

A Survey of Federated Evaluation in Federated Learning

Behnaz Soltani¹, Yipeng Zhou^{1*}, Venus Haghghi¹ and John C.S. Lui²

¹ School of Computing, Macquarie University, Sydney, Australia

² Department of Computer Science and Engineering, The Chinese University of Hong Kong, HKSAR
{behnaz.soltani@hdr., yipeng.zhou@, venus.haghghi@hdr.}mq.edu.au, csloi@cse.cuhk.edu.hk

Abstract

In traditional machine learning, it is trivial to conduct model evaluation since all data samples are managed centrally by a server. However, model evaluation becomes a challenging problem in federated learning (FL), which is called *federated evaluation* in this work. This is because clients do not expose their original data to preserve data privacy. Federated evaluation plays a vital role in client selection, incentive mechanism design, malicious attack detection, etc. In this paper, we provide the first comprehensive survey of existing federated evaluation methods. Moreover, we explore various applications of federated evaluation for enhancing FL performance and finally present future research directions by envisioning some challenges.

1 Introduction

Recently, federated learning (FL) has emerged as a privacy-preserving framework, in which clients collaboratively train shared machine learning models without exposing their own local data during the training process [McMahan *et al.*, 2017]. FL can extensively exploit massive data samples scattered on decentralized clients such as Internet-of-Things (IoTs) and mobile devices for model training [Zhou *et al.*, 2019]. With FL, clients only expose model information rather than original data samples for training models. More specifically, the FL server distributes the global model to selected clients; participants train local models iteratively on their own data and send their local models to the server; the server aggregates the local models to generate the updated global model. The above steps are repeated for a certain number of iterations.

In traditional machine learning, it is trivial to conduct model evaluation with centrally collected data samples from clients. Yet, the model evaluation problem becomes very challenging in FL since all data samples are owned and privately retained by clients. Without owning any data, the server cannot manipulate data for model evaluation.

In FL, model evaluation plays a significant role in model training, which is much more complicated than traditional

machine learning. On the one hand, evaluating a model accurately in FL is essential for designing incentive mechanisms by reasonably rewarding each participating client [Zhang *et al.*, 2021a; Gao *et al.*, 2021; Deng *et al.*, 2021], devising efficient client selection strategies [Li *et al.*, 2021a; Lai *et al.*, 2021], detecting malicious attacks [Che *et al.*, 2022; Gao *et al.*, 2021] and deriving personalized models [Jia *et al.*, 2019; Huang *et al.*, 2021] based on evaluation results. On the other hand, each FL client has a local model trained on their own local data implying that each individual FL client can be evaluated independently. However, without the knowledge about clients' data, it is a challenging problem to evaluate the importance of clients.

To make model evaluation feasible in FL, tremendous efforts have been dedicated by existing works. We propose two different ways to categorize existing methods. Firstly, its architecture can be categorized as: centralized federated evaluation and decentralized federated evaluation. The former one assumes that a single FL server or task owner evaluates the quality of FL models. The latter one recruits a number of independent clients to conduct federated evaluation of models in a distributed fashion. Secondly, federated evaluation methods can be categorized based on their evaluation approaches such as data-level evaluation, utility-based approach, Shapley values approach, and statistical metric-based approach.

To the best of our knowledge, there are no existing works that explore federated evaluation in different scenarios. To bridge this gap, we review existing methods, survey the applications of federated evaluation results, discuss the challenges of federated evaluation and envision potential future work.

2 Federated Evaluation Architecture

In this section, we briefly introduce the workflow of FL and discuss two kinds of federated evaluation architectures: centralized architectures and decentralized architectures.

2.1 FL System

In a FL system, there are typically multiple decentralized clients that participate in the training process. Each client owns a training dataset and a test dataset. The objective of FL clients is to collaboratively train a shared model. FL is usually conducted for multiple global iterations (a.k.a. rounds). At the beginning of each global iteration, a global model is

*Corresponding author.

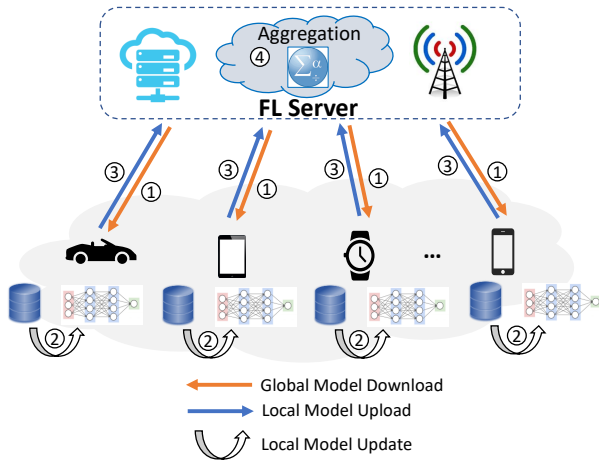


Figure 1: An overview of the federated learning process.

distributed by the server to participating clients. On each participating client, the global model is updated with their local dataset to obtain a local model. Then, each client returns model information (e.g. model parameters and model gradients) to the server. The server aggregates collected models from participating clients to update the global model.

In Fig. 1, we present a snapshot of FL for a particular global iteration. It is worth noting that:

- In FL, each individual client contributes a local model trained based on private local data. It implies that each model can be evaluated independently.
- The FL server cannot touch data samples owned by clients, and hence the server is unable to directly conduct evaluation of models. To overcome this shortcoming, the FL server can exploit its auxiliary data or employ statistical methods to evaluate clients.

Based on Fig. 1, we broadly discuss how existing works conduct federated evaluation to evaluate models including both the global model and local models in a FL system.

2.2 Centralized Federated Evaluation

Clients can be evaluated using a single FL server or task owner based on statistical information uploaded by clients. The most straightforward way is to evaluate clients based on their data sizes [Feng *et al.*, 2019]. The server can evaluate clients more accurately if clients can share label distributions to the server [Ma *et al.*, 2021a]. More complicated statistical metrics can be designed to evaluate local models on the server side based on model information (e.g. model parameters and model gradients), which are uploaded from clients to the server [Deng *et al.*, 2021; Gao *et al.*, 2021].

