FedES: Federated Early-Stopping for Hindering Memorizing Heterogeneous Label Noise

Bixiao Zeng[1,2], **Xiaodong Yang**[1], **Yiqiang Chen**[*1,2,3], **Zhiqi Shen**[4], **Hanchao Yu**[5] and **Yingwei Zhang**[1]

[1]Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Peng Cheng Laboratory
[4]Nanyang Technological University
[5]Bureau of Frontier Sciences and Education, Chinese Academy of Sciences
{zengbixiao19b, yangxiaodong, yqchen}@ict.ac.cn, zqshen@ntu.edu.sg, {yuhanchao, zhangyingwei}@ict.ac.cn

Abstract

Federated learning (FL) facilitates collaborative model training across distributed clients while maintaining privacy. Federated noisy label learning (FNLL) is more of a challenge for data inaccessibility and noise heterogeneity. Existing works primarily assume clients are either noisy or clean, which may lack the flexibility to adapt to diverse label noise across different clients, especially when entirely clean or noisy clients are not the majority. To address this, we propose a general noise-robust federated learning framework called Federated Early-Stopping (FedES), which adaptively updates critical parameters of each local model based on their noise rates, thereby avoiding overfitting to noisy labels. FedES is composed of two stages: federated noise estimation and parameter-adaptive local updating & global aggregation. We introduce a signed distance based on local and global gradients during a federated round to estimate clients' noise rates without requiring additional information. Based on this measure, we employ various degrees of early-stopping during local updating on the clients, and further, a noise-aware global aggregation is employed to achieve noise-robust learning. Extensive experiments conducted on varying synthetic and real-world label noise demonstrate the superior performance of FedES over the state-of-the-art methods.

1 Introduction

Federated learning (FL) empowers collaborative deep learning model training without the need to share sensitive data, making it particularly valuable in privacy-centric domains such as healthcare and finance [McMahan *et al.*, 2017; Dayan *et al.*, 2021; Rieke *et al.*, 2020]. Many FL systems utilize datasets with associated labels, but labeling data is an expensive and resource-intensive endeavor [Song *et al.*, 2022]. Differences in multi-source labeling skills and class distribution result in varying noise rates, known as noise heterogeneity [Fang and Ye, 2022]. For example, doctors in different hospitals may face varying cases and have differing diagnostic skills, leading to varying misdiagnosis rates [Ju *et al.*, 2022; Bernhardt *et al.*, 2022; Karimi *et al.*, 2020].

Existing federated noisy label learning (FNLL) addresses noise heterogeneity by distinguishing noisy clients from clean ones. Some methods discard clients based on specific criteria, assuming that the number of clients needed for training is smaller than the total participating clients ($S < N$) [Nagalapatti and Narayanam, 2021; Deng *et al.*, 2021]. However, discarding clients, even if noisy, may result in the loss of valuable information since they could contain clean data. Furthermore, the clients who are retained may not be entirely clean. Alternatively, other methods detect noisy clients and employ de-noise strategies like pseudo-labeling or knowledge distillation to them [Xu *et al.*, 2022; Wu *et al.*, 2023]. Nevertheless, these approaches still treat clients as either noisy or clean, potentially leading to suboptimal performance. One straightforward way to address noise heterogeneity adaptively is to tailor de-noise strategies for each client, such as data selection. However, removing the noisy samples in each client may still lose valuable information and leave residual noise [Tuor *et al.*, 2021; Zeng *et al.*, 2022; Zeng *et al.*, 2023]. So far, there has been limited exploration of adaptive approaches to handling noise heterogeneity in the context of FNLL.

In the landscape of federated learning systems, the question arises: *How can a de-noise strategy be adapted to clients with varying noise rates without causing information loss or noise residue?* For example, early-stopping explores the dynamic optimization policies during the training of deep neural networks (DNNs) [Rolnick *et al.*, 2017; Li *et al.*, 2020; Nguyen *et al.*, 2019; Tanaka *et al.*, 2018], focusing on the

