FedNoRo: Towards Noise-Robust Federated Learning by Addressing Class Imbalance and Label Noise Heterogeneity

Nannan Wu[1] , **Li Yu**[1] , **Xuefeng Jiang**[2] , **Kwang-Ting Cheng**[3] and **Zengqiang Yan**[1*]

[1]School of Electronic Information and Communications, Huazhong University of Science and Technology

[2]Institute of Computing Technology, Chinese Academy of Sciences

[3]School of Engineering, Hong Kong University of Science and Technology

{wnn2000, hustlyu}@hust.edu.cn, jiangxuefeng21b@ict.ac.cn, timcheng@ust.hk, z_yan@hust.edu.cn

Abstract

Federated noisy label learning (FNLL) is emerging as a promising tool for privacy-preserving multi-source decentralized learning. Existing research, relying on the assumption of class-balanced global data, might be incapable to model complicated label noise, especially in medical scenarios. In this paper, we first formulate a new and more realistic federated label noise problem where global data is class-imbalanced and label noise is heterogeneous, and then propose a two-stage framework named FedNoRo[1] for noise-robust federated learning. Specifically, in the first stage of FedNoRo, per-class loss indicators followed by Gaussian Mixture Model are deployed for noisy client identification. In the second stage, knowledge distillation and a distance-aware aggregation function are jointly adopted for noise-robust federated model updating. Experimental results on the widely-used ICH and ISIC2019 datasets demonstrate the superiority of FedNoRo against the state-of-the-art FNLL methods for addressing class imbalance and label noise heterogeneity in real-world FL scenarios.

1 Introduction

Federated Learning (FL), allowing individual clients to train a deep learning model collaboratively without data sharing, has been widely studied in privacy-conscious occasions, e.g. medical [Kaissis *et al.*, 2020] and financial [Zheng *et al.*, 2021] applications. Most existing federated learning frameworks are based on the paradigm of fully supervised learning [McMahan *et al.*, 2017; Li *et al.*, 2020b], which implicitly assumes that each participant's data is labeled correctly. However, building a completely clean dataset with high-quality annotation is costly in realistic medical scenarios, as labeling medical data is time-consuming and labor-intensive requiring expertise. Consequently, it would unavoidably introduce noisy labels when hiring non-professionals to label or using automatic labeling techniques [Irvin and others, 2019].

Therefore, it is more common that the labels of some clients are clean while others are not in real-world FL scenarios.

Due to the existence of noisy clients, developing a noise-robust FL framework is of great importance, where accurately identifying the noisy clients is the first and the most crucial step. Existing methods for noisy client detection propose to calculate an average indicator (*e.g.* loss) over all samples of each client as its feature and filter out the clients with abnormal features as noisy clients, which assumes clean clients' features are independent and identically distributed (IID) while noisy clients' are outliers following the small-loss trick [Han *et al.*, 2018]. Specifically, Xu *et al.* [2022] calculated the average LID value [Houle, 2013] of each client and identified the clients with larger average LID values as noisy clients. Similarly, Wang *et al.* [2022] replaced the LID value with the confidence score and identified noisy clients with smaller confidence scores. Unfortunately, these indicators can be less effective to deal with noisy clients in real-world FL scenarios, due to the following observations:

1. Data is highly class-imbalanced from the global perspective [Wang *et al.*, 2021; Shang *et al.*, 2022]. Under class imbalance, using a global indicator for all classes is highly sensitive to clients' label distributions which may vary dramatically.

2. Data is heterogeneous across clients [Li *et al.*, 2020b]. As each client collects its own data independently under FL, there may exist severe data variations, affecting the calculation of indicators across clients.

3. Label noise is heterogeneous across clients where the heterogeneity of noise varies in both strength and pattern. The former represents different noise rates across clients, and the latter indicates various forms of label noise related to clients' local data distributions.

Suffering from class imbalance, the federated model in FL can bias to the global majority classes [Kang *et al.*, 2020; Zhou *et al.*, 2020; Cao *et al.*, 2019; Menon *et al.*, 2021], resulting in large indicator variations across classes and making noisy clients' indicators less distinguishable. For instance, in clinical scenarios, given one cancer-specialized hospital *A* and one general hospital *B*, *A* would enroll more malignant cases and *B* will enroll more healthy cases. Therefore, *A* is more likely to produce an abnormal client-wise feature (*e.g.*,

large loss values similar to noisy clients) due to class imbalance (*i.e.*, healthy $\gg$ malignant). In addition to data heterogeneity, the client-wise feature can also be affected by heterogeneous label noise, making noisy client detection harder. For instance, in clinical scenarios, given two hospitals C and D with different patient diversity (*i.e.*, various label distributions), one benign case is more likely to be wrongly diagnosed as healthy by C and malignant by D. Though both labels are wrong, the loss values (*i.e.*, indicators for noisy client detection) produced by C would be much smaller than D, due to class imbalance (*i.e.*, healthy $\gg$ malignant). In summary, class imbalance would divert the client-wise feature from being only related to whether a client is noisy to being related to both its local data distribution and label noise pattern, making the above identification methods based on global indicator struggle in realistic FL scenarios.

To address this, one straightforward way is to combine class-balancing [Kang *et al.*, 2020; Zhou *et al.*, 2020; Cao *et al.*, 2019; Menon *et al.*, 2021] with noisy label learning. However, those methods can only mitigate the bias in class prior probability while the bias in learning difficulty of class-specific features [Yi *et al.*, 2022] is unsolved. Due to relatively limited training data, the class-specific features of minority classes would be more difficult to learn compared to those of the majority classes, resulting in imbalance in the feature space. When adopting feature learning [Karthik *et al.*, 2021; Yi *et al.*, 2022] to alleviate the bias in learning difficulty, it can be constrained by each client's limited local data and computing resources. Till now, how to address class imbalance in federated noisy label learning (FNLL) is under-explored.