If test data is available at the server, the server can evaluate local models using test data directly, though the assumption that the server holds test data is strong and impracticable in many scenarios. A number of centralized evaluation methods are introduced as follows. In [Zhang *et al.*, 2021a], local models are evaluated using the validation set on the server.

A key idea here is that if a high-quality local model participates in the aggregation process, the loss value of the global model should be decreased. FIFL [Gao *et al.*, 2021] defines the marginal test loss to detect malicious clients in FL. It uses the first Taylor’s first-order expansion to simplify its calculation. Therefore, the similarity distance between the local gradients and the server gradient obtained from the test dataset at the server is computed for detecting abnormal local models. In [Deng *et al.*, 2021], local models are evaluated based on the difference between the average loss value of the global model on the test dataset and the average training loss value of the local models. The evaluation process also takes historical records into account. It allocates larger weights to recent records due to their higher level of informativeness. Clients are selected with the aim of maximizing the sum of evaluated qualities subject to a budget. The test performance difference with or without a local model is another metric to evaluate the importance of clients [Kang *et al.*, 2019].

Even if the test data is not available on the server, clients can still be centrally evaluated using their local model information. For example, the difference between local model parameters before and after a training round is considered as a quality metric evaluated by the server [Zhao *et al.*, 2022].

2.3 Decentralized Federated Evaluation

Federated evaluation can be performed on multiple decentralized clients or third parties instead of a single server because of two main reasons: serverless FL and decentrally distributed test data across clients.

In decentralized FL (DFL), there is no dedicated server to conduct centralized evaluation of models [Su *et al.*, 2022]. For example, blockchain-based FL consists of miners and devices without relying on a central server [Kim *et al.*, 2019]. Miners, possibly from clients or third parties’ devices such as base stations, are responsible for evaluating local models in a decentralized manner to exclude malicious attackers. Only local models that are verified by miners can be recorded in a generated block with a consensus algorithm. A hierarchical blockchain-based architecture is proposed by [Kang *et al.*, 2020a] which utilizes multiple consortium blockchains as subchains to conduct decentralized model evaluation based on model accuracy. A committee-based serverless FL is designed by [Che *et al.*, 2022], in which honest clients are selected as committee members to decentrally evaluate local models based on the difference between the local gradients and the committee gradients. Similarly, Refiner [Zhang *et al.*, 2021c] selects a committee of randomly selected validators to evaluate local models based on loss values.

When test data is privately distributed on decentralized clients, model evaluation using test data must be performed in a distributed manner on clients. In this case, it is common to select clients to evaluate model performance using their own test data. The test results can be collected by the server for other purposes. For example, Oort [Lai *et al.*, 2021] selects participants to serve the developer-specified criteria on testing data. FedFomo [Zhang *et al.*, 2020] and L2C [Li *et al.*, 2022] are personalized FL algorithms, which locally evaluate models collected from other clients to locally customize aggregation weights so as to pursue personalized models.

3 Federated Evaluation Approaches

Due to the inaccessibility of local data in FL, it is challenging to directly evaluate local models. Therefore, we introduce various approaches for indirect evaluation of models.

3.1 Data-level Evaluation

Although accessing original data is forbidden in FL, it is still possible to obtain data quantity information from FL clients. To a certain extent, the local model quality can be evaluated by data quantity information.

In original FL, FedAvg simply uses the local dataset size to determine the aggregation weight of a particular client's local model [McMahan *et al.*, 2017]. However, the non-IID (Identical and Independently Distributed) distribution of data across clients can degrade the utility of the model. In [Feng *et al.*, 2019], the server negotiates with clients about the sizes of their data, and in return, clients receive rewards based on their data sizes. The goal is to maximize the total amount of training data in order to achieve higher learning accuracy.

Later on, more advanced methods are proposed that evaluate the quality of clients based on their data distribution. In [Li *et al.*, 2021a], prior to the training process, the server quantifies the intersection between the label sets of clients and the target label set. Clients with an intersection higher than a specified threshold are considered as relevant clients. To preserve privacy, the calculation of intersection is performed using a private set intersection (PSI) method [Agrawal *et al.*, 2003]. Next, the server selects clients for training based on their high statistical homogeneity and content diversity. Statistical homogeneity is evaluated using the similarity between a uniform distribution generated on the server and clients' distributions based on homomorphic encryption. Content diversity is evaluated by computing the similarity of clients' data using a noisy content sketch, which is obtained as follows. Each client generates a content embedding vector for each sample using a general deep learning model. The client's data is encoded into a low-dimensional vector based on JL-transformation [Biswas *et al.*, 2019] as a content sketch.

Label quantity information can also be utilized to evaluate clients. In a grouping-based mechanism, clients are divided into multiple groups based on their label quantity information shared with the server. Only clients in the same group are selected for training [Ma *et al.*, 2021a]. This approach introduces a new metric named Group Earth Mover's Distance (GEMD) inspired by Earth Mover's Distance (EMD) [Zhao *et al.*, 2018] to evaluate the difference between the global data distribution and the selected local distributions. A smaller GEMD implies that the data distribution is closer to IID. A pair-wise grouping mechanism is proposed in which each client is initially considered as a separate group. Based on GEMD, each group is merged into a pair iteratively to complement missing labels. The objective is to make the aggregated data distribution close to the global distribution.

3.2 Model Utility

Similar to traditional machine learning, the quality of a local model can be evaluated based on its utility, which can be measured in terms of the loss value or model accuracy.

A new client selection framework called Oort has been introduced in [Lai *et al.*, 2021], which tries to select the most significant clients for training in each global iteration. Its metric to evaluate a client's importance is the loss value obtained by training the model with local data on each client. Based on the aggregate training loss across all data samples, Oort can dedicatedly select the most important clients to participate in FL. Similarly, FedSAE [Li *et al.*, 2021b] evaluates the importance of each client based on the local training loss and the number of local samples. According to the importance values, the server determines the selection probability of each client per global iteration.