*Corresponding author

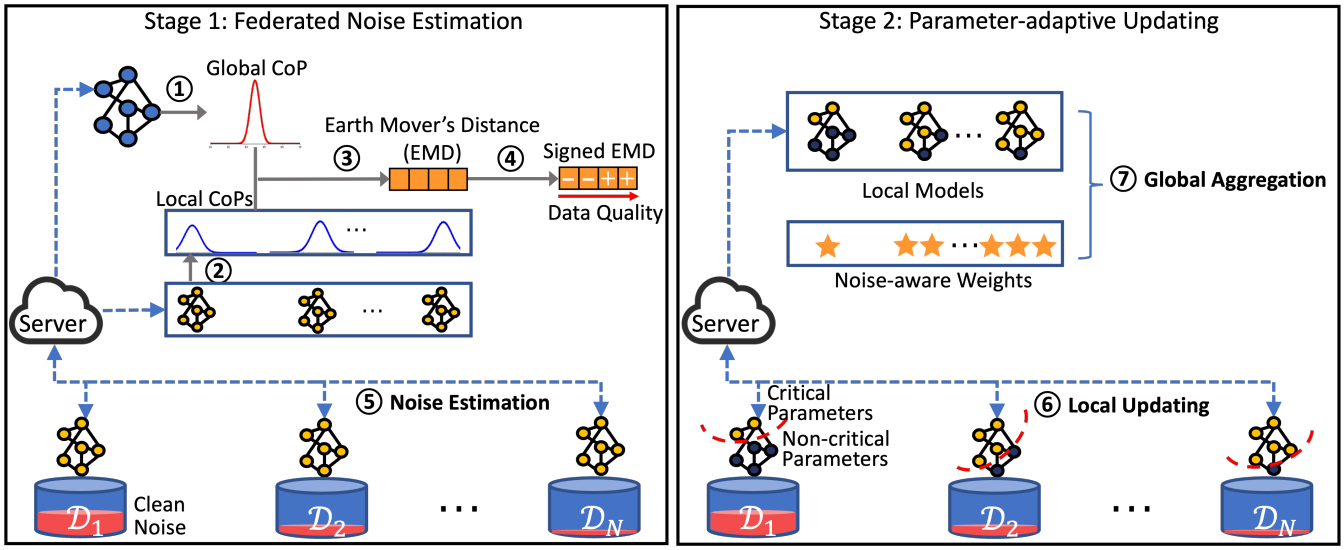


Figure 1: Overview of the proposed two-stage framework FedES.

memorization effect that DNNs tend to first memorize clean labels and then memorize noisy ones [Arpit *et al.*, 2017]. Although stopping training at a certain time point has shown promising results, it remains difficult to avoid memorizing noisy labels from the beginning [Han *et al.*, 2020].

Early-stopping on a non-critical segment of DNNs has arisen to prevent DNNs from memorizing noisy labels [Bai *et al.*, 2021; Xia *et al.*, 2020]. Some methods involve stopping the training of noise-sensitive layers, while other methods involve stopping the training of non-critical parameters. However, these methods all require some prior knowledge, such as the noise rate of training data. Unfortunately, in federated learning, noise rates remain unknown and exhibit variations among heterogeneous clients.

In this paper, we introduce a general noise-robust FL framework to handle clients with varying noise rates, as depicted in Figure 1. To ensure privacy, we introduce a signed distance to estimate clients' noise rates by measuring the distribution of criticality of parameters (CoP) of local models. Extensive experiments have shown a positive correlation between the amount of clean data and critical parameters [Frankle and Carbin, 2018; Xia *et al.*, 2021], suggesting more clean data need more critical model parameters to memorize them. Hence, clients with more clean data will exhibit more large CoP values. The computation of CoP is related to the parameter and its gradient, thus no additional information needs to be requested from clients. Based on this measure, we then employ various degrees of early-stopping during local updating on clients. To enhance training stability and mitigate negative noise impact, we introduce a noise-aware global aggregation using weights based on estimated noise rates.

The primary contributions of our work can be summarized as follows:

1. We present a general noise-robust framework, FedES, to handle noise heterogeneity where clients have varying noise rates instead of a binary noisy-vs-clean problem.
2. We present a general noise-generation approach for modeling federated label noise, incorporating varying noise rates for clients with a continuous spectrum.
3. We estimate each client's noise rate via a signed EMD based on the local and global gradient, without requiring additional information from clients.
4. We demonstrate that FedES outperforms state-of-the-art FL methods on both varying synthetic federated label noise and real-world label noise.

2 Related Work

2.1 Federated Methods

Detecting noisy clients and then handling them separately from clean ones is a common practice for addressing the FNLL problem. For example, S-FedAvg [Nagalapatti and Narayanam, 2021] discards clients with lower Shapley-based marginal contribution. Similarly, Fair [Deng *et al.*, 2021] discards clients with lower loss differences. However, these methods assume clients are either highly clean or noisy, leading to information loss in the 'noisy' clients and noise residue in the 'clean' clients. Without discarding clients, [Xu *et al.*, 2022] conducts pseudo-labeling for detected noisy clients. FedNoRo [Wu *et al.*, 2023] employs knowledge distillation for better noisy label learning on detected noisy clients. They somewhat alleviate the information loss but still, handle clients with varying noise rates in a binary way. A more general way is to conduct data selection for each client. [Tuor *et al.*, 2021] filters out noisy training samples early before training, while [Zeng *et al.*, 2022] iteratively removes them. However, both methods may cause information loss and leave some noise residues. Recently, an FNLL library named FedNoisy [Liang *et al.*, 2023] has orga-

nized multiple centralized de-noise strategies that can be employed by clients locally [Zhang *et al.*, 2017; Han *et al.*, 2018; Wang *et al.*, 2019]. However, such centralized strategies face limitations in estimating label noise in small client data and do not produce a noise-robust weighting scheme for global aggregation.