In this paper, we first formulate a new FNLL problem to model more realistic FL scenarios under class imbalance, and then propose a two-stage framework FedNoRo for noise-robust learning. Specifically, in the first stage, instead of using a global indicator, we propose to identify noisy clients according to the client-wise per-class average loss values calculated by a warm-up model trained by FedAvg [McMahan *et al.*, 2017]. As each class is considered independently, this detection method will be less affected by class imbalance and heterogeneity, leading to better detection performance. In the second stage, different learning strategies are employed for clean and noisy clients respectively, where cross-entropy loss is adopted for training on clean clients and knowledge distillation (KD) [Hinton *et al.*, 2015] is employed to minimize the negative influence of noisy labels on noisy clients. In addition to client-level training, a distance-aware aggregation function is proposed to better balance the importance of clean and noisy clients for global model updating in the server. In the local training phase of both the two stages, logit adjustment (LA) [Menon *et al.*, 2021] is imposed to fight against data heterogeneity and class imbalance. The main contributions are summarized as follows:

- A new FNLL problem where both class-imbalanced global data and heterogeneous label noise are considered to model real FL scenarios.

- A new label noise generation approach for multi-source data, where the synthetic label noise is heterogeneous

and instance-dependent.

- A two-stage FL framework, named FedNoRo, to address both class imbalance and label noise heterogeneity. In FedNoRo, noisy clients are identified based on abnormal per-class loss values, and noise-robust training strategies are used for effective federated model updating.

- Superior performance against the state-of-the-art FNLL approaches on real-world multi-source medical datasets.

2 Related Work

2.1 Federated Learning

Federated learning (FL) has drawn great attention for privacy-preserving applications [McMahan *et al.*, 2017]. Existing studies mainly focus on the challenges of FL applications, including data heterogeneity [Li *et al.*, 2020b; Karimireddy *et al.*, 2020; Li *et al.*, 2021a; Li *et al.*, 2021b; Yan *et al.*, 2020; Zhu *et al.*, 2021; Li and Zhan, 2021], data quality [Liu *et al.*, 2021; Liang *et al.*, 2022; Jiang *et al.*, 2022a], privacy protection [Agarwal *et al.*, 2018] and communication efficiency [Konečnỳ *et al.*, 2016; Sattler *et al.*, 2019].

2.2 Noisy Label Learning

Noisy label learning is of great importance as it is difficult to collect a large number of high-quality clean labels in real-world applications. Existing studies can be roughly divided into two types: selection-based [Han *et al.*, 2018; Yu *et al.*, 2019; Li *et al.*, 2020a; Wei *et al.*, 2022b; Zhao *et al.*, 2022; Yi *et al.*, 2022] and selection-free [Wang *et al.*, 2019; Zhang *et al.*, 2017; Wei *et al.*, 2022a; Lukasik *et al.*, 2020; Lukov *et al.*, 2022]. The former assumes that clean and noisy samples behave differently in a certain indicator (*e.g.* loss value) which can be used for identification. The latter focuses on designing special loss functions or regularizations that are robust to noisy labels.

2.3 Federated Noisy Label Learning

Federated noisy label learning (FNLL) is an emerging research topic. Chen *et al.* [2020] and Yang *et al.* [2021] leveraged clean datasets to assist FNLL, which may be infeasible in practice. RoFL [Yang *et al.*, 2022] updated local models with shared class-wise centroids to avoid inconsistent decision boundaries caused by noisy labels. FedLSR [Jiang *et al.*, 2022b] introduced regularization in the local training phase to alleviate the effect of noisy labels. Fang *et al.* [2022] combined FNLL with the heterogeneous model problem and proposed a noise-tolerant loss function and a client confidence re-weighting scheme. It treated all clients equally without explicitly filtering out noisy clients. In contrast, FedCorr [Xu *et al.*, 2022] and FedNoiL [Wang *et al.*, 2022] are selection-based methods, which first select noisy clients through designed indicators and then update the federated model by appropriate training strategies. Unfortunately, assuming class-balanced global data makes them less effective when dealing with realistic FL scenarios with class imbalance.
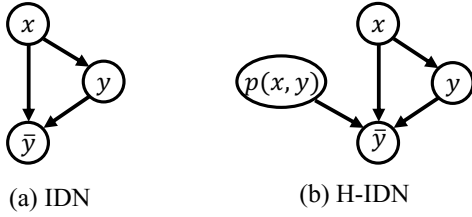
Figure 1: Graphical models of standard IDN and the proposed H-IDN.

3 Problem Definition

3.1 Preliminaries

Given a multi-class image classification task where $x_i \in \mathcal{X}$ is the input image and $y_i \in \mathcal{Y} = [C]$ is the corresponding ground-true label. The goal is to use a training dataset $D_{tr} = \{(x_i, y_i)\}_{i=1}^N$ of size N which is assumed to be class-imbalanced (*i.e.* $p_{tr}(y) \neq \frac{1}{C}$) to train a model $f(\cdot)$ for classification. In FL, D_{tr} can not be built directly due to privacy concerns, and it is divided into K non-overlapped subsets distributed over K clients. For any private dataset $D_k = \{(x_i, y_i)\}_{i=1}^{N_k}$ of size N_k in client $k \in [K]$, each image-label pair (x_i, y_i) obeys the private joint probability distribution $p_k(x, y) = p_k(y)p_k(x \mid y)$. Following [Shen *et al.*, 2021; Xu *et al.*, 2022], we assume the conditional distribution $p_k(x \mid y)$ is identical, while the local class prior $p_k(y)$ varies across clients (*i.e.* $p_{k_1}(y) \neq p_{k_2}(y)$ for $k_1 \neq k_2$) due to data heterogeneity. To simulate noisy labels, the ground-true label y in some clients is replaced by a noisy label $\overline{y}$.

3.2 Noise Model

To realistically model label noise in multi-source data, we construct a heterogeneous label noise model. In terms of the strength of label noise, we first define the global noise rate ρ as the proportion of noisy clients. Then, we assume local noise rate η_i (*i.e.*, the proportion of noisy samples) follows the uniform distribution $U(\eta^l, \eta^u)$ for any noisy client i as [Xu *et al.*, 2022]. For the pattern of label noise, we propose heterogeneous instance-dependent noise (H-IDN) for better modeling. As shown in Fig. 1 and Definition 1, local data distribution is an additional factor to manipulate label noise compared to standard instance-dependent noise (IDN) [Chen *et al.*, 2021].

