abidin, A. Z., et al. (2021). Federated learning for predicting clinical outcomes in patients with COVID-19. *Nature Medicine*, *27*(10), 1735–1743.

Dimitriadis, G., Neto, J. P., & Kampff, A. R. (2018). t-SNE visualization of large-scale neural recordings. *Neural Computation*, *30*(7).

Erlich, Y., & Narayanan, A. (2014). Routes for breaching and protecting genetic privacy. *Nature Reviews Genetics*, *15*(8), 409.

Fernandez, C. B., Nurmi, P., & Hui, P. (2021). Seeing is believing?: Effects of visualization on smart device privacy perceptions. In *Proc. of ACM multimedia conference* (pp. 4183–4192). ACM.

Ganju, K., Wang, Q., Yang, W., Gunter, C. A., & Borisov, N. (2018). Property inference attacks on fully connected neural networks using permutation invariant representations. In *Proc. of ACM SIGSAC conference on computer and communications security* (pp. 619–633). ACM.

Geiping, J., Bauermeister, H., Dröge, H., & Moeller, M. (2020). Inverting gradients - how easy is it to break privacy in federated learning? In *Proc. of conference on neural information processing systems (NIPS), Vol. 33* (pp. 16937–16947). Curran Associates, Inc..

Ghorbani, A., & Zou, J. Y. (2019). Data Shapley: Equitable valuation of data for machine learning. In *Proc. of international conference on machine learning (ICML), Vol. 97* (pp. 2242–2251). PMLR.

Gildenblat, J., & contributors (2021). Pytorch library for CAM methods. https://github.com/jacobgil/pytorch-grad-cam.

Gu, Y., Bai, Y., & Xu, S. (2022). CS-MIA: Membership inference attack based on prediction confidence series in federated learning. *Journal of Information Security and Applications*, *67*, Article 103201.

Guo, Y., Liu, F., Cai, Z., Zeng, H., Chen, L., Zhou, T., et al. (2021). PREFER: point-of-interest recommendation with efficiency and privacy-preservation via federated edge learning. *ACM Interaction, Mobile, Wearable and Ubiquitous Technology*, *5*(13), 1–25.

Karegar, F., & Fischer-Hübner, S. (2021). Vision: A noisy picture or a picker wheel to spin? Exploring suitable metaphors for differentially private data analyses. In *Proc. of European symposium on usable security* (pp. 29–35). ACM.

Kumov, V., & Samorodov, A. (2020). Recognition of genetic diseases based on combined feature extraction from 2D face images. In *Conference of open innovations association (FRUCT)* (pp. 1–7). IEEE.

Li, Q., Wei, X., Lin, H., Liu, Y., Chen, T., & Ma, X. (2021). Inspecting the running process of horizontal federated learning via visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 1–15.

Li, Y., Zhou, Y., Jolfaei, A., Yu, D., Xu, G., & Zheng, X. (2021). Privacy-preserving federated learning framework based on chained secure multiparty computing. *IEEE Internet Things of Journal*, *8*(8), 6178–6186.

Lim, S., Hibschman, J., Zhang, H., & O'Rourke, E. (2018). Ply: A visual web inspector for learning from professional webpages. In *Proc. of UIST* (pp. 991–1002). ACM.

Lim, W. Y. B., Huang, J., Xiong, Z., Kang, J., Niyato, D., Hua, X.-S., et al. (2021). Towards federated learning in UAV-enabled internet of vehicles: A multi-dimensional contract-matching approach. *IEEE Transactions on Intelligent Transportation Systems*, *22*(8), 5140–5154. http://dx.doi.org/10.1109/TITS.2021.3056341.

Liu, D. C., & Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, *45*(1–3), 503–528.

Lu, Z., Jiang, X., & Kot, A. C. (2018). Deep coupled ResNet for low-resolution face recognition. *IEEE Signal Processing Letters*, *25*(4), 526–530.

Melis, L., Song, C., Cristofaro, E. D., & Shmatikov, V. (2019). Exploiting unintended feature leakage in collaborative learning. In *Proc. of IEEE symposium on security and privacy (S&P)* (pp. 691–706). IEEE.

Meng, L., Wei, Y., Pan, R., Zhou, S., Zhang, J., & Chen, W. (2021). VADAF: Visualization for abnormal client detection and analysis in federated learning. *ACM Transactions on Intelligent Systems and Technology*, *11*(26), 1–23.

Miao, Y., Xue, M., Chen, C., Pan, L., Zhang, J., Zhao, B. Z. H., et al. (2021). The audio auditor: User-level membership inference in internet of things voice services. *Proceedings on Privacy Enhancing Technologies*, *2021*(1), 209–228.

Mike (2018). Federated learning: distributed machine learning with data locality and privacy. https://blog.fastforwardlabs.com/2018/11/14/federated-learning-distributed-machine-learning-with-data-locality-and-privacy.html.

Muchagata, J., Vieira-Marques, P., & Ferreira, A. (2019). Mhealth applications: Can user-adaptive visualization and context affect the perception of security and privacy? In *Proc. of international conference on enterprise information systems* (pp. 444–451). SciTePress.