In [Wu and Wang, 2022], an optimal aggregation mechanism is proposed to reduce overall data heterogeneity by excluding adverse local models with the objective of enlarging the expected decrement of the global loss. The proposed method iteratively removes a local model and compares the expected inner product between the local gradients and the global gradient before and after excluding the local model by assuming that the inner product implicitly represents the difference between local data distributions and global data distribution. To ensure that excluding local models leads to a faster convergence, global losses are measured for both global models (i.e. with and without the aforementioned local model) using a test dataset. Based on the change of loss values, the decision is made for removing a local model. A tier-based FL system is proposed in [Chai *et al.*, 2020], which divides clients into tiers based on their response latencies. Clients from the same tier are selected to participate in FL in case that the training process is slowed down by a slow client. However, a tier-based client selection can incur training bias since a faster tier is prone to be selected with a higher probability. To eliminate bias, the global model is evaluated by each tier to estimate the importance of that tier. The selection probability of a tier is adjusted based on the test accuracy obtained at different rounds.

Blockchain-based FL is proposed to make the FL process traceable and tamper-proofing without relying on a single server [Kim *et al.*, 2019; Qu *et al.*, 2020]. Blockchain-based FL widely adopts model utility to evaluate the quality of models contributed by clients. Refiner [Zhang *et al.*, 2021c] introduces a FL system implemented upon Ethereum, a public blockchain to deal with self-interested and malicious devices. A committee of randomly selected validators are employed to evaluate local models and prevent corrupted local models from participating in the aggregation process. Local models are evaluated by computing the loss function on the validation dataset provided by the FL task owner. If the loss values of the local models are lower than a specified threshold, they will be considered qualified and included in the aggregated global model. A hierarchical blockchain framework has been introduced in [Kang *et al.*, 2020a], which consists of a public blockchain as a main blockchain, and multiple consortium blockchains as subchains to store local models for model quality evaluation. In a subchain, the miners evaluate the quality of local models by evaluating their accuracy on a test dataset provided by the FL task owner. Local models are qualified if their accuracy is higher than a defined threshold, which will be recorded in a pending block later.

3.3 Shapley Values

FL can be regarded as a cooperative game played by multiple FL clients. It is proposed that Shapley values (SV) [Jia *et al.*, 2019; Ghorbani and Zou, 2019], a method in cooperative game theory, can efficiently evaluate the merit of each FL client. The computation of SV is based on the average contribution (in terms of model utility) of a data source to every subset of data sources. In the computation of SV, it is unnecessary to consider the order of data sources during training. However, the computation complexity of SV is exponential, which makes it unaffordable in reality. Different variants have been devised to approximately compute SV in FL.

A variant named federated SV has been introduced in [Wang *et al.*, 2020] which computes SV from local models without extra communication costs. It also captures the effect of the order of participation because data sources employed earlier have more impact on the model performance compared to those used at the end of the training. The federated SV is estimated using algorithms such as permutation sampling-based approximation and group testing-based approximation. However, federated SV may lead to unfairness on a large scale since only a subset of clients are selected in each round and non-selected ones receive zero credit in the corresponding global iteration. Therefore, clients with identical local data may receive different credits. Completed federated SV has been introduced in [Fan *et al.*, 2022] to address the aforementioned challenges. It introduces a utility matrix that consists of the contributions of all possible subsets of clients across all training iterations. Due to the partially observed utility matrix caused by the partial selection of clients in each round, the goal is to complete missing entries in the utility matrix. To achieve this, a low-rank matrix completion problem is designed. A group-based SV computation for blockchain-based FL is proposed in [Ma *et al.*, 2021b]. It divides clients into several groups according to a permutation sample, and the aggregated global model is obtained for each group. A new model is generated by aggregating different group models. Finally, the SV of each group is estimated based on model utility and assigned to its client members.

3.4 Statistical Metric

Models can be indirectly evaluated based on statistical metrics. The most widely used one is model distance metrics such as distances between model parameters, gradients, or model performances.

[Zhao *et al.*, 2022] evaluates a local model using the model parameter difference metric before and after a training round. Clients providing higher model parameter divergence are considered to have higher quality, as clients with IID data have larger parameter differences than those with non-IID data. However, clients with large amounts of data can prolong the duration of a round, causing other clients to wait for those slow clients to complete their training. Therefore, a size-related ratio is added to the divergence metric to take into account both data distribution and data size. [Zhang *et al.*, 2021b] further proposes to use the model parameter divergence between a local model and a model trained on an auxiliary IID dataset residing on the server to evaluate the degree of non-IID datasets. Clients with a lower degree of non-

IID data lead to a lower divergence, which can accelerate the FL convergence.

The quality-aware framework proposed in [Deng *et al.*, 2021] designs a mechanism to remove unreliable local models from the aggregation process. To evaluate the model quality, the framework measures the median, mean, and standard deviation of the cosine similarity between the local model parameters and the global model parameter. Since clients are selected based on the loss reduction during the learning process, the majority of the received updates are expected to be of high quality. Therefore, if the mean is greater than the median, then the similarity values of low-quality models are higher than the median; otherwise, they are lower. The distance between local and global gradients is computed in [Gao *et al.*, 2021] using the square of the Euclidean norm to evaluate the contribution of clients. To differentiate between positive and negative contributions, a threshold is set based on the gradient distance. Clients with a gradient distance above the threshold are considered to have a negative contribution, while those below it have a positive contribution.

FOCUS [Chen *et al.*, 2020a] is proposed to evaluate local models by evaluating the quality of local data labels. Each client evaluates the performance of the global model on its local dataset and sends the evaluation results to the server. The server evaluates each local model on its benchmark dataset and calculates the cross entropy between the two sets of evaluation results to measure the quality of the clients' local labels. In [Wang *et al.*, 2019], a deletion method is introduced where data samples from each client are deleted, the model is retrained, and the difference in prediction results between the new global model and the original one is computed to determine the contribution of each client.

