# Reliable and Interpretable Personalized Federated Learning

Zixuan Qin, Liu Yang*, Qilong Wang, Yahong Han, Qinghua Hu
College of Intelligence and Computing, Tianjin University
Tianjin Key Lab of Machine Learning, Tianjin, China
{qinzixuan1958@, yangliuyl@, qlwang@, yahong@, huqinghua@}tju.edu.cn

## Abstract

*Federated learning can coordinate multiple users to participate in data training while ensuring data privacy. The collaboration of multiple agents allows for a natural connection between federated learning and collective intelligence. When there are large differences in data distribution among clients, it is crucial for federated learning to design a reliable client selection strategy and an interpretable client communication framework to better utilize group knowledge. Herein, a reliable personalized federated learning approach, termed RIPFL, is proposed and fully interpreted from the perspective of social learning. RIPFL reliably selects and divides the clients involved in training such that each client can use different amounts of social information and more effectively communicate with other clients. Simultaneously, the method effectively integrates personal information with the social information generated by the global model from the perspective of Bayesian decision rules and evidence theory, enabling individuals to grow better with the help of collective wisdom. An interpretable federated learning mind is well scalable, and the experimental results indicate that the proposed method has superior robustness and accuracy than other state-of-the-art federated learning algorithms.*

## 1. Introduction

Federated learning is a new machine learning technique with various applications in data privacy protection and data security [19, 21, 37]. It can be viewed as social learning involving multiple agents coinciding with collective intelligence [3, 11, 13]. Unlike ordinary federated learning, personalized federated learning can address the problem of data heterogeneity among clients and thereby improve their capabilities in relatively more realistic scenarios [4, 18]. However, designing a reliable and interpretable federated learning framework remains a significant challenge in the field of federated learning. FedProx [32], SCAFFOLD [15], MOON [17] used the global model to impose different constraints on the client's local training process. Consequently, the knowledge of the global model was better absorbed.

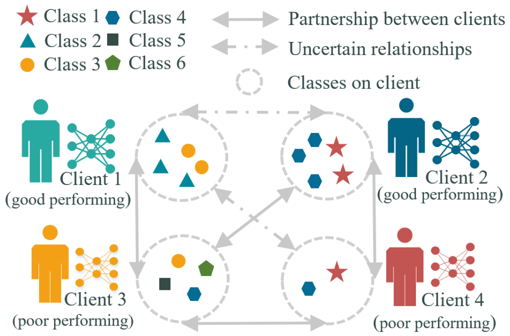


Figure 1. Cooperation between clients. Uninterpretable simple aggregation produces a global model that is not helpful for all clients because the information containing classes 2 and 3 may be negative for client 4, as well as for clients 1 and 4. Clients 1 and 2 can well identify classes 1, 2, 3, and 4; however, unreliable random selection does not guarantee the participation of clients 1 and 2 in aggregation, whereas the simultaneous selection of less-capable clients such as clients 3 and 4 does. In this case, smart customers cannot offer more help to the not-so-smart ones.

Although [6, 33, 40] solved the problem of client heterogeneity in a personalized federated learning framework by incorporating techniques such as clustering and knowledge distillation. [22, 35] propose a certain degree of interpretable client aggregation from the perspective of client contribution to the group. However, their selection and training of clients are often unreliable and uninterpretable, resulting in uncertainty in the training process and a tendency to ignore synergies between clients when the number of clients is large and the data distribution widely varies. Consequently, the collective intelligence is underutilized, as shown in Fig. 1.

Herein, we propose a reliable and interpretable federated learning framework, RIPFL. Specifically, we introduce Dempster–Shafer evidence theory (DST) [20, 34] to quan-

[*] Corresponding author.

tify the uncertainty and performance of each client and provide reliable client selection strategies. To reliably explain client choices and aggregation methods without wasting collective intelligence, RIPFL ensures that all smart clients participate in the aggregation process while a small number of nonsmart individuals participate, to enable that nonsmart clients can adequately gain more valuable collective knowledge from smart clients. Moreover, a method that can reliably integrate social and personal information is proposed. The proposed framework is primarily applicable to situations where the number of clients is large and the tasks solved by clients are complex. The main contributions of this paper are as follows.

- This paper proposes a reliable and interpretable personalized FL architecture from the perspective of social learning, which consists of interpretable local training, reliable clients selection and division, and effective federated aggregation.
- To reliably select the required clients, this paper introduces evidence theory to the local training of clients, thus quantifying the uncertainty of each client and providing interpretable training methods.
- A Bayesian-rule-based evidence fusion method is introduced by considering the global model as the prior information of clients when there are differences in the data distribution among clients. Consequently, the knowledge of the global model is not forgotten by clients in local training.

# 2. Background and Related Work

## 2.1. Personalized Federated Learning

As personalized federated learning shifts from traditional server-centered federated learning to each-client-centered federated learning, it is better suited for cases with a Non-IID data distribution. FedProx [32] restricted the local update direction by imposing a global constraint on the client training loss. [6, 15] placed different constraints on client losses based on [32]. MOON [17] utilized the similarity between model representations to correct the local training of clients. FedPer [4] aggregated the entire model base layer while retaining the last few layers as personalized layers for local updates. Clustered federated learning [33] aggregated individuals with similar data distributions among clients for training. [22] proposed APPLE, a cross-silo FL framework that adaptively learns how much each client benefits from other clients' models. Recently, knowledge distillation [1,38,40] has been widely applied to federated learning, [5, 23, 36, 39] solving the problem of data heterogeneity among clients from different perspectives using different technologies. The federated learning framework proposed herein is also personalized with the objective of allowing each individual to improve their performance.