2.2 Criticality of Parameters (CoP)

The memorization effect reveals that DNNs tend to first memorize clean labels and then memorize noisy ones [Arpit *et al.*, 2017]. Based on this, the parameters that contribute to optimality at an early stage are important for clean labels. Informally, CoP [Xia *et al.*, 2021] is a metric for quantifying the criticality of model parameters, which can be computed as follows:

$$g_i = \left|\nabla l\left(\mathrm{w}_i\right) \times \mathrm{w}_i\right|, i \in [m], \tag{1}$$

where w_i denotes the parameter, $l(\mathrm{w}_i)$ denotes the loss function such as cross entropy, $\nabla l(\mathrm{w}_i)$ denotes the gradient of $l(\mathrm{w}_i)$ and m is the number of all parameters. If g_i is large, w_i is viewed as a critical parameter. $\nabla l(\mathrm{w}_i)$ indicates how much the output of the network would change in the parameter. However, a large gradient alone doesn't necessarily imply that the parameter is critical if the parameter itself is small. In other words, even though the gradient is large, the small value of the parameter means that it doesn't have a substantial impact on the output of a DNN. Generally, more clean labels require more critical model parameters in the early stage.

3 Problem Definition

3.1 Preliminaries

Let's consider a federated learning system that has N clients and a distributed dataset $\mathbb{D} = \{\mathbb{D}_n\}_{n=1}^{N}$. Here, $\mathbb{D}_n = \left\{\left(x_n^i, y_n^i\right)\right\}_{i=1}^{|\mathbb{D}_n|}$ represents the local dataset for client n, and $\mathcal{W}_n$ denotes the model parameters of client n. Each client has a noise rate of τ_n, which is defined as:

$$\tau_n = \frac{|y_n^i \neq y_n^{*i}|}{|\mathbb{D}_n|}, i \in [1, |\mathbb{D}_n|] \tag{2}$$

where y_n^{*i} represents the correct label for instance x_n^i. In this way, data quality is defined as:

$$q_n = 1 - \tau_n, \tag{3}$$

which is the proportion of clean labels to total data. To estimate the noise rates and support the parameter-adaptive updating, two major foundations are used: the federated noise model and FedAvg updating process. In the following subsections, we will provide more details on these two foundations.

3.2 Federated Noise Model

We propose a federated noise model to simulate label noise in FL systems. Our model captures varying noise rates across clients, avoiding binary classification of clients as clean or noisy [Xu *et al.*, 2022]. With a bell-like shape $\mathcal{N}(\mu, \sigma)$, our model characterizes the noise rate distribution among clients. Using $\tau \sim \mathcal{N}(\mu, \sigma)$ offers insights into the central tendencies and variabilities in noise rates across clients. Additionally, the independence of clients allows us to apply the central limit theorem (CLT) [Fischer, 2011], suggesting that the distribution of noise rates across independent clients is likely to approximate a normal distribution.

The generation of noise rates for N clients can be expressed as:

$$\tau_n = \min(\max(\tau, 0), 1), \tau \sim \mathcal{N}(\mu, \sigma), \tag{4}$$

where τ_n represents the n-th client and is clipped by the range $[0, 1]$. With a specified noise rate, one can easily create both symmetric and asymmetric label noise [Wang *et al.*, 2019]. In the case of symmetric label noise, instances undergo random label flipping, with an equal probability determined by the noise rate, ensuring an even likelihood for each class. Conversely, for asymmetric label noise, instances are more inclined to receive class-conditioned labels, introducing a biased likelihood for incorrect labels, again based on the specified noise rate.

3.3 FedAvg Updating Process

FedAvg proposes a standard round of federated learning that involves local updating and global aggregation [McMahan *et al.*, 2017]. Let $\mathcal{W}(t)$ be the global model (i.e., DNN) at round t and distribute it to each client. The local update process for client n is expressed as:

$$\mathcal{W}_n(t+1) = \mathcal{W}(t) - \eta \nabla L_n(\mathcal{W}(t)), \tag{5}$$

where η is the learning rate and $\nabla L_n(\mathcal{W}(t))$ represents the local gradient on $\mathcal{W}(t)$. The global update process can then be expressed as:

$$\begin{aligned} \mathcal{W}(t+1) &= \sum_{n}^{N} \frac{|\mathbb{D}_n|}{|\mathbb{D}|} \mathcal{W}_n(t+1) \\ &= \mathcal{W}(t) - \eta \sum^{N} \frac{|\mathbb{D}_n|}{|\mathbb{D}|} \nabla L_n(\mathcal{W}(t)), \end{aligned} \tag{6}$$

where the weight of each local model is determined by the data size $|\mathbb{D}_n|$. It can be stated that the aggregation of the model parameters is equal to the weighted average of local gradients.