A reputation system can be maintained to evaluate the reliability and quality of local models in FL. A client's reputation can be determined using the combination of a direct reputation (i.e. the reputation evaluated by the task requester) and an indirect reputation (i.e. the reputation evaluated by other requesters). Reputation values are used to conduct client selection in [Kang *et al.*, 2019; Kang *et al.*, 2020b], where the reputation of each client is calculated using the multiweight subjective logic model [Liu *et al.*, 2011]. The model considers three weights: interaction effects (i.e. positive or negative interaction evaluated by quality measurement), interaction timeliness, and interaction frequency. To evaluate the quality of local models, attack detection mechanisms such as Reject on Negative Influence (RONI) [Shayan *et al.*, 2018] and FoolsGold [Fung *et al.*, 2018] are used for IID and non-IID data distributions, respectively. RONI evaluates local models by computing the difference between the performance with and without a local model on a dataset specified by the task publisher. The corresponding local model is discarded from the aggregation process if the performance difference falls below a certain threshold. FoolsGold evaluates clients based on the gradient diversity of their local models. Clients uploading similar gradients in each round are identified as unreliable workers which may contribute unreliable models, and are excluded from the aggregation process. In the collaborative FL proposed in [Zhang *et al.*, 2021a], each individual client can be a task requester or a participant. To evaluate the contribu-

tion of local models, all local models and the global model are recorded at each global iteration. A local model that moves more towards the target model has a higher contribution to the global model, leading to a faster convergence speed. To measure the contribution of each client in each round, first, the direction vector is obtained between the initial global model and the final global model. Then, the local model is projected onto the direction vector and multiplied by the absolute value of the cosine of the angle between the local model and the direction vector. Eventually, the sum of the contributions of each client's local model across multiple rounds is calculated to obtain the overall contribution of clients.

In serverless FL, local models can also be evaluated based on statistical metrics. A committee-based serverless FL framework is proposed in [Che *et al.*, 2022] where honest clients are selected as committee members to filter local gradients for defending against Byzantine attacks or accelerating FL convergence. To exclude attackers, the local gradients close to the committee gradients are selected for model aggregation based on the Euclidean distance. This is because the Euclidean distance between a malicious gradient and an honest gradient is larger than the distance between two honest gradients. However, this strategy may degrade the performance of FL since honest clients with large gradient differences have less opportunity to participate in aggregation. Therefore, to accelerate FL convergence in a non-attack scenario, clients with different local updates are accepted. To obtain the final score of each client calculated using the Euclidean distance, the committee members broadcast their evaluation scores to each other. To reach a consensus, a client is selected randomly as the primary client, which sends a request to the other committee members to confirm the correctness of its aggregation set. If so, the aggregation process is performed on the committee clients, and if the result is consistent with the request, it is sent back to the primary client. If the primary client receives a sufficient number of consistent results, a consensus is reached. Otherwise, the primary committee member is reallocated and the process is repeated.

In FL, it is possible that there are multiple learning tasks such as personalization. The challenge for training multiple tasks is to customize the training progress for each individual task based on federated evaluation results. A personalized FL framework is introduced in [Huang *et al.*, 2021] where clients have separate personalized target models for learning at the server. It investigates pair-wise collaborations among clients. Specifically, the server collects local personalized model parameters from clients to update the model for each client by a weighted convex combination of received model parameters. The weights (i.e. contributions) of clients for each personalized target model are evaluated by the similarity between the model parameters of two clients. Therefore, clients with similar model parameters have more contributions to each other. Each client can request its respective model parameters from the server to optimize its local personalized model using their private data. A similar personalized FL architecture (i.e. a personalized target model for each client) has been introduced in [Ma *et al.*, 2022] which assigns different weights for model layers when aggregating personalized models. More specifically, local features are more related to shallow model layers

while global features correspond to deeper model layers. The proposed method employs layer-wise aggregation to achieve higher performance for personalized model training. A hypernetwork [Ha *et al.*, 2016] for each client is employed on the server to generate the aggregation weight for each layer of different clients. Clients with a similar data distribution have higher weights for aggregation. FedDist algorithm [Sannara *et al.*, 2021] uses distances between neurons to measure how different neurons are between the global and local models. First, the server performs weighted averaging which is similar to the generation of the initial aggregated model in FedAvg. Then, the pairwise Euclidean distance is computed for each neuron in a layer between the local models and the aggregated model to identify diverging neurons. If the Euclidean distance between a specific neuron in any local model and the aggregated model is above a predetermined threshold the neuron is added to the aggregated model in order to improve model generalization. Thus, neurons that are specific to clients are incorporated into the aggregated model.

A personalized FL algorithm based on local memorization has been proposed in [Marfoq *et al.*, 2022]. It combines an aggregated global model with a k-nearest neighbors (kNN) model on each client. The global model is employed to compute the shared representation used by the local kNN. Each client computes and stores a representation-label pair for each sample. At inference time, the client queries the representation-label pair to obtain its k-nearest neighbors based on the model distance. Finally, the personalized model for a sample is obtained by interpolating the nearest neighbor distribution with the global model.

Mutual Information (MI) between model parameters or gradients is another kind of useful statistical metric for evaluating local models. In [Uddin *et al.*, 2020], a novel FL mechanism is introduced that exploits MI for both the client-side weight update and the server-side aggregation. The clients' model weights are updated by minimizing the MI between their local models and the aggregated global model. To extract distinct information, the correlation between two models must be minimized, which leads to minimized MI between them. For the model aggregation step, clients send MI values between their local models and the global model to the server. The server defines local models that are either similar to other models (with too high MI values) or significantly different from others (with too low MI values) as outliers. The FL server ranks the uploaded MI values to select the top useful local models for aggregation. Model-contrastive FL (MOON) [Li *et al.*, 2021c] utilizes contrastive learning at the model level. MOON is built upon FedAvg which incorporates modifications in the local training phase. Its local objective is to decrease the distance between the representation learned by each local model and the global model while increasing the distance between the representation learned by the current local model and the previous local model.

Class imbalance in FL has been investigated in [Yang *et al.*, 2021] without having access to raw data in order to evaluate the importance of local models. The class distribution of clients can be obtained using their updated gradients, assuming that a balanced auxiliary dataset exists on the server for the classification problem. The correlation between the gra-

dients with respect to the corresponding classes brought by auxiliary data on the server and class distribution [Anand *et al.*, 1993] allows for the calculation of the class imbalance of each client using the Kullback-Leibler (KL) divergence. The statistics of class distributions are learned using Combinatorial Multi-Armed Bandit (CMAB) [Chen *et al.*, 2013], and the client selection process is considered as a CMAB problem in order to identify clients with minimal class imbalance.