## 2.2. Collective Intelligence

Collective intelligence, a shared or group intelligence, is a process of acquiring the opinions of numerous people and transforming them into decisions [13, 31]. In groups, individuals often simultaneously select different learning strategies [16,28], including social and asocial learning. Social learning requires extensive decision-making knowledge from other individuals and can be understood as 'copying' behavior to some extent, whereas asocial learning is more self-reliant. Better-performing individuals primarily engage in asocial learning, whereas less-capable individuals tend to improve their abilities by replicating the knowledge of better-performing learners through extensive social learning. [2, 3, 11, 14] reported the use of various conditions to better stimulate collective intelligence. Herein, we interpret the federated learning framework from the perspective of collective intelligence.

## 2.3. Dempster-Shafer Evidence Theory

Dempster-Shafer evidence theory (DST) is a generalization of Bayesian theory to subjective probabilities [7] and has been applied to reliably quantify uncertainty [10, 20]. Subjective logic formalizes the notion of evidence distribution of DST in a discriminative framework as the Dirichlet distribution [12]. In the classification task, for the input, the sample is classified as having a certain belief mass supported by each class. For the class matching the true label of the sample, belief mass should be increased for more credible judgments [34]. Herein, evidence theory is used to quantify the uncertainty in client classification tasks, and evidence generated by different models is fused to better learn group knowledge.

# 3. Methodology

## 3.1. Interpretable Objectives

RIPFL aims to enable clients to more reliably select more appropriate partners to work with, thereby leveraging interpretable learning strategies and achieving improved training results. The overall framework of this method is shown in Figure 2. [8, 9] proposed that interpretable machine learning should be reliably interpretable from data to the model; hence, to better illustrate the interpretability of the proposed method, we first present our objective.

Let $Q$ be an evaluation metric for interpretability and $E$ be a specific method for achieving interpretability. For a single client training, we need to find a method $E$ that enables the client $C$ to solve the task $T$ on the basis of the given data $D$ while having the maximum understanding of the model $M$, then the interpretable objective is

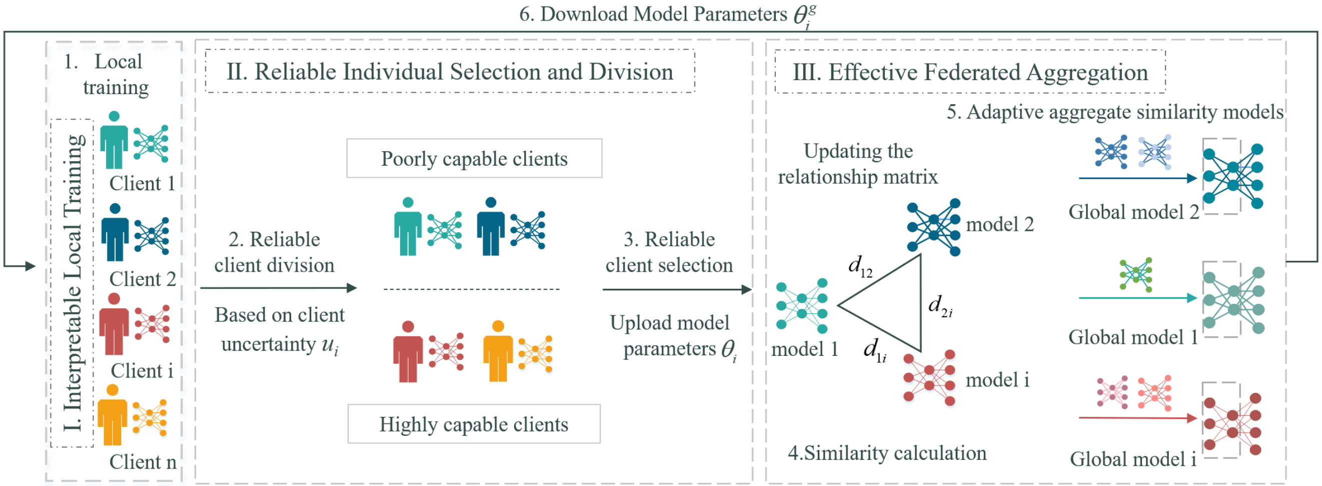

6. Download Model Parameters $\theta_i^g$
1. Local training
I. Interpretable Local Training
Client 1
Client 2
Client i
Client n
II. Reliable Individual Selection and Division
2. Reliable client division
Based on client uncertainty $u_i$
Poorly capable clients
Highly capable clients
3. Reliable client selection
Upload model parameters $\theta_i$
III. Effective Federated Aggregation
5. Adaptive aggregate similarity models
Updating the relationship matrix
model 2
model 1
model i
$d_{12}$
$d_{2i}$
$d_{1i}$
4.Similarity calculation
Global model 2
Global model 1
Global model i

Figure 2. Framework of RIPFL. Each global iteration round comprises three parts and six steps. After the clients complete their local training, the server reliably selects and divides them. Then, clients upload model parameters, select partners interpretably by social learning strategies, and perform model aggregation. In step 3, all clients with low uncertainty are selected, and some clients with high uncertainty are randomly selected. In step 6, we improve the parameter passing method of [4] for application to the scenarios in this paper. For most iteration rounds, the model is downloaded by passing the parameters of the front part (e.g., the layer inside the dashed box), and for a few rounds, the rear part of the model parameters containing the fully connected layer is passed.

$Argmax_E Q(E|M, C, D, T)$.

In personalized federated learning, $C_s$ denotes the $s$-th client involved in learning, $D_s$, $T_s$, $M_s$ represent the data, task and model of the $s$-th client respectively, then the optimization objective of personalized federated learning is

$$Argmax_{E_{FL}} \sum_{s=1}^{N} Q(E_{FL}|M_s, C_s, D_s, T_s), \qquad (1)$$

where $E_{FL}$ denotes the interpretable federated learning method required to solve the problem. So the process from client training to final aggregation is to seek an interpretable method such that $Q$ is maximum.