Definition 1 (H-IDN). *IDN is heterogeneous if noise transition probability is a function of local data distribution, i.e.,* $\Pr(\overline{Y} = \overline{y} \mid Y = y, X = x) = M_{\overline{y}, y, x}(p_i(x, y))$, *where* $M_{\overline{y}, y, x}$ *denotes the noise transition matrix of instance* x.

In this way, the proposed H-IDN is capable of modeling more complicated and realistic label noise.

3.3 Noise Generation

To simulate the noise model in Sec. 3.2, we present a noise generator to produce heterogeneous label noise from multi-source clean data as described in Algorithm 1. Following [Chen *et al.*, 2021], neural networks are used as annotators to simulate the noisy labeling process in noisy clients. Specifically, we first randomly select ρK out of K clients as noisy

---

Algorithm 1 Noise Generation.

Input: Number of clients K; clean local datasets $\{D_k\}_{k=1}^K$; global noise rate ρ; local noise rate distribution parameters η^l, η^u.

1: $\mathcal{I}$ = Randomly select ρK elements from $[K]$.
2: **for** i in $\mathcal{I}$ **do**
3:     Initialize a network f_i.
4:     Train f_i on the local dataset $D_i = \{(x_j, y_j)\}_{j=1}^{N_i}$.
5:     Compute classification probabilities $p(Y \mid x) \in [0,1]^{N_i \times C}$ for all samples in D_i.
6:     Compute the misclassification probability $\widetilde{p}(x) \in [0,1]^{N_i}$ for each sample in D_i (Eq. 1).
7:     $\eta_i \sim U(\eta^l, \eta^u)$.
8:     $\mathcal{N}$ = Randomly select $\eta_i N_i$ elements from $[N_i]$ with the probability $\widetilde{p}(x) / \sum \widetilde{p}(x)$.   ▷ *Normalization*
9:     **for** t in $\mathcal{N}$ **do**
10:        $\overline{y}_t$ = Randomly select a different label from $\mathcal{Y}$ with the probability $p(Y \mid x_t)$.
11:        Flip y_t to $\overline{y}_t$.   ▷ *Add label noise*
12:     **end for**
13: **end for**

Output: Local datasets after adding label noise $\{D_k\}_{k=1}^K$.

---

clients and independently train a neural network based on each selected client's original data (*i.e.*, clean data). Then, the trained neural network is used to produce the classification probabilities of all samples belonging to the client. Given each instance x_j in the i-th noisy client and the corresponding classification probability $p(Y \mid x_j) \in [0,1]^C$, its misclassification probability is determined by

$$\widetilde{p}(x_j) = 1 - p(Y = y_j \mid x_j), \tag{1}$$

and totally $\eta_i N_i$ samples would be chosen as noisy samples based on the normalized misclassification probability $\widetilde{p}(x) \in [0,1]^{N_i}$. According to Eq. 1, hard samples are more likely to be selected as noisy samples. For each selected noisy sample x_j, it is assigned a noisy label $\overline{y}$ different from the original clean label y from the probability distribution $p(Y \mid x_j)$. In this way, noisy samples are more likely to be assigned with wrong labels of higher prediction confidence. As the noise of different clients is generated by different neural networks, it fits the definition of a heterogeneous noise pattern. Following the above generation process, generated noisy labels are more reasonable and challenging.

4 Methodology

To address class imbalance and label noise heterogeneity, we propose a two-stage framework named FedNoRo as illustrated in Fig. 2. In the first stage, a warm-up model is firstly trained based on FedAvg (*i.e.* steps 1 and 2), which is used to calculate the per-class loss values of each client (*i.e.* step 3). Then, client-wise per-class loss values are collected and processed in the server and used to identify noisy clients by Gaussian Mixture Model (GMM) (*i.e.* steps 4 and 5). According to the client identification results of the first stage, different training strategies are employed for clean and noisy clients respectively in the second stage. Specifically, the
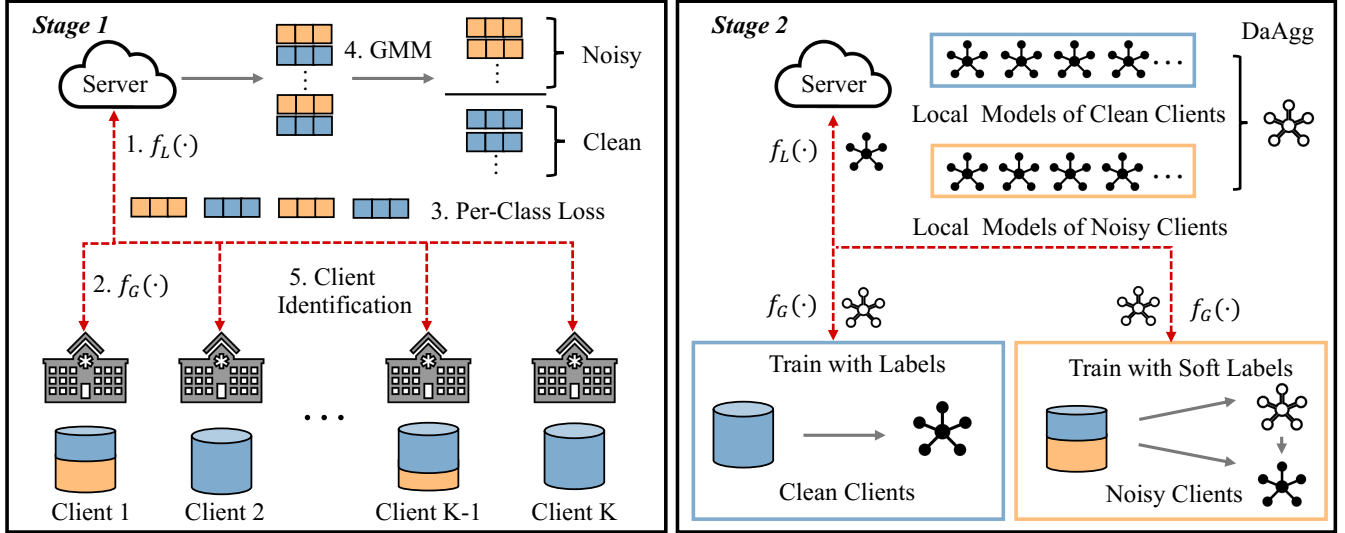
Figure 2: Overview of the proposed two-stage framework FedNoRo.

vanilla cross-entropy loss is adopted for clean clients while knowledge distillation (KD) is employed for better noisy label learning for noisy clients. Moreover, a distance-aware model aggregation function is defined to reduce the negative impact of noisy clients on global model updating. It should be noted that, in both stages, logit adjustment (LA) [Menon *et al.*, 2021] is introduced to local training to combat data heterogeneity and class imbalance.