Nanayakkara, P., Bater, J., He, X., Hullman, J., & Rogers, J. (2022). Visualizing privacy-utility trade-offs in differentially private data releases. *Proceedings on Privacy Enhancing Technologies*, *2022*(2), 601–618.

Ren, H., Deng, J., & Xie, X. (2022). GRNN: Generative regression neural network—A data leakage attack for federated learning. *ACM Transactions on Intelligent Systems and Technology*, *13*(4), 1–24.

Shao, C., Yang, Y., Juneja, S., & Seetharam, T. G. (2022). IoT data visualization for business intelligence in corporate finance. *Information Processing & Management*, *59*(1), Article 102736.

Shen, M., Wang, H., Zhang, B., Zhu, L., Xu, K., Li, Q., et al. (2021). Exploiting unintended property leakage in blockchain-assisted federated learning for intelligent edge computing. *IEEE Internet of Things Journal*, *8*(4), 2265–2275.

Soumelidou, A., & Tsohou, A. (2020). Effects of privacy policy visualization on users' information privacy awareness level. *Information Technology & People*, *33*(2), 502–534.

Sun, R., Li, Y., Shah, T., Sham, R. W. H., Szydlo, T., Qian, B., et al. (2022). FedMSA: A model selection and adaptation system for federated learning. *Sensors*, *22*(19).

Sun, L., & Lyu, L. (2021). Federated model distillation with noise-free differential privacy. In *Proc. of international joint conference on artificial intelligence (IJCAI)* (pp. 1563–1570). Morgan Kaufmann.

Sun, L., Qian, J., & Chen, X. (2021). LDP-FL: practical private aggregation in federated learning with local differential privacy. In *Proc. of international joint conference on artificial intelligence (IJCAI)* (pp. 1571–1578). Morgan Kaufmann.

Tao, F. (2019). FATE-board: FATE's visualization toolkit. https://github.com/FederatedAI/FATE-Board.

Velykoivanenko, L., Niksirat, K. S., Zufferey, N., Humbert, M., Huguenin, K., & Cherubini, M. (2022). Are those steps worth your privacy? Fitness-tracker users' perceptions of privacy and utility. *ACM Interaction, Mobile, Wearable and Ubiquitous Technology*, *5*(4), 1–41.

Verizone (2021). 2021 Data breach investigations report. https://www.verizon.com/business/resources/reports/dbir/.

Wang, X., Chen, W., Xia, J., Wen, Z., Zhu, R., & Schreck, T. (2022). HetVis: A visual analysis approach for identifying data heterogeneity in horizontal federated learning. http://dx.doi.org/10.48550/arXiv.2208.07491, CoRR.

Wei, X., Li, Q., Liu, Y., Yu, H., Chen, T., & Yang, Q. (2019). Multi-agent visualization for explaining federated learning. In *Proc. of international joint conference on artificial intelligence (IJCAI)* (pp. 6572–6574). Morgan Kaufmann.

Wei, W., Liu, L., Loper, M., Chow, K. H., Gursoy, M. E., Truex, S., et al. (2020). A framework for evaluating gradient leakage attacks in federated learning. CoRR https://arxiv.org/abs/2004.10397.

Wilkinson, D., Bahirat, P., Namara, M., Lyu, J., Alsubhi, A., Qiu, J., et al. (2020). Privacy at a glance: The user-centric design of glanceable data exposure visualizations. *Proceedings on Privacy Enhancing Technologies*, *2020*(2), 416–435.

Willemsen, B. (2021). Hype cycle for privacy, 2021. https://www.gartner.com/en/documents/4003504-hype-cycle-for-privacy-2021.

Wu, D., Deng, Y., & Li, M. (2022). FL-MGVN: federated learning for anomaly detection using mixed gaussian variational self-encoding network. *Information Processing & Management*, *59*(2), Article 102839.

Yang, H., Wang, Y., & Li, B. (2022). Individual property inference over collaborative learning in deep feature space. In *IEEE international conference on multimedia and expo (ICME)* (pp. 1–6). IEEE, http://dx.doi.org/10.1109/ICME52920.2022.9859857.

Zhang, X., Chen, X., Hong, M., Wu, S., & Yi, J. (2022). Understanding clipping for federated learning: Convergence and client-level differential privacy. In *Proc. of machine learning research, Vol. 162* (pp. 26048–26067). PMLR.

Zhang, M., Ren, Z., Wang, Z., Ren, P., Chen, Z., Hu, P., et al. (2021). Membership inference attacks against recommender systems. In *ACM conference on computer and communications security* (pp. 864–879). ACM.

Zhang, W., Tople, S., & Ohrimenko, O. (2021). Leakage of dataset properties in multi-party machine learning. In *Proc. of USENIX security symposium* (pp. 2687–2704). USENIX Association.

Zhang, J., Zhang, J., Chen, J., & Yu, S. (2020). GAN enhanced membership inference: A passive local attack in federated learning. In *Proc. of IEEE international conference on communications* (pp. 1–6).

Zhao, B., Mopuri, K. R., & Bilen, H. (2020). iDLG: Improved deep leakage from gradients. CoRR http://arxiv.org/abs/2001.02610.

Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. In *Proc. of conference on neural information processing systems (NIPS)* (pp. 14747–14756). Curran Associates, Inc..