### 3.2. Interpretable Local Training

Let $E_{loc}$ denote the interpretable local training method. The interpretable objectives for each client in this period are $Argmax_{E_{loc},M_s} \sum_{s=1}^{N} Q(E_{loc}|M_s, D_s, T_s)$. At this stage, it is important to design reliable interpretable training models from the data to quantify the uncertainty. Thus, this paper adopts a reliable classification method based on DST [34]. For the classification problem to be solved by each client, the $k$-th class is given a belief mass $b_k$ to determine the sample labels, and an overall quality of uncertainty $u$ is given for each sample. For a $K$ classification problem, the $k$ belief mass and uncertainty are nonnegative and add to 1, i.e., $u + \sum_{k=1}^{K} b_k = 1$, where $u \geq 0$ and $b_k \geq 0$ for $k = 1, ..., K$. A belief mass $b_k$ for a class $k$ is computed using the class evidence. Let $e_k \geq 0$ be the evidence derived for the $k$-th class. Then, the belief $b_k$ and uncertainty $u$ are computed as $b_k = \frac{e_k}{S}$, $u = \frac{K}{S}$, where $S = \sum_{k=1}^{K}(e_i + 1)$. The evidence for the $k$-th class, $e_k$, is determined by the Dirichlet distribution parameter $\alpha$, i.e., $\alpha_k = e_k + 1$. Given an opinion, the expected probability for the $k$-th class is the mean of the corresponding Dirichlet distribution and computed as $\hat{p}_k = \frac{\alpha_k}{S}$. Clearly, the more evidence there is for a category, the smaller is the corresponding uncertainty. This explains how the model makes judgments about the probability of classification for each data sample from the perspective of subjective beliefs, while the evidence is generated with a complete understanding of the data.

Federated learning can produce adequate results and relies heavily on the client's use of group information such as global models [30]. However, during local iterative training, this global knowledge can significantly decrease or even disappear with increasing number of training rounds, especially when the data distribution between clients is very different. Therefore, to better fuse local and global information, the following Bayesian evidence fusion theorem [20] is introduced.

**Theorem 1** *Given the prior Dirichlet distribution* $p(\mathbf{z}|\boldsymbol{\beta}) = Dir(\mathbf{z}|\boldsymbol{\beta})$ *and distribution parameters collected from observed training samples* $\boldsymbol{\gamma}$*, the posterior distribution* $p(\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\gamma}) = Dir(\mathbf{z}|\boldsymbol{\beta} + \boldsymbol{\gamma})$.

Theorem 1 provides the basis for fusing two types of evidence information. The client receives a global model before local training and can generate global evidence $e_k^g$, which can be considered prior information of clients, using local data through the global model. Similarly, the local

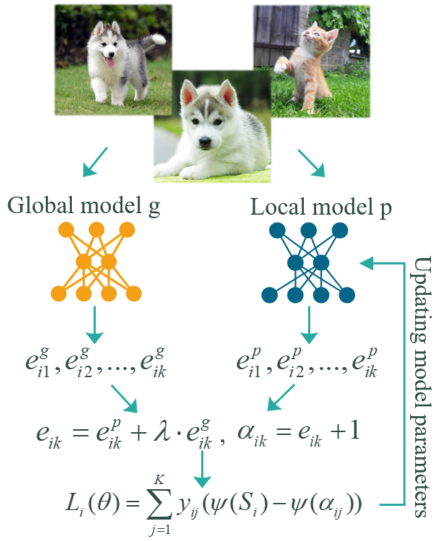

Global model g
Local model p
Updating model parameters
$e^g_{i1}, e^g_{i2}, \ldots, e^g_{ik}$
$e^p_{i1}, e^p_{i2}, \ldots, e^p_{ik}$
$e_{ik} = e^p_{ik} + \lambda \cdot e^g_{ik}$ , $\alpha_{ik} = e_{ik} + 1$
$L_i(\theta) = \sum_{j=1}^{K} y_{ij}(\psi(S_i) - \psi(\alpha_{ij}))$

Figure 3. Local training and evidence fusion process. After the client receives the global model, a prior evidence $e^g_k$ is generated using local sample $i$, and the global model is not involved in the update to maintain social information. Simultaneously, the client updates the individual model while generating the evidence $e^p_k$. Eventually, the two parts of evidence are fused, thereby generating evidence containing social and individual information and keeping the global model information continuously acting in the local training process.

model generates individual evidence $e^p_k$ after starting the update, and this part of evidence is acquired by observing the training sample. According to Theorem 1, for each sample posterior, evidence $e_k$ containing social and asocial information can be generated and calculated as

$$e_k = e^p_k + \lambda \cdot e^g_k, \tag{2}$$

where $k$ represents the class and the parameter $\lambda$ is used to control the ratio of prior evidence to locally observed evidence. Consequently, there is a different impact on the posterior evidence. The local training and evidence fusion process is shown in Fig. 3.

To make better use of the fused evidence while making model $M_s$ training interpretable, this paper introduces evidence classification loss. To obtain nonnegative outputs, the softmax layer of the normal DNN is replaced with an activation function layer (i.e., $exp(\cdot)$), and the output is used as an evidence vector to predict the Dirichlet distribution. For sample $x_i$, the evidence $e^p_k$ is obtained after the network as $e^p_k = exp(f(x_i|\theta))$, $\theta$ is the parameter of the client personal model. Similarly, the prior evidence $e^g_k$ generated by the global model $\theta_g$ can be obtained. According to Eq. (2), the fusion evidence $e_k$ and the posterior Dirichlet distribution parameter $\alpha_k$ can be calculated.