4 Applications of Federated Evaluation

Federated evaluation can enhance FL performance from multiple aspects. In this section, we discuss the applications of federated evaluation results in FL to illustrate its importance.

4.1 Understanding Global Model

It is vital to evaluate the global model performance on test datasets during training to understand the performance of FL and determine the cut-off accuracy.

In the ideal scenario, the server holds the test dataset and can centrally evaluate the global model. However, in most cases, the test data is not available on the FL server. As a result, the server resorts to conducting federated evaluation on selected clients. However, randomly selecting testing clients can lead to data deviations from the target distribution and may result in biased testing results. In [Lai *et al.*, 2021], a method is proposed to evaluate data deviations from the global distribution before selecting clients to test the global model. If data characteristics are not available, the proposed method estimates the number of participants in a way that bounds the deviation. If data characteristics are provided, the method iteratively select clients with the most number of samples until pre-defined conditions are met.

4.2 Incentive Mechanism Design

FL is unsealed in the sense that clients can depart the system at any time. Without centrally owning data by the server, it is difficult to force clients to contribute their models. Due to computation and communication costs for model training, clients are inherently reluctant to contribute to FL altruistically without any rewards. Thus, incentive mechanisms are indispensable to motivate clients to participate in FL. Pioneering works have designed incentive mechanisms to prevent free-riders and encourage the contribution of high-quality models from clients by allocating rewards to clients in accordance with their contributions [Deng *et al.*, 2021; Zhang *et al.*, 2021a; Hu *et al.*, 2022a; Hu *et al.*, 2022b].

A critical challenge in designing incentive mechanisms is how to allocate rewards among clients based on their contributions in a reasonable manner. The evaluation results of local models can exactly guide the allocation of rewards to incentivize clients. There are multiple works that have investigated incentive mechanisms in FL. In a simple way, data-level approaches simply consider the data size to determine the reward of a client [Feng *et al.*, 2019]. More complicated methods can employ model distance metrics to design incentive mechanisms. Fair [Deng *et al.*, 2021] establishes a reverse auction mechanism in which clients submit their bids to the server for participating in FL. It formulates a learning

quality maximization problem (LQM) to maximize the sum of the qualities of all selected participants within the learning budget and uses a greedy algorithm to solve this problem. Another incentive mechanism based on reputation and reverse auction is proposed in [Zhang *et al.*, 2021a]. Participants are rewarded by combining their bids and reputation scores. The reputation of clients is evaluated based on a model distance metric. FIFL [Gao *et al.*, 2021] designs an incentive mechanism that rewards clients based on their contributions and reputations. It uses a model distance metric based on gradient similarity to evaluate clients' contributions. Some existing works use Shapley values to evaluate the contribution of clients and determine reward allocation [Wang *et al.*, 2020; Fan *et al.*, 2022; Ma *et al.*, 2021b]. In Refiner [Zhang *et al.*, 2021c], clients are rewarded based on their contributions which are evaluated using both a model distance metric and marginal performance loss.

4.3 Client Selection

Due to the limited processing capacity of the server in FL, only a limited number of clients can be selected to participate in FL at each global iteration. How to select clients to participate in FL is a crucial problem that can significantly influence the model utility [Yang *et al.*, 2021; Li *et al.*, 2021a; Zhao *et al.*, 2022; Zhang *et al.*, 2021b]. Advanced client selection schemes can be devised based on federated evaluation results. More specifically, clients that can contribute more valuable models should be selected with higher priority. A well-designed client selection scheme can not only improve the final model utility but also shorten the training time by expediting the convergence of FL [Soltani *et al.*, 2022]. For example, Oort [Lai *et al.*, 2021] selects clients based on their local model utility evaluated on clients when making client selection decisions. Similarly, FedSAE [Li *et al.*, 2021b] considers local training loss as a utility metric to select clients. In a data-level approach, the server can select clients based on statistical homogeneity and content diversity of their data [Li *et al.*, 2021a]. The grouping-based scheduling method proposed in [Ma *et al.*, 2021a], divides clients into several groups to complement their missing labels. Clients within the same group are selected for training. Some existing works develop model distance metrics for designing client selection schemes. Fair [Deng *et al.*, 2021] employs the loss reduction to select participants in a way that maximizes the sum of the qualities of all participants. Methods proposed in [Zhao *et al.*, 2022; Zhang *et al.*, 2021b] utilize the divergence of model parameters to evaluate the quality of local models.

4.4 Malicious Attack Detection

In FL, malicious clients may easily launch attacks such as poisoning attacks by tampering with data labels or model gradients to deteriorate model utility. Since data is not exposed by FL clients, malicious attack detection algorithms designed for conventional machine learning are not applicable for FL.

Malicious attackers in FL can be identified and excluded from the model aggregation through accurate and efficient evaluation of local models. FIFL [Gao *et al.*, 2021] employs