4 Methodology

To address clients with varying noise rates, we propose a general noise-robust framework named FedES (see Algorithm 1). Our method comprises two stages: federated noise estimation and parameter-adaptive updating. In the first stage, a global model is pre-trained by FedAvg, which is used to calculate the global and local parameter-wise CoPs. Then, the distance between global and local parameter-wise CoPs is used to estimate clients' noise rates. In the second stage, according to the noise estimation results, different degrees of early-stopping are employed during local updating. Moreover, a noise-aware global aggregation is deployed to maintain the updating stability and reduce the negative impact of noise.

4.1 Federated Noise Estimation

In Section 2.2, we discussed the positive correlation between the amount of clean data and critical parameters. To estimate

Algorithm 1: FedES: Federated Early-Stopping

Input: N (number of clients), T (number of training rounds), $\mathbb{D} = \{\mathbb{D}_n\}_{n=1}^{N}$ (dataset), $\mathcal{W}^{\text{pre}}(s_1)$ (global model before the last pre-training round), $\mathcal{W}_n^{\text{pre}}(s_1+1)$ (local model after the last pre-training round), $\mathcal{W}(1)$ (initialized global model in the second stage)
Output: Global $\mathcal{W}(T)$
$\mathcal{W}^{\text{pre}}(s_1+1) = \sum_{n=1}^{N} \frac{|\mathbb{D}_n|}{\sum_k |\mathbb{D}_k|} \mathcal{W}_n^{\text{pre}}(s_1+1)$
// Federated noise estimation
$\mathbf{g}_s \leftarrow |(\mathcal{W}^{\text{pre}}(s_1+1) - \mathcal{W}^{\text{pre}}(s_1)) * \mathcal{W}^{\text{pre}}(s_1)|$ // global CoP
for *each client* $n = 1$ *to* N ***in parallel*** **do**
  $\mathbf{g}_n \leftarrow |(\mathcal{W}_n^{\text{pre}}(s_1+1) - \mathcal{W}^{\text{pre}}(s_1)) * \mathcal{W}^{\text{pre}}(s_1)|$ // local CoP
  $d_n \leftarrow \text{EMD}_{\text{signed}}(\mathbf{g}_n, \mathbf{g}_s)$ // signed EMD
  $\rho_n \leftarrow \frac{d_n - \min(\boldsymbol{d})}{\max(\boldsymbol{d}) - \min(\boldsymbol{d})}$ // estimated data quality
end
// Parameter-adaptive local updating
for *each round* $t = 1, ..., T$ **do**
  $\mathcal{W}_n \leftarrow \mathcal{W}(t)$
  for *each client* $n = 1, ..., N$ ***in parallel*** **do**
    for *each local iteration* $t' = 1, 2, ...E$ **do**
      $D_n \leftarrow$ select a minibatch of size $B \subseteq \mathbb{D}_n$
      $\mathcal{W}_n \leftarrow \mathcal{W}_n - \eta \rho_n \mathcal{M}_n \nabla L(D_n, \mathcal{W}_n)$
      //$\mathcal{M}_n$ obtained in Eq. 15
    end
  end
  // Noise-aware global aggregation
  $\mathcal{W}(t+1) = \sum_{n=1}^{N} \frac{|\mathbb{D}_n|\rho_n}{\sum_k |\mathbb{D}_k|\rho_k} \mathcal{W}_n$
end
return $\mathcal{W}_n(T)$

noisy clients, we examine the distribution of CoP for each client. For a more accurate computation of CoP, we first pre-train a global model $\mathcal{W}^{\text{pre}}(s_1+1)$ for s_1 rounds by FedAvg (s_1 will be discussed in Section 5.4).

Based on the global update process described in Eq. 6, the parameter-wise global CoP expressed as the product of the parameter and its gradient is given by:

$$\mathbf{g}_s \leftarrow |(\mathcal{W}^{\text{pre}}(s_1+1) - \mathcal{W}^{\text{pre}}(s_1)) * \mathcal{W}^{\text{pre}}(s_1)|, \quad (7)$$

where the global gradient is essentially the average of all local gradients. Similarly, based on the local update process described in Eq. 5, the local CoP is given by:

$$\mathbf{g}_n \leftarrow |(\mathcal{W}_n^{\text{pre}}(s_1+1) - \mathcal{W}^{\text{pre}}(s_1)) * \mathcal{W}^{\text{pre}}(s_1)|, \quad (8)$$

where the parameter-wise local gradient denoted by $\mathcal{W}_n^{\text{pre}}(s_1+1) - \mathcal{W}^{\text{pre}}(s_1)$ can be easily obtained in a federated round. The learning rate η can be omitted for clarity as it only serves as a scaling factor. This method allows us to obtain both local and global CoPs by computing parameter-wise gradients during a federated round. It eliminates the need for additional information.