Generated posterior Dirichlet parameters can generate $Dir(p_i|\alpha_i)$ for a given sample $i$, where $p_i$ represents the class assignment probability on a simplex. After the modification of the ordinary cross-entropy loss, the loss for a single sample $i$ is calculated as follows:

$$\begin{aligned}\mathcal{L}_i^{edl}(\theta) &= \int \left[\sum_{j=1}^{K} -y_{ij}\log(p_{ij})\right] \frac{1}{B(\alpha_i)} \prod_{j=1}^{K} {p_{ij}}^{\alpha_{ij}-1} d\mathbf{p}_i \\ &= \sum_{j=1}^{K} y_{ij}(\psi(S_i) - \psi(\alpha_{ij})),\end{aligned} \tag{3}$$

where $y_{ij}$ is a one-hot vector encoding the ground-truth class of observation $x_i$ with $y_{ij} = 1$ and $y_{ij} = 0$ for all $k \neq j$, $S_i = \sum_{j=1}^{K} \alpha_{ij}$; $\psi(\cdot)$ is the digamma function. Eq. (3) means assigning the sum of all categories of evidence generated by the prediction to the true classes as much as possible and providing positive feedback. However, the above losses do not guarantee that incorrect labels will yield less evidence; hence, to shrink the evidence for incorrect labels to 0 to the maximum possible extent, the following KL dispersion term is introduced:

$$\begin{aligned}KL[D(\mathbf{p}_i|\tilde{\alpha}_i)||D(\mathbf{p}_i|\mathbf{1})] &= \log(\frac{\Gamma(\sum_{\mathrm{k}=1}^{\mathrm{K}} \tilde{\alpha}_{\mathrm{ik}})}{\Gamma(\mathrm{K}) \prod_{\mathrm{k}=1}^{\mathrm{K}} \Gamma(\tilde{\alpha}_{\mathrm{ik}})}) \\ &+ \sum_{k=1}^{K} (\tilde{\alpha}_{ik} - 1)[\log(\tilde{\alpha}_{\mathrm{ik}}) - \log(\sum_{\mathrm{j}=1}^{\mathrm{K}} \tilde{\alpha}_{\mathrm{ij}})],\end{aligned} \tag{4}$$

where $\mathbf{1}$ represents the parameter vector of $K$ ones, $\Gamma(\cdot)$ is the gamma function, $\tilde{\alpha}_i = \mathbf{y}_i + (1 - \mathbf{y}_i) \odot \alpha_i$ is the Dirichlet adjustment parameter that prevents excessive penalties for all evidence of the ground-truth class. Therefore, for a single sample $i$, the loss is:

$$\mathcal{L}_i(\theta) = \mathcal{L}_i^{edl}(\theta) + \lambda_t KL[D(\mathbf{p}_i|\tilde{\alpha}_i)||D(\mathbf{p}_i|\mathbf{1})], \tag{5}$$

where $\lambda_t$ is the balance parameter, and Eq. (5) can be considered an interpretable optimization method $E_{local}$. The optimization objective $F_s(\theta_s)$ for the $s$-client with data $D_s$ is as follows:

$$F_s(\theta_s) = \min_{\theta_s} \frac{1}{|D_s|} \sum_{i=1}^{|D_s|} \mathcal{L}_i(\theta_s), \tag{6}$$

where $\theta_s$ denotes the personal model parameters of the $s$-th client. For client $c$, we determine a way $E_{loc}$ to collect evidence from data $D_s$ that supports the choice decision. Consequently, we have an interpretable model $M_s$ when the client is faced with the classification task $T_s$. Finally, the goal of local training interpretability is reached.

### 3.3. Reliable Individual Selection and Division

Assume that $E_{sel}$ represents the interpretable method of client selection. The interpretable objectives of this period is $Argmax_{E_{sel},M_{sel}} Q(E_{sel}|M_{sel}, C_g)$, where $C_g$ denotes all clients participating in federated aggregation and $M_{sel}$ denotes the client selection model. In federated learning, clients are randomly selected to participate in aggregation. For client selection, the ratio of well-performing clients to poor-performing clients is not considered, as excessive poor-performing clients degrade the global information generated by aggregation and do not provide sufficient help to each client [13]. To solve this problem, RIPFL selects all intelligent clients, i.e., $C_1 = C_{better}$. Meanwhile, a fraction of unintelligent individuals is randomly selected to participate in the aggregation, i.e., $C_2 \in C_{poor}$. The clients involved in each global iteration are $C_g = C_1 \cup C_2$. Compared to normal random selection, this approach ensures that numerous better-performing individuals are involved in federated training each time, thus helping as many poor-performing individuals as possible. Simultaneously, the RIPFL determination of client performance is reliable because sample uncertainty can be determined by evidence theory, and the sum of the uncertainty $u_s$ of all samples on the client $s$ reliably reflects the client's ability (i.e., the smaller the $u_s$, the more intelligent is the individual).

### 3.4. Effective Federated Aggregation

As a unified global model does not help all clients in the same manner, providing all clients with the same global model is not the best solution. Let $E_{agg}$ denote the specific method for solving the interpretability of client aggregation; then, the interpretable objectives of this period are $Argmax_{E_{agg},M_{agg}} Q(E_{agg}|M_{agg}, C_g, T)$, where $T = \sum_{s=1}^{N} T_s$ and $M_{agg}$ is the federated aggregation model. From collective intelligence, each client selects individuals who are well suited to their own tasks and models to better help themselves [3, 11]. The model similarity $d_{sk}$ between clients $s$ and $k$ is calculated as $d_{sk} = cos(\theta_s, \theta_k)$, where $\theta_s$ is the model parameter, and $cos(a, b)$ denotes the calculation of the cosine similarity between a and b. This value can be considered the similarity of model $M_s$ and task $T_s$ contained in the individual. It is updated as the iteration proceeds and stored in the relationship matrix.

For client $s$, the sequence of similarity of the remaining clients participating in this round of federated aggregation to themselves is $v_s = [d_{s1}, d_{s2}, \ldots, d_{sk}]$ from largest to smallest, where client $k$ must satisfy $C_k \in C_g$. After the sequence $v_s$ is sorted, the first $c$ client models that are more similar to the $s$-client are selected from the sequence after elimination for aggregation, and the aggregated global model $\theta_s^g$ is sent to client for the next round of iterations. i.e., $\theta_s^g = \frac{1}{c}\sum_{k=1}^{c} \theta_k$. Here, $c$ can be a fixed constant adjusted to different group sizes; however, the introduction of uncertainty values allows $c$ to be adaptively adjusted to the performance of different individuals. For individuals, higher uncertainty implies that they are less capable of solving the task and therefore require more group information to help themselves. By contrast, individuals with lower uncertainty are more capable and require less group information. Therefore, the number of similar models selected by the client is proportional to the uncertainty $u_s$, i.e., $c = \zeta(u_s)$, where $\zeta(\cdot)$ is the ordinary mapping function. As the value of $u_s$ ranges from 0 to 1, mapping operation is required to convert it to a larger value based on the real group size.