Evaluation Methods	Ref.	Applications	Key Ideas	Architecture
Data-level Evaluation	[Feng <i>et al.</i> , 2019]	Incentive Mechanism	Obtaining revenue based on data size	Centralized
	[Li <i>et al.</i> , 2021a]	Client Selection	Filtering clients based on their labels and selecting clients with high statistical homogeneity and content diversity	Centralized
	[Zhao <i>et al.</i> , 2018]	Client Selection	Pair-wise grouping mechanism to complement missing labels	Centralized
	[Lai <i>et al.</i> , 2021]	Understanding Global Model	Preserving the deviation target of the data formed by the participants from the global distribution	Decentralized
Model Utility	[Lai <i>et al.</i> , 2021]	Client Selection	Sampling based on the aggregate training loss computed locally on the client and duration of the training round	Centralized
	[Li <i>et al.</i> , 2021b]	Client Selection	Selecting clients based their training loss	Centralized
	[Chai <i>et al.</i> , 2020]	Client Selection	Dividing clients into tiers based on their response latencies and selecting tiers based on their accuracy	Centralized
	[Li <i>et al.</i> , 2021a]	Client Selection	Measuring the gradient upper bound norm of the aggregated global loss with respect to the pre-activation outputs	Centralized
	[Zhang <i>et al.</i> , 2021c]	Incentive Mechanism, attack Detection	Evaluating local models on the validation set and computing marginal loss	Decentralized
Shapley Values	[Kang <i>et al.</i> , 2020a]	Attack Detection	A Proof-of-Verifying consensus scheme to verify the quality of local models on test data	Decentralized
	[Wang <i>et al.</i> , 2020; Fan <i>et al.</i> , 2022]	Incentive Mechanism	Contribution evaluation based on Shapley Value	Centralized
	[Ma <i>et al.</i> , 2021b]	Incentive Mechanism	Contribution evaluation based on Shapley Value	Decentralized
Statistical Metric	[Zhao <i>et al.</i> , 2022]	Client Selection	Evaluating weight difference before and after a round of training	Centralized
	[Zhao <i>et al.</i> , 2021b]	Client Selection	Evaluating weight divergence between a local model and an auxiliary model	Centralized
	[Deng <i>et al.</i> , 2021]	Attack Detection	Statistical analysis of the similarity values between the local models and the global model	Centralized
	[Gao <i>et al.</i> , 2021]	Attack Detection, Incentive Mechanism	The measurement of the distance between the local and benchmark gradients, and the local and global gradients	Centralized
	[Wang <i>et al.</i> , 2019]	Incentive Mechanism	Computing the difference between the new global model and original one after deleting some data points	Centralized
	[Zhang <i>et al.</i> , 2021a]	Client Selection, Incentive Mechanism	Reputation-based evaluation based on the direction vector between the initial and final global model	Centralized
	[Kang <i>et al.</i> , 2019; Kang <i>et al.</i> , 2020b]	Client Selection, Incentive Mechanism, Attack Detection	Reputation-based evaluation using multiweight subjective logic based on attack detection mechanisms	Centralized
	[Che <i>et al.</i> , 2022]	Attack Detection	Evaluating the distance between the local gradients and the committee gradients	Decentralized
	[Huang <i>et al.</i> , 2021]	Personalization	Pair-wise similarity evaluation between clients	Centralized
	[Ma <i>et al.</i> , 2022]	Personalization	Layer-wise similarity evaluation for identifying aggregation weights	Centralized
	[Sannara <i>et al.</i> , 2021]	Personalization	Evaluating pairwise distances of neurons between the local models and the initial global model	Centralized
	[Marfoq <i>et al.</i> , 2022]	Personalization	Interpolating nearest neighbor distribution with the global model	Centralized
	[Uddin <i>et al.</i> , 2020]	Personalization	Client-side weight update and server-side aggregation based on Mutual Information (MI)	Centralized
	[Yang <i>et al.</i> , 2021]	Client Selection	Evaluating the class imbalance of clients using the Kullback-Leibler (KL) divergence	Centralized
	[Deng <i>et al.</i> , 2021]	Incentive Mechanism, Client Selection	Maximization of the sum of the qualities of local models based on loss reduction within the learning budget	Centralized

Table 1. A summary of existing federated evaluation works.

gradient similarity as a model distance metric to detect abnormal gradients and malicious attackers. In [Kang *et al.*, 2019; Kang *et al.*, 2020b], a client contributing a very low-quality model is considered as a malicious attacker and excluded from the aggregation process based on a model distance metric, i.e. the performance difference for IID datasets and gradient diversity of local models for non-IID datasets.

4.5 Personalized Federated Learning

It is well-known that data distribution across different clients is often non-IID, which can lead to poor generalization performance of the global model on all local data distributions [Li *et al.*, 2020; Tan *et al.*, 2022]. Personalized federated learning (PFL) is proposed to address this problem [Tan *et al.*, 2022]. For PFL, each individual client has a learning objective slightly different from other clients. Thus, each individual client seeks to closely collaborate with clients owning a more similar data distribution. How to learn the similarity between data distributions can be accomplished using federated evaluation. For example, in [Huang *et al.*, 2021], the distance between models is used as a metric to evaluate the similarity of data distributions on different clients. A personalized model aggregation algorithm is devised which enables each client to assign higher weights to more similar clients when aggregating models. Similarly, in a layer-wised personalized FL [Ma *et al.*, 2022], each layer from different clients is assigned different aggregation weights based on the similarity between the data distribution of clients.

5 Summary and Challenges

Federated evaluation is indispensable for FL to achieve high-performance models without accessing clients’ data. In Table 1, we summarize a number of existing federated evalua-

tion studies, which are categorized based on their approaches. Their applications, evaluation architectures, and the key idea of each work are briefly illustrated.

In spite of tremendous efforts made by existing works, there are several challenges calling for more significant novel contributions, which are summarized as follows.

1. Differentially private FL, which injects zero-mean noises to obfuscate exposed information [Zhou *et al.*, 2023], can greatly complicate federated evaluation. Differentially private (DP) noises will disturb the accurate evaluation of model quality. Meanwhile, the noise scale will be amplified by the number of model exposure times, which considerably restricts the number of times to evaluate a local model.
2. If federated evaluation tasks are offloaded to clients or third parties, it is difficult to guarantee that these evaluators will return genuine and accurate evaluation results. They can easily attack federated evaluation by returning falsified evaluation results.
3. For fully decentralized FL (DFL), clients contact each other in an ad hoc manner to exchange model parameters [Su *et al.*, 2022]. Without the coordination of a server, federated evaluation becomes more difficult since the collected information of each DFL client is very limited to fully support evaluation of models.
4. In online FL, data is continuously collected and generated by clients [Chen *et al.*, 2020b]. To conduct federated evaluation in online FL, it is required to continuously track the change of model evaluation results with the arrival of new data.

Acknowledgments

This work was supported by ARC DP210101723.