4.1 Stage 1: Noisy Client Detection

As discussed in Section 1, class imbalance destroys the law that the client-wise average indicator (*e.g.* loss value) of clean clients is independent and identically distributed (IID), causing failure to identify noisy clients with client-wise indicators. To address this, the key is to re-find an IID indicator for clean clients that can detect noisy clients as outliers.

For the identification of noisy clients, we first train a warm-up model for T_1 (will be discussed in 5.4) rounds by FedAvg

$$w_g = \sum_{i=1}^{K} \frac{N_i}{\sum_{j=1}^{K} N_j} w_i, \qquad (2)$$

where w_g and w_i denote the weights of the global model and the i-th local model respectively and K is the number of clients for aggregation. Then, in the local training phase, vanilla cross-entropy loss is adopted to train the local model combined with LA as $f(\cdot) + \log \pi$, where π represents the local class prior (*i.e.* label distribution). LA is to make the local model treat each class equally rather than being biased, regardless of the biased class prior in local data (*i.e.* local class imbalance). In this way, different local models with data heterogeneity are more homogeneous by pursuing a unified learning goal regardless of their biased local class priors, which in turn helps model aggregation. Furthermore, as all local models participating

in the aggregation are unbiased, the global model is more likely to be unbiased. However, such an unbiased global model is proved to be unsuitable for noisy client identification through client-wise average indicators [Xu *et al.*, 2022; Wang *et al.*, 2022]. It is because, though LA is helpful to alleviate the bias in class priors, the bias in learning difficulty of class-specific features is unsolved [Yi *et al.*, 2022].

To complement this, per-class indicators are calculated, instead of using an averaged indicator for samples of all classes. It is based on our observation in clinical practice that, though inter-class samples are non-IID under class imbalance, intra-class samples are more IID. That's because medical imaging data is captured strictly following standard protocols. Given certain diseases, patients exhibit similar symptoms, and the imaging equipment required for diagnosis is strictly fixed, leading to images with the same or similar styles even at different hospitals. In other words, the IID assumption is satisfied by separately focusing on each specific class c on each client's private data. Therefore, given any class c, clients containing noisy samples in class c can be effectively identified as outliers (*e.g.*, larger loss values). In addition, under label noise heterogeneity, noisy samples across clients may be distributed to different noisy labels, and relying on any single class indicator is insufficient to identify noisy clients. Thus, the average loss values of all classes on each client i denoted as $l_i = (l_i^1, l_i^2, ..., l_i^C)^T \in \mathbb{R}^C$ are used for identification.

Given K per-class average loss values $L = [l_1, ..., l_K]$, noisy client detection is implemented in an unsupervised learning way. Considering the class-missing problem in heterogeneous data, a specific class c may not exist in client i, leading to a missing average loss value l_i^c in l_i. Such missing values in L cannot be used for unsupervised learning and must be filled. Actually, if class c is missing in client i, it will surely have no noisy samples in c. According to the small-

loss trick [Han *et al.*, 2018], the cleaner a sample is, the easier it is to learn and the smaller its loss is. Therefore, l_i^c is simply replaced by the minimum value of class c across all clients (*i.e.*, $\min_{i=1}^K l_i^c$). Given the bias of learning difficulty of different classes, loss value ranges across classes can vary dramatically, and the average loss values of all clients for any class should be uniformly normalized to $[0, 1]$ by

$$l_i^c = \frac{l_i^c - \min_i l_i^c}{\max_i l_i^c - \min_i l_i^c}. \tag{3}$$

To this end, all classes contribute equally to identifying noisy clients instead of being dominated by certain classes. Then, a two-component Gaussian Mixture Model (GMM) is deployed onto the processed client-wise per-class loss vectors, and K clients are partitioned into two subsets: S_c (*i.e.*, clean clients corresponding to the Gaussian distribution with a small norm mean vector) and S_n (*i.e.*, noisy clients corresponding to the Gaussian distribution with a large norm mean vector). It should be noted that this noisy client detection process is privacy-preserving as only average loss values are uploaded to the server.

4.2 Stage 2: Noise-Robust Training

After noisy client detection, customized training and aggregation strategies are proposed for clean and noisy clients respectively for noise-robust federated learning. Similar to the first stage, in the local training phase, LA is introduced to both clean and noisy clients to address data heterogeneity and global class imbalance. For clean clients, the vanilla cross-entropy loss is adopted to train each local model based on clean labels. For noisy clients, inspired by the label-soften-based method which has shown great potential in noisy label learning [Wei *et al.*, 2022a; Lukasik *et al.*, 2020; Lukov *et al.*, 2022], a knowledge distillation (KD) [Hinton *et al.*, 2015] based training method is applied. More specifically, given any x and its output logit from the global model $f_G(x)$, a targeted probability distribution is calculated as

$$y_G = \mathrm{softmax}(\frac{f_G(x)}{T}), \tag{4}$$

where T is the temperature to control the uniformness of y_G and is set as 0.8 by default in this paper. Then, the total training loss for each local model in noisy clients is written as

$$\mathcal{L} = \lambda \mathcal{L}_{KL}(y_p, y_G) + (1 - \lambda)\mathcal{L}_{CE}(y_p, \overline{y}), \tag{5}$$

where y_p represents the prediction results of the local model, $\mathcal{L}_{KL}$ is the Kullback-Leibler divergence, and λ is a trade-off coefficient. Inspired by semi-supervised learning [Laine and Aila, 2017], λ grows from 0 to λ_{max} (set as 0.8 by default) using a Gaussian ramp-up curve. Considering the memorization effect of neural networks on noisy datasets [Han *et al.*, 2018; Jiang *et al.*, 2022b], this strategy would be robust for noisy clients. In the early training stage, each local model tends to learn simple patterns (more likely to be clean samples), and the training process is dominated by the hard labels $\overline{y}$ and the cross entropy loss. As the neural network begins to fit complex patterns (more likely to be noisy samples), soft labels y_G gradually dominate the training process to avoid the model overfitting to noisy labels.