This flexible and interpretable aggregation approach $E_{agg}$ is not limited by a single global model. It ensures that each client can aggregate a different amount of social information to generate a personalized global model that is beneficial to them, depending on their performance. Finally, the global optimization objectives are as follows.

$$\min_{\theta} F(\theta) = \min_{\theta_s, s\in[C]} F(\theta_1, \ldots, \theta_N) = \min_{\theta_s, s\in[C]} \sum_{s=1}^{N} F_s(\theta_s) \tag{7}$$

Following the $t$-th global iteration, $\theta_s^{(t+1)} = \theta_s^{g\,(t)}$, where $\theta^{(t)}$ is the model parameter for the $t$-th round. The complete federated frame comprises local training, client selection, and model aggregation; hence, the interpretable method $E_{FL} = E_{loc} \cup E_{sel} \cup E_{agg}$, and up to this point, we complete the interpretable goal proposed in Eq. (1).

### 3.5. Generalization Bound for RIPFL

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the input and output spaces, respectively. This paper focuses on the multiclassification problem, where $\mathcal{Y}$ is a finite set of classes. Consider a hypothesis of the form $h : \mathcal{X} \to \Delta^{\mathcal{Y}}$, where $\Delta^{\mathcal{Y}}$ represents the simplex over $\mathcal{Y}$. Thus, $h(x)$ is the probability distribution that can be assigned to the class on $x \in \mathcal{X}$. According to Eq. (5), the loss of $h \in \mathcal{H}$ for a labeled sample $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is given by $\mathcal{L}(h(x), y)$, where $\mathcal{H}$ denotes a family of such hypotheses $h$ with $VC$-dimension d. We denote the risk and empirical risk of hypothesis $h$ to distribution $\mathcal{D}$ and empirical distribution $\widehat{\mathcal{D}}$ as $\epsilon_{\mathcal{D}}(h)$ and $\widehat{\epsilon}_{\mathcal{D}}(h)$; $\epsilon_{\mathcal{D}}(h)$ is denoted as $\epsilon_{\mathcal{D}}(h) = \mathbb{E}_{\mathcal{D}}[\mathcal{L}(h(x), y)]$.

RIPFL client local training introduces interpretable models $m \in \mathcal{M}$, where $\mathcal{M}$ is a class of interpretable models. However, not every client model is sufficiently simple to be well interpreted on its own data. Therefore, $\widehat{\epsilon}_{\mathcal{D}_s}(m)$ is used to denote the empirical risk that $m$ cannot be interpreted using complex models in the face of complex data. The data distribution of client $s$ is denoted by $\mathcal{D}_s$, where $n_s$ is the number of samples. For client $s$, $m_s$ is the client model. According to standard learning–theoretic tools [26], with a probability of at least $1 - \delta$, the minimizer of empirical risk and risk satisfies

$$\epsilon_{\mathcal{D}_s}(h) \le \widehat{\epsilon}_{\mathcal{D}_s}(h) + 2\sqrt{\frac{2d\log(2n_s) + \log(4/\delta)}{n_s}} + \widehat{\epsilon}_{\mathcal{D}_s}(m_s). \tag{8}$$

For two distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ and a class of hypotheses $\mathcal{H}$, $d_{\mathcal{H}}(\mathcal{D}_1, \mathcal{D}_2)$ denotes the divergence between two distributions. In the federation system, multiple clients are involved in training, and the mixture of all client distributions is the global distribution. However, in the framework proposed herein, each client selects models similar to its own and uses its own aggregation model, instead of using a unified global model after the aggregation of all clients, i.e., $m_s^{(t+1)} = m_s^{g(t)}$, $\mathcal{D}_s^g = \sum_{s=1}^{\zeta(u_s)} \frac{1}{\zeta(u_s)} \mathcal{D}_s$, $\zeta(u_s) \leq N$; $m^{(t)}$ is the model for the $t$-th round. The mixture distribution then corresponds to the global hypothesis $h_s^g = \sum_{s=1}^{\zeta(u_s)} \frac{1}{\zeta(u_s)} h_s$, where $h_s$ is a hypothesis for the $s$-th client. Based on the global model generalization proposed by [24, 27], with a probability of at least $1 - \delta$, the risk of a single client in the federation system can be given as

$$\begin{aligned} \epsilon_{\mathcal{D}_s}(h_s) \leq \widehat{\epsilon}_{\mathcal{D}_s}(h_s^g) + 2\sqrt{\frac{2d\log(2n_s^g) + \log(4/\delta)}{n_s^g}} \\ + d_{\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_s^g) + \widehat{\epsilon}_{\mathcal{D}_s}(m_s), \end{aligned} \tag{9}$$

where $n_s^g$ is the total number of samples of the clients participating in the aggregation for client $s$. According to Eq. (7), the final optimization objective is

$$\min \frac{1}{N} \sum_{s=1}^{N} \epsilon_{\mathcal{D}_s}(h_s). \tag{10}$$

Assuming that $n = \sum_{s=1}^{N} n_s$ denotes the total number of samples, combining Eq. (9) and Eq. (10) yields the generalization bound for the entire framework as

$$\begin{aligned} \sum_{s=1}^{N} \frac{1}{N} \min_{h_s \in \mathcal{H}} \epsilon_{\mathcal{D}_s}(h_s) \leq \sum_{s=1}^{N} \frac{1}{N} \widehat{\epsilon}_{\mathcal{D}_s}(h_s^g) + \sum_{s=1}^{N} \frac{1}{N} \widehat{\epsilon}_{\mathcal{D}_s}(m_s) \\ + \sum_{s=1}^{N} \frac{1}{N} d_{\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_s^g) + 2\sqrt{\frac{2d\log(2n) + \log(4/\delta)}{n}}. \end{aligned} \tag{11}$$

# 4. Experiment

In this section, through specific experiments on two different real data sets, we verify that the proposed RIPFL method has better robustness.