Considering that higher data quality is associated with a CoP distribution having many large values, and the shape of distributions with varying noise rates differs from the global distribution, Earth Mover's Distance (EMD) is suitable for measuring the shape difference between local and global CoP [Rubner *et al.*, 2000]. As EMD quantifies the mass required to transition between distributions, the distance concerning the CoP of a low-data-quality client may be the same as that of a high-data-quality client (a horizontally flipped version of a low-data-quality client).

To resolve this conflict, a signed EMD is required. We first gather $N+1$ parameter-wise local CoPs and global CoP as $G = [\mathbf{g}_1, ..., \mathbf{g}_N, \mathbf{g}_s]$. Then, we use an $N+1$-component Gaussian Mixture Model (GMM) to fit and predict G. We exploit the vector of means of predicted results of G, which is denoted as $[\boldsymbol{\mu}_1, ...\boldsymbol{\mu}_N, \boldsymbol{\mu}_{N+1}]$. Thus, the signed EMD between local and global CoP can be calculated by:

$$\begin{aligned} d_n &= \text{sgn}\left(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{N+1}\right) \cdot \text{EMD}\left(\mathbf{g}_n, \mathbf{g}_s\right) \\ &= \text{sgn}\left(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{N+1}\right) \cdot \inf_{\pi \in \Pi(\mathbf{g_n}, \mathbf{g_s})} \mathbb{E}_{(x,y)\sim\pi}[d(x,y)] \end{aligned} \quad (9)$$

where $\text{EMD}(\cdot,\cdot)$ denotes the minimum-cost transportation plan π^* of moving the distribution mass from one to another, and the cost is a ground distance metric $d(x,y)$. $\text{sgn}(\cdot)$ is the sign of the difference between $\boldsymbol{\mu}_n$ and $\boldsymbol{\mu}_{N+1}$, which is expressed as

$$\text{sgn}(x) = -[x < 0] + [x > 0], \quad (10)$$

where $[\cdot]$ is the Iverson bracket.

In this way, clients with high data quality have a large proportion of critical model parameters that can be reflected in a large d_n (positive). Conversely, clients with low data quality have a large proportion of non-critical model parameters that can be reflected in a small d_n (negative). Based on this, the data quality can be generated by:

$$\rho_n = \frac{d_n - \min(\boldsymbol{d})}{\max(\boldsymbol{d}) - \min(\boldsymbol{d})}, \quad (11)$$

where the d_n is scaled to the range of $[0, 1]$ using min-max scaling. The experimental results on data quality estimation will be presented in Section 5.4.

4.2 Parameter-adaptive Updating

With the data quality ρ_n, different degrees of early-stopping are employed for local updating. The CoP ranking for each client is computed as follows:

$$\mathbf{g}_n^{\downarrow} = \left[g_n^{\downarrow}[1], \ldots, g_n^{\downarrow}[m_n^c], \ldots, g_n^{\downarrow}[m]\right], \quad (12)$$

$$g_n^{\downarrow}[1] \geq \ldots \geq g_n^{\downarrow}[m_n^c] \geq \ldots \geq g_n^{\downarrow}[m] \quad (13)$$

where $g_n^{\downarrow}[i]$ is the sorted CoP for each parameter, m_n^c is the number of critical model parameters which is given by:

$$m_n^c = \rho_n * m, \quad (14)$$

where m is the number of model parameters.

The critical and non-critical parameters are then determined as follows:

$$\mathcal{M}_n[i] = \begin{cases} 1, & \text{if } g_n^{\downarrow}[1] \geq g[i] \geq g_n^{\downarrow}[m_n^c] \\ 0, & \text{otherwise} \end{cases}, \quad (15)$$

which is a parameter-wise indicator. For each client, only critical model parameters undergo gradient decay. The process of selective gradient decay (SeGD) for each client is carried out in the following manner:

$$\mathcal{W}_n\left(t'+1\right) \leftarrow \mathcal{W}_n\left(t'\right) - \eta\rho_n \mathcal{M}_n \odot \nabla L\left(\mathcal{W}_n\left(t'\right)\right), \quad (16)$$

where t' represents the current iteration, the gradient decay coefficient ρ_n is set to prevent overconfident descent steps. It can be seen that clients with low data quality have fewer critical model parameters and should be stopped early. Conversely, clients with high data quality have more critical model parameters and should be updated.