In the model aggregation phase, models from clean and noisy clients are of different importance. Hence, inspired by [Liang *et al.*, 2022], distance-aware model aggregation DaAgg is proposed where a client-wise distance metric is defined as

$$d(i) = \min_{j \in S_c} \|w_i - w_j\|_2, \tag{6}$$

w_i denotes the weights of the i-th local model. This metric measures the distance between a model w_i and the nearest model of clean clients. Note that $d(i)$ is equal to 0 if client i is clean. To ensure the boundness of $d(i)$, it is further normalized to $[0, 1]$ as $D(i) = \frac{d(i)}{\max_j d(j)}$. Then, local models are aggregated to update the global model by

$$w_g = \sum_{i=1}^K \frac{N_i e^{-D(i)}}{\sum_{j=1}^K N_j e^{-D(j)}} w_i. \tag{7}$$

In this way, the aggregation weight of any clean client i is constant due to $D(i) = 0$, while the weights of noisy clients are multiplied by a scaling factor in $[0, 1]$. Moreover, the weights would decrease as the model distance increases, adaptively adjusting the contributions of noisy clients in global model updating for noise-robust federated learning.

5 Experimental Results

5.1 Dataset and Evaluation Metric

Two widely-used datasets are included for evaluation:

1. ICH dataset [Flanders *et al.*, 2020] consisting of 67969 brain CT slices for the diagnosis of five intracranial hemorrhage (ICH) subtypes. The dataset is pre-processed following [Liu *et al.*, 2021; Jiang *et al.*, 2022a].

2. ISIC 2019 dataset [Tschandl *et al.*, 2018; Codella and others, 2018; Combalia *et al.*, 2019] containing 25331 images for the classification of eight skin diseases.

Both datasets are randomly divided into the training and test sets following a 7:3 split. As both datasets are highly class-imbalanced, the divided training and test sets are also class-imbalanced, satisfying the proposed setting. Considering the imbalanced test sets, the official ISIC Challenge 2019 metric, balanced accuracy (BACC), is adopted for evaluation.

5.2 Experimental Setup

Data Partition and Noise Generation

Following the previous work on FNLL [Xu *et al.*, 2022], we use Bernoulli distribution and Dirichlet distribution to generate heterogeneous data partitions. Specifically, we set $\alpha_{Dir} = 2.0, \Phi_{ij} \sim \mathrm{Bernoulli}(0.9), K = 20$ and $\alpha_{Dir} = 1.5, \Phi_{ij} \sim \mathrm{Bernoulli}(0.99), K = 20$ for the ICH and ISIC 2019 datasets respectively. For better visualization, we use dot size to represent the amount of data and plot the data distributions as shown in Fig. 3. For evaluation under various settings of label noise, ρ is set as 0.2, 0.3, 0.4, and 0.6, and (η^l, η^u) is set as (0.3,0.5) and (0.5,0.7) for the ICH dataset, while ρ is set as 0.4 and 0.6 and (η^l, η^u) is set as (0.5,0.7) for the ISIC 2019 dataset.

Category	Methods	ρ	0.0	0.2		0.3		0.4		0.6	
		(η^l, η^u)	(0.0, 0.0)	(0.3, 0.5)	(0.5, 0.7)	(0.3, 0.5)	(0.5, 0.7)	(0.3, 0.5)	(0.5, 0.7)	(0.3, 0.5)	(0.5, 0.7)
FL	FedAvg	Best	69.34	65.81	64.53	62.49	60.82	60.52	58.38	58.46	54.77
		Last	68.92	65.33	63.97	62.10	60.23	60.05	56.79	56.88	50.35
	FedProx	Best	68.16	64.58	63.64	61.81	61.91	60.85	58.50	60.21	57.57
		Last	67.29	63.80	62.65	61.46	61.00	59.99	57.85	59.35	56.63
	FedLA	Best	**73.56**	69.45	69.28	66.84	64.84	66.60	62.39	63.90	58.78
		Last	**73.07**	68.82	68.45	66.20	64.03	63.90	60.51	61.86	54.99
Denoise FL	RoFL	Best	-	42.57	40.42	40.64	39.87	40.35	35.60	39.68	35.35
		Last	-	42.19	40.30	40.95	39.27	40.25	35.39	40.18	35.47
	RHFL	Best	-	57.48	56.91	56.72	55.74	55.26	54.30	54.63	51.00
		Last	-	57.05	56.46	55.75	52.19	54.71	52.08	52.53	49.64
	FedLSR	Best	-	55.99	55.28	54.44	52.27	52.48	48.20	50.89	43.30
		Last	-	55.69	54.77	53.95	51.70	51.92	46.77	49.96	39.81
	FedCorr	Best	-	57.90	56.68	55.02	54.61	53.62	50.82	50.89	47.13
		Last	-	57.68	55.86	54.52	53.62	52.60	49.59	50.24	46.13
Joint	FedCorr+LA	Best	-	64.78	64.29	62.58	62.78	62.19	59.58	57.88	54.53
		Last	-	63.99	63.55	61.90	61.87	61.30	60.06	57.17	53.92
Ours	FedNoRo	Best	-	**70.59**	**70.64**	**70.14**	**69.35**	**70.69**	**69.30**	**67.55**	**63.83**
		Last	-	**70.18**	**69.81**	**69.29**	**68.47**	**70.14**	**68.87**	**67.10**	**63.29**

Table 1: Quantitative BACC (%) comparison results on the ICH dataset under different noise rates. The best results are marked in bold.
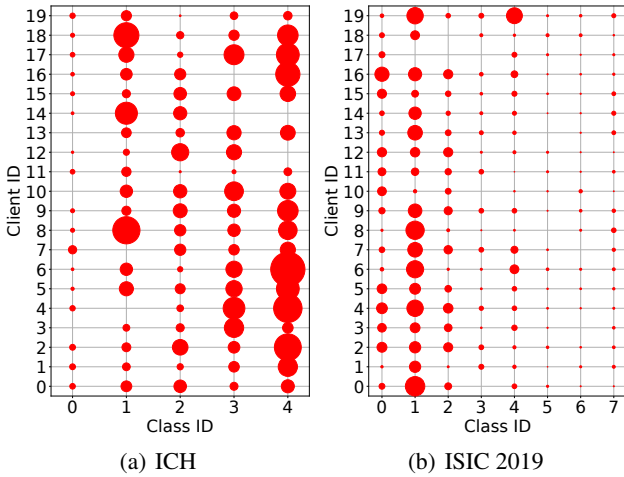

(a) ICH   (b) ISIC 2019

Figure 3: Visualization of data distributions of the two datasets.