## 4.1. Implementation Details

### 4.1.1 Compared Methods

FedAvg [25] is the classical federated learning algorithm. FedPer [4], FedProx [32], MOON [17], and Fed-Rod [6] are all personalized federated learning techniques, with the different local training losses employed to address the heterogeneity among clients being the difference between them. CFL [33] reduces the impact of irrelevant data for each client by clustering clients with similar data distributions such that multiple clustering centers coordinate for training. APPLE [22] is a cross-silo personalized FL framework where each client adaptively selects the aggregator to maximize its own benefits.

### 4.1.2 Datasets

The CIFAR10 and CIFAR100 datasets are widely used to test the ability of models in federated learning to accomplish classification tasks. Herein, we follow the data partitioning approach of [4]. The experiment controls the degree of variation in the data distribution by controlling the maximum number of sampled classes $\sigma$ for each client. The larger $\sigma$ is, the more classes a client can randomly receive.

### 4.1.3 Model and Hyperparameters

An experiment was conducted with PyTorch [29] to implement RIPFL and other baselines. A simple CNN network containing two convolutional layers and three fully connected layers is used for CIFAR10, and a ResNet18 network is used for CIFAR100. For CIFAR10, the number of local training rounds is 5, and the number of global rounds is 200. For cifar100, the number of local training rounds is 10, and the number of global rounds is 200. The Adam optimizer is used with a uniform learning rate of 0.0003.

## 4.2. Performance Comparison

### 4.2.1 Performance in Different Environments

The test accuracy in this experiment is the average of the accuracy of all clients that participated in federated learning on the local test dataset. Tab. 1 indicates that RIPFL produces better results in all cases; particularly, it outperforms other methods by at least 2.14% and 2.55% in the cases where $\sigma = 60$, the dataset is CIFAR100, and the numbers of clients are 30 and 50, respectively. This indicates that RIPFL has a more significant advantage as the task becomes complex and the number of clients increases.

With increasing discrepancy in data distribution between clients, the number of clients increases, number of classes per client increases, and number of samples per class decreases, FedPer causes the clients to rapidly overfit on the training set as a result of not passing the classification layer parameters during fine-tuning, resulting in a significant decrease in test accuracy. CFL exhibits poor performance in all cases because when the class overlap between clients is less and the sample size of each class is smaller, CFL needs to use most clients as cluster centers, which results in ineffective clustering. For the same case, the use of Fed-Rod with class balance loss to resolve the sample size of classes and class differences between clients is not feasible. APPLE reduces the communication overhead by limiting the

| Dataset | CIFAR10 | | | | | | CIFAR100 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Number of clients | $N = 30$ | | | $N = 50$ | | | $N = 30$ | | | $N = 50$ | | |
| Non-IID | $\sigma = 4$ | $\sigma = 5$ | $\sigma = 6$ | $\sigma = 4$ | $\sigma = 5$ | $\sigma = 6$ | $\sigma = 40$ | $\sigma = 50$ | $\sigma = 60$ | $\sigma = 40$ | $\sigma = 50$ | $\sigma = 60$ |
| FedAvg [25] | 78.19 | 74.75 | 71.27 | 75.49 | 72.42 | 71.14 | 65.97 | 63.89 | 61.29 | 62.45 | 59.56 | 56.41 |
| FedPer [4] | 78.31 | 75.32 | 72.45 | 76.88 | 74.81 | 71.05 | 60.05 | 55.20 | 50.94 | 50.05 | 46.15 | 43.70 |
| FedProx [32] | 76.38 | 73.58 | 70.16 | 74.32 | 72.02 | 70.75 | 67.27 | 64.41 | 62.03 | 62.94 | 60.42 | 58.10 |
| MOON [17] | 78.84 | 74.29 | 72.54 | 76.17 | 73.89 | 71.11 | 67.65 | 65.32 | 62.40 | 62.36 | 61.04 | 58.50 |
| CFL [33] | 64.33 | 67.73 | 67.48 | 57.23 | 60.37 | 60.05 | 57.10 | 56.93 | 56.38 | 50.77 | 51.48 | 52.81 |
| APPLE [22] | 77.14 | 72.64 | 69.58 | 70.48 | 67.17 | 66.22 | −− | −− | −− | −− | −− | −− |
| Fed-Rod [6] | 77.65 | 74.67 | 70.95 | 75.04 | 69.02 | 65.90 | 65.88 | 63.50 | 61.72 | 60.45 | 56.73 | 53.01 |
| **RIPFL** | **79.11** | **76.43** | **74.52** | **78.57** | **76.16** | **73.21** | **68.73** | **66.84** | **64.54** | **63.65** | **62.51** | **61.05** |

Table 1. Test accuracy (%) of different FL methods on CIFAR10 and CIFAR100, where $N$ denotes the number of clients. APPLE with a more complex network on a larger dataset would lead to a large overhead, and experiments are only performed on CIFAR10 owing to the limitations of the experimental equipment.

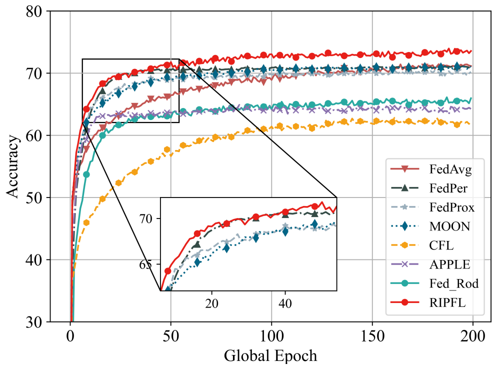


Figure 4. Convergence rate of each method on CIFAR10. $N = 50$ and $\sigma = 6$.