In the following analysis, we show how parameter-adaptive local updating based on ρ_n is equivalent to adapting the underlying loss function across clients in FL.

Theorem 1. *Let $\mathcal{W}_n$ and $\widehat{\mathcal{W}}_n$ refer to two sets of model parameters that use selective gradient decay and standard one, respectively. When an update is performed at iteration t', two different model parameters are obtained: $\mathcal{W}_n(t'+1)$ and $\widehat{\mathcal{W}}_n(t'+1)$. The extent of regularization, i.e., the loss difference at the next iteration, is controlled by ρ_n as follows:*

$$\begin{aligned} &L(\mathcal{W}_n(t'+1)) - L(\widehat{\mathcal{W}}_n(t'+1)) \\ &= \eta\Big(\sum_{\mathcal{M}_n[i]=1} (1-\rho_n)\nabla l_i^2 + \sum_{\mathcal{M}_n[i]=0} 1 * \nabla l_i^2\Big), \end{aligned} \quad (17)$$

where ∇l_i represents a specific entry in $\nabla L(\mathcal{W}_n(t'))$. For the sake of clarity, the proof can be found in Appendix A.

The following corollaries illustrate the extent of regularization on varying ρ_n:

Corollary 1. *For $\rho_n = 1,\ \forall i, \mathcal{M}_n[i] = 1$: all model parameters are critical and updated in a standard way, therefore $L(\mathcal{W}_n(t'+1)) - L(\widehat{\mathcal{W}}_n(t'+1)) = 0$.*

Corollary 2. *For $\rho_n = 0,\ \forall i, \mathcal{M}_n[i] = 0$: all model parameters are non-critical and early stopped, therefore the $L(\mathcal{W}_n(t'+1)) - L(\widehat{\mathcal{W}}_n(t'+1)) = \eta\|\nabla L(\mathcal{W}_n(t'))\|^2$, which is the maximum of $L(\mathcal{W}_n(t'+1)) - L(\widehat{\mathcal{W}}_n(t'+1))$.*

Corollary 3. *For $\rho_n \in (0,1)$: apparently, $L(\mathcal{W}_n(t'+1)) - L(\widehat{\mathcal{W}}_n(t'+1)) \in (0, \eta\|\nabla L(\mathcal{W}_n(t'))\|^2)$. Decreasing ρ_n increases the loss difference from the traditional update process, as it multiplies a larger proportion of gradients by 1 and the remaining small proportion by the increased (1-ρ_n). ρ_n controls the regularization effect that prevents the model from memorizing noise, so a small ρ_n exactly suggests stronger regularization.*

In global aggregation, we assign different weights to models based on their noise rates. High-quality data from clients with large ρ_n leads to significant updates. To maintain the influence of substantial updates in the right direction while reducing the negative impact of noise, we propose noise-aware aggregation (NaAgg):

$$\mathcal{W}(t+1) = \sum_{n=1}^{N} \frac{|\mathbb{D}_n|\,\rho_n}{\sum_k |\mathbb{D}_k|\,\rho_k} \mathcal{W}_n(t+1). \quad (18)$$

This ensures that clients contributing significantly to the update maintain their influence after global aggregation, emphasizing the importance of cleaner data in guiding the optimization direction.

5 Experimental Results

5.1 Dataset

Two groups of datasets are used for evaluation: CIFAR-10/100 with synthetic label noise and Clothing1M with real-world label noise:

1. CIFAR-10/100 datasets [Krizhevsky *et al.*, 2009]: CIFAR-10 has 60,000 32x32 color images in 10 categories, with each category having 6,000 images. CIFAR-100 has 60,000 32x32 color images in 100 categories, grouped into 20 supercategories with 5 specific categories each. Both datasets are divided into train/test sets with a ratio of 5:1.
2. Clothing1M dataset [Xiao *et al.*, 2015]: Clothing1M has 1M clothing images in 14 classes. The dataset contains real-world label noise due to it being collected from multiple online shopping websites. As commonly practiced, Clothing1M is divided into train/test sets with a ratio of 4:1.

Note that for both datasets, 20% of the test set is reserved as a benchmark dataset so that some comparison methods can use it as a reference to evaluate noisy clients.

5.2 Experimental Setup

Data Partition and Noise Generation

We consider the Dirichlet distribution [Liang *et al.*, 2023; Li *et al.*, 2021] and the Federated noise model described in Section 3.1 to generate the $N = 20$ heterogeneous-noise clients. We consider both IID and non-IID data partitions in our work. In the case of IID partitions, we randomly distribute the entire dataset $\mathbb{D}$ among N clients [Kairouz *et al.*, 2021]. For non-IID partitions, we sample $p_k \sim \mathrm{Dir}_N(\beta)$ and assign a proportion $p_{k,j}$ of instances of class k to client j, where $\mathrm{Dir}(\beta)$ is the Dirichlet distribution with a concentration parameter β (default = 0.5). For evaluation under different settings of label noise, we use $\mathcal{N} \sim (0.3, 0.2)$ and $\mathcal{N} \sim (0.5, 0.2)$ to generate the varying noise rates of clients. As detailed in Section 3.1, both symmetric and asymmetric label noise are considered. Note that for the Clothing1M dataset, the noise is in its original form and not artificially generated.