Methods	ρ	0.0	0.4	0.6
	(η^l, η^u)	(0.0, 0.0)	(0.5, 0.7)	(0.5, 0.7)
FedAvg	Best	69.09	58.23	54.35
	Last	65.44	61.83	49.88
FedProx	Best	72.20	64.46	58.55
	Last	**69.60**	60.13	49.87
FedLA	Best	**72.55**	66.34	61.20
	Last	68.72	61.18	56.11
RoFL	Best	-	28.45	28.86
	Last	-	27.79	28.29
RHFL	Best	-	46.06	46.67
	Last	-	44.04	45.09
FedLSR	Best	-	30.15	27.24
	Last	-	29.11	26.08
FedCorr	Best	-	42.54	38.40
	Last	-	41.12	37.17
FedCorr+LA	Best	-	60.38	55.40
	Last	-	59.16	54.27
FedNoRo (ours)	Best	-	**68.59**	**66.00**
	Last	-	**64.67**	**60.65**

Table 2: Quantitative BACC (%) results on the ISIC 2019 dataset under different noise rates. The best results are marked in bold.

Implementation Details

ResNet-18 [He et al., 2016] is used as the backbone model for the ICH and ISIC 2019 dataset (pre-trained by ImageNet). In terms of federated settings, the maximal communication round is set as 100 and the local epoch is set as 5 for the ICH dataset while set as 300 and 1 respectively for the ISIC 2019 dataset. T_1 for warm-up training is set to 10 and 20 on the two datasets respectively. Other settings are the same for both datasets, including an Adam optimizer with momentum terms of (0.9, 0.999), a weight decay of 5e-4, a constant learning rate of 3e-4, and a batch size of 16.

5.3 Comparison with SOTA Methods

A series of state-of-the-art methods are selected for comparison, including conventional FL methods (*i.e.*, FedAvg [McMahan et al., 2017], FedProx [Li et al., 2020b], and FedLA [Menon et al., 2021]) and de-noise FL methods (*i.e.*, RoFL [Yang et al., 2022], RHFL [Fang and Ye, 2022], FedLSR [Jiang et al., 2022b], and FedCorr [Xu et al., 2022]). We also combine LA with FedCorr for comparison.

Quantitative results, including both the best BACC across all rounds (denoted as Best) and the averaged BACC over the last 10 rounds (denoted as Last), on the two datasets are summarized in Tabs. 1 and 2. FedLA generally achieves the best performance among conventional FL methods, benefiting from alleviating class imbalance. Meanwhile, all conventional FL methods would suffer from performance degradation with the increase in noise rate. Unfortunately, though designed for noisy label learning, the four de-noise FL methods fail to achieve performance improvement. Without a specific solution to class imbalance, label noise learning will be seriously affected by class imbalance, resulting in even worse performance compared to the baseline. It explains why introducing LA to FedCorr is beneficial. But as analyzed before, class imbalance cannot be completely eliminated by LA, simply combing LA with FedCorr is sub-optimal. Comparatively, FedNoRo consistently outperforms other methods by a large margin. Furthermore, the performance of FedNoRo is more stable with the increase in noise rate, demonstrating the noise-robust nature of FedNoRo.

Indicator	LA	Per-Class	Norm.	Re (%)	Pr (%)	MR (%)
ICH, $\rho = 0.3$, $(\eta^l, \eta^u) = (0.3, 0.5)$						
LID	✗	✗	✗	**100.00**	54.54	0.00
Conf.	✗	✗	✗	16.66	8.33	0.00
Loss	✗	✗	✗	94.58	57.12	0.00
Loss	✓	✗	✗	**100.00**	47.14	0.00
Loss	✗	✓	✗	97.19	90.63	85.77
Loss	✓	✓	✗	99.48	97.78	96.93
Loss	✓	✓	✓	99.70	**98.76**	**98.28**
ICH, $\rho = 0.4$, $(\eta^l, \eta^u) = (0.3, 0.5)$						
LID	✗	✗	✗	87.50	63.63	0.00
Conf.	✗	✗	✗	37.50	25.00	0.00
Loss	✗	✗	✗	78.80	60.57	0.00
Loss	✓	✗	✗	89.41	80.46	0.00
Loss	✗	✓	✗	65.77	**100.00**	37.58
Loss	✓	✓	✗	81.10	**100.00**	47.09
Loss	✓	✓	✓	**90.23**	**100.00**	**88.82**

Table 3: Ablation study of the first stage in FedNoRo.

5.4 Ablation Study

How Each Step in Noisy Client Detection Works
Additional experiments are conducted on the ICH dataset to analyze how each step in the first stage of FedNoRo affects noisy client detection as summarized in Tab. 3. Here, *Per-Class* is to calculate certain indicators separately for each class, and *Norm.* is to apply min-max normalization to the *Per-Class* indicator according to Eq. 3. For a fair comparison, the number of training rounds in the first stage of all methods is the same as Sec. 5.2. Given certain client-wise indicators under a specific setting, we repeatedly adopt GMM (10000 different random seeds) to identify noisy clients. Then, the average recall (*i.e.*, Re), precision (*i.e.*, Pr), and matching ratio (*i.e.*, MR: ratio of successfully-identified noisy clients) of 10,000 times of identifications are used to measure whether the method can effectively detect noisy clients.