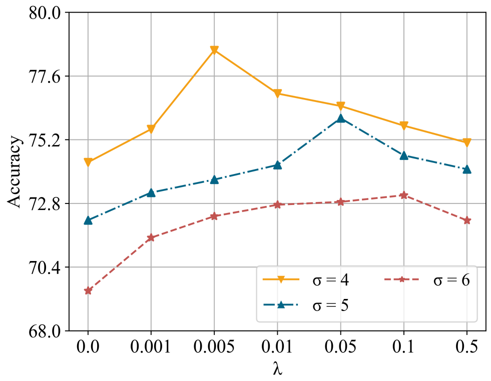


Figure 5. Performance of the RIPFL method on CIFAR10 for different $\lambda$ and $\sigma$ cases, $N = 50$.

number of core models each client can download from other clients. However, with increasing complexity of the task, the amount of social information required by the clients increases, and this limitation degrades the performance of the algorithm.

#### 4.2.2 Convergence Rate

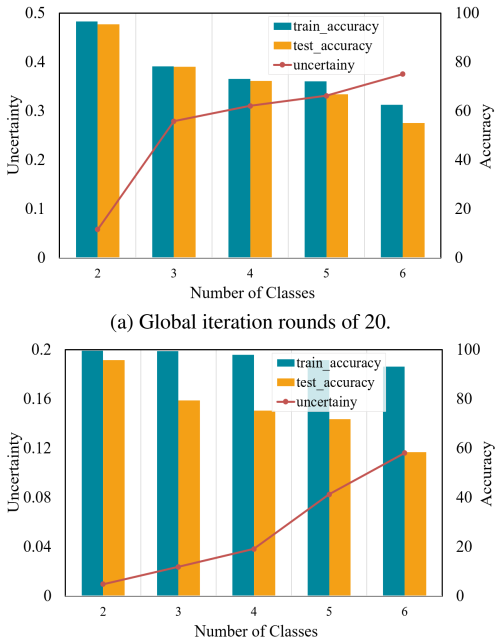
Fig. 4 reflects the convergence rate of each method within 200 rounds of communication. RIPFL achieves the best performance, followed by FedPer. FedPer converges faster than FedAvg, MOON, and FedProx because the base layer part of the model is uploaded during aggregation, while the final parts of the parameters are fine-tuned individually. RIPFL converges the fastest because it makes a reliable client selection while uploading partial parameters of the model. Consequently, poorly performing clients can gain knowledge from numerous good clients and thus have faster convergence.

### 4.3. Effect of Evidence Fusion

To show the influence of mixed evidence incorporating global prior evidence on the training effect, the experiments show the variation of accuracy with $\lambda$ for different $\sigma$. As shown inFig. 5, with increasing $\sigma$, the task becomes more difficult. The client requires help from stronger social information to facilitate social learning, and therefore, the parameter $\lambda$ to achieve the best performance increases with it. Clearly, if $\lambda$ is excessively small, clients tend to forget social knowledge during training and do not receive help from other clients, resulting in poor performance. If $\lambda$ is exceedingly large, the global model containing other clients' knowledge has excessive influence on the local model, consequently decreasing the accuracy rate.

### 4.4. Reliability Verification

To verify the reliability of client selection, the relationship between the number of classes included in the clients, accuracy, and client uncertainty is given. Fig. 6a shows that the uncertainty of clients varies with train/test set ac-



(a) Global iteration rounds of 20.

(b) Global iteration rounds of 60.

Figure 6. Train/Test accuracy (%), uncertainty, and number of classes included in the client on CIFAR10 with different global rounds ($N = 30, \sigma = 6$). The experiments randomly select one client from the clients containing different numbers of classes and show the variation of each metric in different training rounds.

curacy. Furthermore, they show a negative correlation, with the uncertainty decreasing with increasing accuracy. Fig. 6b shows that uncertainty can reliably distinguish the performance of different clients on the test set when the accuracy of the training set is generally high. The higher the number of client inclusion classes, the more complex the task, and the more likely it is to overfit as training proceeds, thus forgetting knowledge from the global model. As such clients with multiple classes tend to have intersection of classes with other clients, more social information is required. By contrast, clients with fewer classes have better classification ability and low uncertainty, and therefore, they do not require much social information.

Moreover, we conducted experiments of RIPFL and FedAvg with the attacks on client data. The accuracy of RIPFL dropped from 74.52% to 73.58% after being attacked, while the accuracy of FedAvg dropped from 71.27% to 68.86%. The performance of RIPFL decreases less than FedAvg, which demonstrates the robustness of RIPFL. From the perspective of aggregation, the uncertainty has been increased after being attacked, and the probability of being selected is reduced. From the perspective of personalization, only similar clients are selected for collaboration, which can also reduce the impact of attacks.

### 4.5. Interpretability Verification

| Client division | Acc% |
|---|---|
| well-performed(15); poor-performed(5) | **74.18** |
| well-performed(10); poor-performed(10) | 72.25 |
| well-performed(5); poor-performed(15) | 71.30 |

Table 2. Interpretability verification. The experiment was conducted on CIFAR10 with a number of 30 clients, from which 20 were selected to participate in the aggregation.

In this work, we account for interpretability with that uncertainty can be used to guide the grouping and selection. Specifically, clients with low uncertainty can obtain good performance, while the uncertainty of each client can be quantified by evidence theory during the local training, which is relevant to the performance. The results of the interpretability verification experiments are shown in Tab. 2. Clients are divided into well-performing (lower uncertainty) and poor-performing (higher uncertainty) groups, with 15 in each group. The numbers of well-performing and poor-performing are 15/5, 10/10, 5/15, respectively. It can be shown that accuracy is higher when more well-performing clients are selected, indicating that well-performing clients are more capable of helping others to improve the performance, while poor-performing clients may reduce the performance of the global model, clearly supporting the observation on our interpretability.