Implementation Details

The ResNet-18 [He *et al.*, 2016] model serves as the backbone for both the CIFAR-10/100 and Clothing1M datasets. For the CIFAR-10/100 dataset, the local epoch is 10. For the Clothing1M dataset, the local epoch is set to 5. The learning rate is 0.01 for the CIFAR-10/100 dataset and 0.001 for the Clothing1M dataset. The batch size is 128 for CIFAR-10/100 and 32 for Clothing1M. Other settings remain the same for both datasets, including 10 rounds for pre-training, a total of 200 communication rounds, and an optimizer of SGD with a weight decay of 5e-4 and momentum of 0.9.

5.3 Comparison with SOTA Methods

We have categorized state-of-the-art methods into two groups based on whether they treat clients as either noisy or clean (i.e., in a binary manner). The first group is the binary denoise methods, which includes S-FedAvg [Nagalapatti and

Category	Method	IID				Non-IID			
		Symmetric		Asymmetric		Symmetric		Asymmetric	
		$\mu = 0.3$	$\mu = 0.5$	$\mu = 0.3$	$\mu = 0.5$	$\mu = 0.3$	$\mu = 0.5$	$\mu = 0.3$	$\mu = 0.5$
Baseline	FedAvg	78.32±0.36	55.59±0.72	81.62±0.32	50.28±0.03	58.75±0.06	32.56±0.82	63.06±0.66	32.52±0.82
Binary De-noise	S-FedAvg	85.42±0.28	63.72±0.95	88.94±0.62	58.82±0.04	66.55±0.17	41.27±0.53	70.62±0.52	40.72±0.98
	Fair	83.56±0.08	64.35±0.83	87.60±0.22	58.35±0.76	64.04±0.58	41.27±0.18	68.32±0.97	40.64±0.97
	FedNoRo	87.55±0.29	71.00±0.11	83.79±0.14	48.16±0.38	59.36±0.61	35.36±0.27	53.97±0.76	47.62±0.68
General De-noise	Fed-SCE	90.19±0.21	83.00±0.34	84.77±0.10	52.50±0.67	83.66±0.38	65.33±0.56	70.92±0.04	23.63±0.30
	Fed-Mixup	88.72±0.15	74.19±0.69	87.77±0.20	54.61±0.52	70.72±0.48	40.07±0.15	66.71±0.17	31.56±0.83
	Fed-Coteaching	85.38±0.17	73.67±0.20	87.15±0.09	58.20±0.59	76.64±0.73	54.77±0.12	72.25±0.78	22.26±0.71
Ours	**FedES**	**93.09±0.93**	**85.40±0.34**	**90.79±0.91**	**60.34±0.36**	**85.74±0.99**	**68.11±0.48**	**74.54±0.65**	**50.59±0.41**

Table 1: Test Accuracy (%) comparison results on CIFAR-10 datasets under varying synthetic federated label noise

Category	Method	IID				Non-IID			
		Symmetric		Asymmetric		Symmetric		Asymmetric	
		$\mu = 0.3$	$\mu = 0.5$	$\mu = 0.3$	$\mu = 0.5$	$\mu = 0.3$	$\mu = 0.5$	$\mu = 0.3$	$\mu = 0.5$
Baseline	FedAvg	46.22±0.60	30.94±0.87	53.01±0.80	31.66±0.63	42.11±0.36	25.84±0.12	51.72±0.83	33.04±0.01
Binary De-noise	S-FedAvg	53.29±0.00	39.78±0.44	60.33±0.88	40.56±0.53	49.60±0.45	34.22±0.24	58.74±0.06	41.07±0.36
	Fair	51.71±0.19	39.92±0.64	58.54±0.43	39.83±0.05	47.11±0.72	34.44±0.04	56.99±0.79	41.45±0.91
	FedNoRo	59.76±0.38	47.14±0.40	61.13±0.13	33.22±0.75	42.73±0.64	30.40±0.15	50.43±0.06	44.97±0.29
General De-noise	Fed-SCE	57.83±0.51	48.01±0.74	58.05±0.36	33.01±0.43	63.17±0.27	50.20±0.45	57.36±0.11	34.63±0.23
	Fed-Mixup	60.14±0.73	47.05±0.56	62.16±0.59	37.08±0.24	55.86±0.12	40.86±0.18	58.27±0.42	37.57±0.29
	Fed-Coteaching	59.22±0.45	44.27±0.33	58.98±0.50	34.64±0.98	58.45±0.02	42.72±0.43	60.59±0.35	39.03±0.16
Ours	**FedES**	**63.13±0.32**	**50.59±0.68**	**65.11±0.09**	**39.58±0.37**	**65.51±0.75**	**52.96±0.76**	**62.72±0.89**	**47.05±0.11**