References

- [Agrawal *et al.*, 2003] Rakesh Agrawal, Alexandre Evfimievski, and Ramakrishnan Srikant. Information sharing across private databases. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 86–97, 2003.
- [Anand *et al.*, 1993] Rangachari Anand, Kishan G Mehrotra, Chilukuri K Mohan, and Sanjay Ranka. An improved algorithm for neural network classification of imbalanced training sets. *IEEE Transactions on Neural Networks*, 4(6):962–969, 1993.
- [Biswas *et al.*, 2019] Chandan Biswas, Debasis Ganguly, Dwai-payan Roy, and Ujjwal Bhattacharya. Privacy preserving approximate k-means clustering. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1321–1330, 2019.
- [Chai *et al.*, 2020] Zheng Chai, Ahsan Ali, Syed Zawad, Stacey Truex, Ali Anwar, Nathalie Baracaldo, Yi Zhou, Heiko Ludwig, Feng Yan, and Yue Cheng. Tifl: A tier-based federated learning system. In *Proceedings of the 29th International Symposium on High-Performance Parallel and Distributed Computing*, pages 125–136, 2020.
- [Che *et al.*, 2022] Chunjiang Che, Xiaoli Li, Chuan Chen, Xi-aoyu He, and Zibin Zheng. A decentralized federated learning framework via committee mechanism with convergence guarantee. *IEEE Transactions on Parallel and Distributed Systems*, 33(12):4783–4800, 2022.
- [Chen *et al.*, 2013] Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, pages 151–159. PMLR, 2013.
- [Chen *et al.*, 2020a] Yiqiang Chen, Xiaodong Yang, Xin Qin, Han Yu, Biao Chen, and Zhiqi Shen. Focus: Dealing with label quality disparity in federated learning. *arXiv preprint arXiv:2001.11359*, 2020.
- [Chen *et al.*, 2020b] Yujing Chen, Yue Ning, Martin Slawski, and Huzefa Rangwala. Asynchronous online federated learning for edge devices with non-iid data. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 15–24. IEEE, 2020.
- [Deng *et al.*, 2021] Yongheng Deng, Feng Lyu, Ju Ren, Yi-Chao Chen, Peng Yang, Yuezhi Zhou, and Yaoyue Zhang. Fair: Quality-aware federated learning with precise user incentive and model aggregation. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2021.
- [Fan *et al.*, 2022] Zhenan Fan, Huang Fang, Zirui Zhou, Jian Pei, Michael P Friedlander, Changxin Liu, and Yong Zhang. Improving fairness for data valuation in horizontal federated learning. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 2440–2453. IEEE, 2022.
- [Feng *et al.*, 2019] Shaohan Feng, Dusit Niyato, Ping Wang, Dong In Kim, and Ying-Chang Liang. Joint service pricing and cooperative relay communication for federated learning. In *2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pages 815–820. IEEE, 2019.
- [Fung *et al.*, 2018] Clement Fung, Chris JM Yoon, and Ivan Beschastnikh. Mitigating sybils in federated learning poisoning. *arXiv preprint arXiv:1808.04866*, 2018.
- [Gao *et al.*, 2021] Liang Gao, Li Li, Yingwen Chen, Wenli Zheng, ChengZhong Xu, and Ming Xu. Fifi: A fair incentive mechanism for federated learning. In *50th International Conference on Parallel Processing*, pages 1–10, 2021.
- [Ghorbani and Zou, 2019] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, pages 2242–2251. PMLR, 2019.
- [Ha *et al.*, 2016] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.
- [Hu *et al.*, 2022a] Miao Hu, Di Wu, Yipeng Zhou, Xu Chen, and Min Chen. Incentive-aware autonomous client participation in federated learning. *IEEE Transactions on Parallel and Distributed Systems*, 33(10):2612–2627, 2022.
- [Hu *et al.*, 2022b] Miao Hu, Wenzhuo Yang, Zhenxiao Luo, Xuezheng Liu, Yipeng Zhou, Xu Chen, and Di Wu. Autofl: A bayesian game approach for autonomous client participation in federated edge learning. *IEEE Transactions on Mobile Computing*, 2022.
- [Huang *et al.*, 2021] Yutao Huang, Lingyang Chu, Zirui Zhou, Lan-jun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized cross-silo federated learning on non-iid data. In *AAAI*, pages 7865–7873, 2021.
- [Jia *et al.*, 2019] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J Spanos. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1167–1176. PMLR, 2019.
- [Kang *et al.*, 2019] Jiawen Kang, Zehui Xiong, Dusit Niyato, Shengli Xie, and Junshan Zhang. Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory. *IEEE Internet of Things Journal*, 6(6):10700–10714, 2019.
- [Kang *et al.*, 2020a] Jiawen Kang, Zehui Xiong, Chunxiao Jiang, Yi Liu, Song Guo, Yang Zhang, Dusit Niyato, Cyril Leung, and Chunyan Miao. Scalable and communication-efficient decentralized federated edge learning with multi-blockchain framework. In *International Conference on Blockchain and Trustworthy Systems*, pages 152–165. Springer, 2020.
- [Kang *et al.*, 2020b] Jiawen Kang, Zehui Xiong, Dusit Niyato, Yuze Zou, Yang Zhang, and Mohsen Guizani. Reliable federated learning for mobile networks. *IEEE Wireless Communications*, 27(2):72–80, 2020.
- [Kim *et al.*, 2019] Hyesung Kim, Jihong Park, Mehdi Bennis, and Seong-Lyun Kim. Blockchained on-device federated learning. *IEEE Communications Letters*, 24(6):1279–1283, 2019.
- [Lai *et al.*, 2021] Fan Lai, Xiangfeng Zhu, Harsha V Madhyastha, and Mosharaf Chowdhury. Oort: Efficient federated learning via guided participant selection. In *15th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 21)*, pages 19–35, 2021.
- [Li *et al.*, 2020] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.
- [Li *et al.*, 2021a] Anran Li, Lan Zhang, Juntao Tan, Yaxuan Qin, Junhao Wang, and Xiang-Yang Li. Sample-level data selection for federated learning. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2021.