## 5. Conclusion

In this study, we developed a reliable and interpretable FL method (RIPFL) for the image classification task in the case of Non-IID data distribution among clients. We reliably quantified client uncertainty in training and designed interpretable client selection and aggregation methods that fully exploit group collaboration. Further, we introduced a Bayesian evidence fusion approach that allows social information to continue to work on local clients, enabling them to grow better with collective intelligence. The experimental results showed that the proposed model exhibited higher performance than state-of-the-art FL methods. The proposed FL framework is suitable for classification problems with large data distribution among customers, complex tasks of customers, and numerous clients.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China [No.62076179, 61925602, 62276186], the National Key R&D Program of China 2022ZD0116500.

## References

[1] Andrei Afonin and Sai Praneeth Karimireddy. Towards model agnostic federated learning using knowledge distillation. *Proceedings of the International Conference on Learning Representations*, 2022. 2

[2] Abdullah Almaatouq, Mohammed Alsobay, Ming Yin, and Duncan J. Watts. Task complexity moderates group synergy. *Proceedings of the National Academy of Sciences*, 2021. 2

[3] Abdullah Almaatouq, Alejandro Noriega-Campero, Abdulrahman Alotaibi, P. M. Krafft, Mehdi Moussaid, and Alex Pentland. Adaptive social networks promote the wisdom of crowds. *Proceedings of the National Academy of Sciences*, 2020. 1, 2, 5

[4] Manoj Ghuhan Arivazhagan, V. Aggarwal, Aaditya Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019. 1, 2, 3, 6, 7

[5] Ravikumar Balakrishnan, Tian Li, Tianyi Zhou, Nageen Himayat, Virginia Smith, and Jeffrey Bilmes. Diverse client selection federated learning via submodular maximization. In *Proceedings of the International Conference on Learning Representations*, 2022. 2

[6] Hong-You Chen and Wei-Lun Chao. On bridging generic and personalized federated learning for image classification. In *Proceedings of the International Conference on Learning Representations*, 2022. 1, 2, 6, 7

[7] Arthur P Dempster. *A Generalization of Bayesian Inference*. Springer Berlin Heidelberg, 2008. 2

[8] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017. 2

[9] Hugo Jair Escalante, Sergio Escalera, Isabelle Guyon, Xavier Baró, Yağmur Güçlütürk, Umut Güçlü, and Marcel van Gerven. *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer Cham, 2018. 2

[10] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *Proceedings of the International Conference on Learning Representations*, 2021. 2

[11] Bertrand Jayles, Hye rin Kim, Ramón Escobedo, Stéphane Cezera, Adrien Blanchet, Tatsuya Kameda, Clément Sire, and Guy Theraulaz. How social information can improve estimation accuracy in human groups. *In Proceedings of the National Academy of Sciences*, 2017. 1, 2, 5

[12] Audun Jøsang. Subjective logic: A formalism for reasoning under uncertainty. 2016. 2

[13] Tatsuya Kameda, Wataru Toyokawa, and R. Scott Tindale. Information aggregation and collective intelligence beyond the wisdom of crowds. *Nature Reviews Psychology*, 2022. 1, 2, 5

[14] Bhargav Karamched, Megan Stickler, William Ott, Benjamin Lindner, Zachary P. Kilpatrick, and Kre šimir Josić. Heterogeneity improves speed and accuracy in social networks. *Phys. Rev. Lett.*, 2020. 2

[15] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the International Conference on Machine Learning*, 2020. 1, 2

[16] Kevin N Laland. *Social learning strategies*. 2004. 2

[17] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2, 6, 7

[18] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *Proceedings of the International Conference on Machine Learning*, 2021. 1

[19] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *arXiv preprint arXiv:1908.07873*, 2019. 1

[20] Xiongkun Linghu, Yan Bai, Yihang Lou, Shengsen Wu, Jinze Li, Jianzhong He, and Tao Bai. Bayesian evidential learning for few-shot classification. *arXiv preprint arXiv:2207.13137*, 2022. 1, 2, 3

[21] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2021. 1

[22] Jun Luo and Shandong Wu. Adapt to adaptation: Learning personalization for cross-silo federated learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2022. 1, 2, 6, 7

[23] Zhengquan Luo, Yunlong Wang, Zilei Wang, Zhenan Sun, and Tieniu Tan. Disentangled federated learning for tackling attributes skew via invariant aggregation and diversity transferring. In *Proceedings of the International Conference on Machine Learning*, 2022. 2

[24] Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020. 6

[25] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2017. 6, 7

[26] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018. 5

[27] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *Proceedings of the International Conference on Machine Learning*, 2019. 6

[28] Lucas Molleman, Pieter van den Berg, and Franz J Weissing. Consistent individual differences in human social learning strategies. *Nature Communications*, 2014. 2

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In

*Proceedings of the International Conference on Neural Information Processing Systems*, 2019. 6

[30] Alfonso Perez-Escudero and Gonzalo de Polavieja. Collective animal behavior from bayesian estimation and probability matching. *Nature Precedings*, 2011. 3

[31] RonSun. *Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation*. Cambridge University Press, 2005. 2

[32] Anit Kumar Sahu, Tian Li, Maziar Sanjabi, Manzil Zaheer, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Third Conference on Machine Learning and Systems*, 2018. 1, 2, 6, 7

[33] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *In IEEE Transactions on Neural Networks and Learning Systems*, 2021. 1, 2, 6, 7

[34] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Proceedings of the International Conference on Neural Information Processing Systems*, 2018. 1, 2, 3

[35] Chengshuai Shi and Cong Shen. Federated multi-armed bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 1

[36] Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 2

[37] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. In *ACM Transactions on Intelligent Systems and Technology*, 2019. 1

[38] Lin Zhang, Li Shen, Liang Ding, Dacheng Tao, and Ling-Yu Duan. Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[39] Xu Zhang, Yinchuan Li, Wenpeng Li, Kaiyang Guo, and Yunfeng Shao. Personalized federated learning via variational bayesian inference. In *Proceedings of the International Conference on Machine Learning*, 2022. 2

[40] Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *Proceedings of the International Conference on Machine Learning*, 2021. 1, 2