As stated in Tab. 3, the first two methods from [Xu *et al.*, 2022] and [Wang *et al.*, 2022], together with the third one, completely ignore class imbalance, resulting in poor detection results. Introducing LA to address class imbalance is somewhat helpful for noisy client detection, leading to slight performance improvement. But it fails to improve the MR performance. Comparatively, when each class is considered independently, both noise heterogeneity and class imbalance are further alleviated, leading to better MR performance. When combing LA and *Per-Class* with min-max normalization, namely FedNoRo, most noisy clients are effectively detected, which lays a solid foundation for the second stage.

How Warm-Up Training Affects Noisy Client Detection
In the first stage, the global model warms up for T_1 rounds with FedAvg before noisy client detection. Due to the lack of prior information on noisy clients in realistic scenarios, it is hard to exactly determine T_1 for optimal performance. To evaluate the effect of T_1, noisy client detection is conducted under different settings of T_1 on the ICH dataset ($\rho = 0.4$, (η^l, η^u)=(0.3,0.5)) as shown in Fig. 4. It is observed that, after a certain number of warm-up training rounds, noisy client detection performance becomes stable in a certain range. In other words, the setting of T_1 would not significantly affect
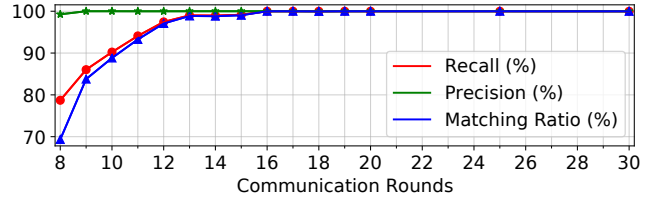

Figure 4: Ablation study of T_1 for warm-up training.

Noisy Clients	De-Noise Strategy		Type	BACC (%)	
	$\mathcal{L}_{KL}$	DaAgg		ICH	ISIC
✓	✗	✗	Best	66.60	61.20
			Last	63.90	56.11
✗	✗	✗	Best	69.67	64.94
			Last	68.07	59.17
✓	✗	✓	Best	69.32	65.72
			Last	68.18	58.60
✓	✓	✗	Best	64.53	60.79
			Last	64.01	43.57
✓	✓	✓	Best	**70.69**	**66.00**
			Last	70.14	60.65

Table 4: Ablation study of the second stage in FedNoRo. Settings: ρ=0.4 and (η^l, η^u)=(0.3,0.5) for the ICH dataset; ρ=0.6, (η^l, η^u)=(0.5,0.7) for the ISIC dataset.

FedNoRo's performance.

How $\mathcal{L}_{KL}$ and DaAgg Benefit Noise-Robust Training
Quantitative results of FedNoRo coupled with various de-noise strategies are summarized in Tab. 4. Without de-noise training, noisy clients with low-quality local models would surely degrade the classification performance. Though relying only on clean clients can get rid of label noise and somewhat improve classification performance, it may miss valuable information from other clients, making the performance sub-optimal. With noisy clients, replacing FedAvg with DaAgg effectively reduces the weights of noisy clients for global model updating, leading to performance improvement. With noisy clients, only adopting L_{KL} in noisy clients is less effective, as the global model may be severely affected by noisy clients containing inaccurate soft labels. Comparatively, jointly using $\mathcal{L}_{KL}$ and DaAgg can better explore valuable information from noisy clients while minimizing negative impact, outperforming the classification performance relying only on clean clients.

6 Conclusion

This paper presents a new federated label noise model where global data is class-imbalanced and label noise across clients is heterogeneous. Compared to existing label noise models assuming balanced global data, the proposed noise model is more realistic and flexible to describe real-world FL scenarios. To address this, we further propose a novel two-stage framework named FedNoRo for noise-robust federated learning. Experimental results on publicly-available datasets demonstrate FedNoRo's superior performance against the state-of-the-art FNLL approaches. We believe both the noise model and the solution would inspire future work on developing applicable FL frameworks for real-world applications.

Acknowledgments

References

[Agarwal *et al.*, 2018] Naman Agarwal, Ananda Theertha Suresh, Felix Xinnan X Yu, Sanjiv Kumar, and Brendan McMahan. cpSGD: Communication-efficient and differentially-private distributed sgd. In *NeurIPS*, 2018.

[Cao *et al.*, 2019] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, 2019.

[Chen *et al.*, 2020] Yiqiang Chen, Xiaodong Yang, Xin Qin, Han Yu, Biao Chen, and Zhiqi Shen. FOCUS: Dealing with label quality disparity in federated learning. *arXiv:2001.11359*, 2020.

[Chen *et al.*, 2021] Pengfei Chen, Junjie Ye, Guangyong Chen, Jingwei Zhao, and Pheng-Ann Heng. Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. In *AAAI*, pages 11442–11450, 2021.

[Codella and others, 2018] Noel CF Codella et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In *ISBI*, pages 168–172, 2018.

[Combalia *et al.*, 2019] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. BCN20000: Dermoscopic lesions in the wild. *arXiv:1908.02288*, 2019.

[Fang and Ye, 2022] Xiuwen Fang and Mang Ye. Robust federated learning with noisy and heterogeneous clients. In *CVPR*, pages 10072–10081, 2022.

[Flanders *et al.*, 2020] Adam E Flanders, Luciano M Prevedello, George Shih, Safwan S Halabi, Jayashree Kalpathy-Cramer, Robyn Ball, John T Mongan, Anouk Stein, Felipe C Kitamura, Matthew P Lungren, et al. Construction of a machine learning dataset through collaboration: The RSNA 2019 brain CT hemorrhage challenge. *Radiology: Artificial Intelligence*, 2(3), 2020.

[Han *et al.*, 2018] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015.

[Houle, 2013] Michael E Houle. Dimensionality, discriminability, density and distance distributions. In *ICDMW*, pages 468–473, 2013.

[Irvin and others, 2019] Jeremy Irvin et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, pages 590–597, 2019.

[Jiang *et al.*, 2022a] Meirui Jiang, Hongzheng Yang, Xiaoxiao Li, Quande Liu, Pheng-Ann Heng, and Qi Dou. Dynamic bank learning for semi-supervised federated image diagnosis with class imbalance. In *MICCAI*, pages 196–206, 2022.

[Jiang *et al.*, 2022b] Xuefeng Jiang, Sheng Sun, Yuwei Wang, and Min Liu. Towards federated learning against noisy labels via local self-regularization. In *CIKM*, pages 862–873, 2022.