- [Li *et al.*, 2021b] Li Li, Moming Duan, Duo Liu, Yu Zhang, Ao Ren, Xianzhang Chen, Yujuan Tan, and Chengliang Wang. Fedvae: A novel self-adaptive federated learning framework in heterogeneous systems. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. IEEE, 2021.
- [Li *et al.*, 2021c] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10713–10722, 2021.
- [Li *et al.*, 2022] Shuangtong Li, Tianyi Zhou, Xinmei Tian, and Dacheng Tao. Learning to collaborate in decentralized learning of personalized models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9766–9775, 2022.
- [Liu *et al.*, 2011] Yining Liu, Keqiu Li, Yingwei Jin, Yong Zhang, and Wenyu Qu. A novel reputation computation model based on subjective logic for mobile ad hoc networks. *Future Generation Computer Systems*, 27(5):547–554, 2011.
- [Ma *et al.*, 2021a] Jiahua Ma, Xinghua Sun, Wenchao Xia, Xijun Wang, Xiang Chen, and Hongbo Zhu. Client selection based on label quantity information for federated learning. In *2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pages 1–6. IEEE, 2021.
- [Ma *et al.*, 2021b] Shuaicheng Ma, Yang Cao, and Li Xiong. Transparent contribution evaluation for secure federated learning on blockchain. In *2021 IEEE 37th International Conference on Data Engineering Workshops (ICDEW)*, pages 88–91. IEEE, 2021.
- [Ma *et al.*, 2022] Xiaosong Ma, Jie Zhang, Song Guo, and Wen-chao Xu. Layer-wised model aggregation for personalized federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10092–10101, 2022.
- [Marfoq *et al.*, 2022] Othmane Marfoq, Giovanni Neglia, Richard Vidal, and Laetitia Kameni. Personalized federated learning through local memorization. In *International Conference on Machine Learning*, pages 15070–15092. PMLR, 2022.
- [McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguerre y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.
- [Qu *et al.*, 2020] Youyang Qu, Longxiang Gao, Tom H Luan, Yong Xiang, Shui Yu, Bai Li, and Gavin Zheng. Decentralized privacy using blockchain-enabled federated learning in fog computing. *IEEE Internet of Things Journal*, 7(6):5171–5183, 2020.
- [Sannara *et al.*, 2021] EK Sannara, François Portet, Philippe Lalande, and VEGA German. A federated learning aggregation algorithm for pervasive computing: Evaluation and comparison. In *2021 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–10. IEEE, 2021.
- [Shayan *et al.*, 2018] Muhammad Shayan, Clement Fung, Chris JM Yoon, and Ivan Beschastnikh. Biscotti: A ledger for private and secure peer-to-peer machine learning. *arXiv preprint arXiv:1811.09904*, 2018.
- [Soltani *et al.*, 2022] Behnaz Soltani, Venus Haghghi, Adnan Mahmood, Quan Z Sheng, and Lina Yao. A survey on participant selection for federated learning in mobile networks. In *Proceedings of the 17th ACM Workshop on Mobility in the Evolving Internet Architecture*, pages 19–24, 2022.
- [Su *et al.*, 2022] Dongyuan Su, Yipeng Zhou, and Laizhong Cui. Boost decentralized federated learning in vehicular networks by diversifying data sources. In *2022 IEEE 30th International Conference on Network Protocols (ICNP)*, pages 1–11. IEEE, 2022.
- [Tan *et al.*, 2022] Alysa Ziyang Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- [Uddin *et al.*, 2020] Md Palash Uddin, Yong Xiang, Xuequan Lu, John Yearwood, and Longxiang Gao. Mutual information driven federated learning. *IEEE Transactions on Parallel and Distributed Systems*, 32(7):1526–1538, 2020.
- [Wang *et al.*, 2019] Guan Wang, Charlie Xiaoqian Dang, and Ziye Zhou. Measure contribution of participants in federated learning. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 2597–2604. IEEE, 2019.
- [Wang *et al.*, 2020] Tianhao Wang, Johannes Rausch, Ce Zhang, Ruoxi Jia, and Dawn Song. A principled approach to data valuation for federated learning. In *Federated Learning*, pages 153–167. Springer, 2020.
- [Wu and Wang, 2022] Hongda Wu and Ping Wang. Node selection toward faster convergence for federated learning on non-iid data. *IEEE Transactions on Network Science and Engineering*, 2022.
- [Yang *et al.*, 2021] Miao Yang, Ximin Wang, Hongbin Zhu, Haifeng Wang, and Hua Qian. Federated learning with class imbalance reduction. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 2174–2178. IEEE, 2021.
- [Zhang *et al.*, 2020] Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M Alvarez. Personalized federated learning with first order model optimization. *arXiv preprint arXiv:2012.08565*, 2020.
- [Zhang *et al.*, 2021a] Jingwen Zhang, Yuezhou Wu, and Rong Pan. Incentive mechanism for horizontal federated learning based on reputation and reverse auction. In *Proceedings of the Web Conference 2021*, pages 947–956, 2021.
- [Zhang *et al.*, 2021b] Wenyu Zhang, Xiumin Wang, Pan Zhou, Weiwei Wu, and Xinglin Zhang. Client selection for federated learning with non-iid data in mobile edge computing. *IEEE Access*, 9:24462–24474, 2021.
- [Zhang *et al.*, 2021c] Zhebin Zhang, Dajie Dong, Yuhang Ma, Yilong Ying, Dawei Jiang, Ke Chen, Lidan Shou, and Gang Chen. Refiner: A reliable incentive-driven federated learning system powered by blockchain. *Proceedings of the VLDB Endowment*, 14(12):2659–2662, 2021.
- [Zhao *et al.*, 2018] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.
- [Zhao *et al.*, 2022] Jianxin Zhao, Xinyu Chang, Yanhao Feng, Chi Harold Liu, and Ningbo Liu. Participant selection for federated learning with heterogeneous data in intelligent transport system. *IEEE Transactions on Intelligent Transportation Systems*, 2022.
- [Zhou *et al.*, 2019] Zhi Zhou, Xu Chen, En Li, Liekang Zeng, Ke Luo, and Junshan Zhang. Edge intelligence: Paving the last mile of artificial intelligence with edge computing. *Proceedings of the IEEE*, 107(8):1738–1762, 2019.
- [Zhou *et al.*, 2023] Yipeng Zhou, Xuezheng Liu, Yao Fu, Di Wu, Jessie Hui Wang, and Shui Yu. Optimizing the numbers of queries and replies in convex federated learning with differential privacy. *IEEE Transactions on Dependable and Secure Computing*, 2023.