[Kaissis *et al.*, 2020] Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311, 2020.

[Kang *et al.*, 2020] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *ICLR*, 2020.

[Karimireddy *et al.*, 2020] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *ICML*, pages 5132–5143, 2020.

[Karthik *et al.*, 2021] Shyamgopal Karthik, Jérome Revaud, and Boris Chidlovskii. Learning from long-tailed data with noisy labels. *arXiv:2108.11096*, 2021.

[Konečnỳ *et al.*, 2016] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv:1610.05492*, 2016.

[Laine and Aila, 2017] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017.

[Li and Zhan, 2021] Xin-Chun Li and De-Chuan Zhan. FedRS: Federated learning with restricted softmax for label distribution non-iid data. In *KDD*, pages 995–1005, 2021.

[Li *et al.*, 2020a] Junnan Li, Richard Socher, and Steven CH Hoi. DivideMix: Learning with noisy labels as semi-supervised learning. In *ICLR*, 2020.

[Li *et al.*, 2020b] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *MLSys*, pages 429–450, 2020.

[Li *et al.*, 2021a] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *CVPR*, pages 10713–10722, 2021.

[Li *et al.*, 2021b] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated learning on non-iid features via local batch normalization. In *ICLR*, 2021.

[Liang *et al.*, 2022] Xiaoxiao Liang, Yiqun Lin, Huazhu Fu, Lei Zhu, and Xiaomeng Li. RSCFed: Random sampling consensus federated semi-supervised learning. In *CVPR*, pages 10154–10163, 2022.

[Liu *et al.*, 2021] Quande Liu, Hongzheng Yang, Qi Dou, and Pheng-Ann Heng. Federated semi-supervised medical image classification via inter-client relation matching. In *MICCAI*, pages 325–335, 2021.

[Lukasik *et al.*, 2020] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *ICML*, pages 6448–6458, 2020.

[Lukov *et al.*, 2022] Tohar Lukov, Na Zhao, Gim Hee Lee, and Ser-Nam Lim. Teaching with soft label smoothing for mitigating noisy labels in facial expressions. In *ECCV*, pages 648–665, 2022.

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, pages 1273–1282, 2017.

[Menon *et al.*, 2021] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2021.

[Sattler *et al.*, 2019] Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Robust and communication-efficient federated learning from non-iid data. *IEEE Transactions on Neural Networks and Learning Systems*, 31(9):3400–3413, 2019.

[Shang *et al.*, 2022] Xinyi Shang, Yang Lu, Gang Huang, and Hanzi Wang. Federated learning on heterogeneous and long-tailed data via classifier re-training with federated features. In *IJCAI*, 2022.

[Shen *et al.*, 2021] Zebang Shen, Juan Cervino, Hamed Hassani, and Alejandro Ribeiro. An agnostic approach to federated learning with class imbalance. In *ICLR*, 2021.

[Tschandl *et al.*, 2018] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):1–9, 2018.

[Wang *et al.*, 2019] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, pages 322–330, 2019.

[Wang *et al.*, 2021] Lixu Wang, Shichao Xu, Xiao Wang, and Qi Zhu. Addressing class imbalance in federated learning. In *AAAI*, pages 10165–10173, 2021.

[Wang *et al.*, 2022] Zhuowei Wang, Tianyi Zhou, Guodong Long, Bo Han, and Jing Jiang. FedNoiL: A simple two-level sampling method for federated learning with noisy labels. *arXiv:2205.10110*, 2022.

[Wei *et al.*, 2022a] Jiaheng Wei, Hangyu Liu, Tongliang Liu, Gang Niu, Masashi Sugiyama, and Yang Liu. To smooth or not? when label smoothing meets noisy labels. In *ICML*, 2022.

[Wei *et al.*, 2022b] Qi Wei, Haoliang Sun, Xiankai Lu, and Yilong Yin. Self-Filtering: A noise-aware sample selection for label noise with confidence penalization. In *ECCV*, pages 516–532, 2022.

[Xu *et al.*, 2022] Jingyi Xu, Zihan Chen, Tony QS Quek, and Kai Fong Ernest Chong. FedCorr: Multi-stage federated learning for label noise correction. In *CVPR*, pages 10184–10193, 2022.

[Yan *et al.*, 2020] Zengqiang Yan, Jeffry Wicaksana, Zhiwei Wang, Xin Yang, and Kwang-Ting Cheng. Variation-aware federated learning with multi-source decentralized medical image data. *IEEE Journal of Biomedical and Health Informatics*, 25(7):2615–2628, 2020.

[Yang *et al.*, 2021] Miao Yang, Hua Qian, Ximin Wang, Yong Zhou, and Hongbin Zhu. Client selection for federated learning with label noise. *IEEE Transactions on Vehicular Technology*, 71(2):2193–2197, 2021.

[Yang *et al.*, 2022] Seunghan Yang, Hyoungseob Park, Junyoung Byun, and Changick Kim. Robust federated learning with noisy labels. *IEEE Intelligent Systems*, 37(2):35–43, 2022.

[Yi *et al.*, 2022] Xuanyu Yi, Kaihua Tang, Xian-Sheng Hua, Joo-Hwee Lim, and Hanwang Zhang. Identifying hard noise in long-tailed sample distribution. In *ECCV*, pages 739–756, 2022.

[Yu *et al.*, 2019] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *ICML*, pages 7164–7173, 2019.

[Zhang *et al.*, 2017] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. *arXiv:1710.09412*, 2017.

[Zhao *et al.*, 2022] Ganlong Zhao, Guanbin Li, Yipeng Qin, Feng Liu, and Yizhou Yu. Centrality and consistency: Two-stage clean samples identification for learning with instance-dependent noisy labels. In *ECCV*, pages 21–37, 2022.

[Zheng *et al.*, 2021] Wenbo Zheng, Lan Yan, Chao Gou, and Fei-Yue Wang. Federated meta-learning for fraudulent credit card detection. In *IJCAI*, pages 4654–4660, 2021.

[Zhou *et al.*, 2020] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*, pages 9719–9728, 2020.

[Zhu *et al.*, 2021] Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *ICML*, pages 12878–12889, 2021.