Table 2: Test Accuracy (%) comparison results on CIFAR-100 datasets under varying synthetic federated label noise

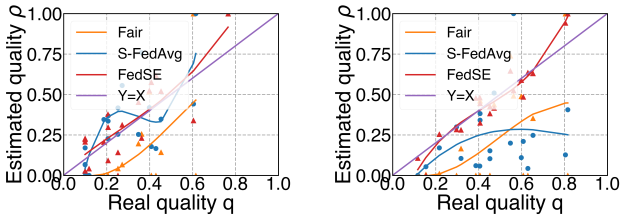
Narayanam, 2021], Fair [Deng *et al.*, 2021], and FedNoRo [Wu *et al.*, 2023]. Particularly, S-FedAvg and Fair exploit a benchmark dataset to identify noisy clients, while FedNoRo exploits client-wise features. The second group is the general de-noise methods, which consist of some well-known centralized methods that can be applied to the client's local environment. These methods are organized by FedNoisy library [Liang *et al.*, 2023] and include Fed-SCE [Wang *et al.*, 2019], Fed-Coteaching [Han *et al.*, 2018], and Fed-Mixup [Zhang *et al.*, 2017].

Tables 1, 2, and 3 show the average (5 trials) and standard deviation of the best test accuracies on CIFAR-10/100 and Clothing1M. FedES achieves the best performance among comparison methods by adapting to clients with varying noise rates. Without a specific solution to noise heterogeneity, the six FNLL methods fail to fully improve performance. All comparison methods suffer from performance degradation as noise rates increase and the shift from IID to Non-IID occurs. Though general de-noise strategies exhibit better performance than binary methods, which divide clients as noisy or clean, simply employing these strategies on clients is sub-optimal. Comparatively, FedES consistently outperforms other methods by a large margin.

5.4 Ablation Study

How Each Step in Federated Noise Estimation Works

CIFAR-10/100 datasets are used to conduct experiments on the first stage of FedES. The aim is to examine the impact of each step on the estimation of noisy clients. Results are summarized in Table 4. This study compares local and global distributions using two methods: Mean and Earth Mover's Distance (EMD). The former uses the means of a Gaussian Mixture Model (GMM) of all local CoPs, and the latter measures the distance between the two distributions. The Sign is applied to EMD to determine if a positive or negative sign should be used. The number of training rounds in the first stage of all methods is kept constant for fairness in comparisons. The methods' effectiveness in estimating noisy clients is evaluated using mean squared error (MSE) between the client-wise indicator ρ_n and actual data quality q_n.



(a) CIFAR-10 (b) CIFAR-100

Figure 2: Comparison on data quality estimation. Settings: CIFAR-10 dataset ($\mu = 0.5$, noise type: asymmetric, data partition: Non-IID) and CIFAR-100 dataset ($\mu = 0.5$ noise type: asymmetric, data partition: Non-IID)

In Table 4, the first two indicators from Fair [Deng *et al.*, 2021] and S-FedAvg [Nagalapatti and Narayanam, 2021] rely on a limited dataset, leading to poor results. Using mean for local distribution improves MSE performance, while EMD can enhance it further by measuring the distance between local and global distribution. FedES, a combination of Sign and EMD, accurately estimates data quality values of noisy clients, laying a strong foundation for parameter-adaptive updating. Compared to Fair and S-FedAvg, which only identify high and low-quality clients, FedES stands out in data quality estimation. See Figure 2 for a visual comparison.

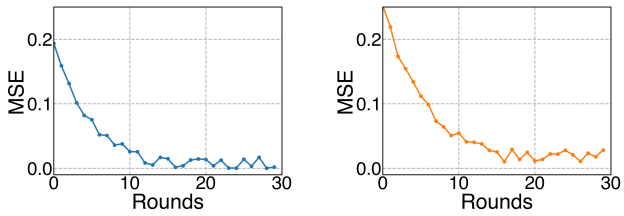
How pre-training Affects Federated Noise Estimation

Baseline	Binary De-noise			General De-noise			Ours
FedAvg	S-FedAvg	Fair	FedNoRo	Fed-Mixup	Fed-Coteaching	Fed-SCE	Fed-ES
70.52±0.23	71.33±0.04	71.25±0.50	71.05±0.14	72.61±0.27	71.35±0.23	72.57±0.12	73.03±0.14

Table 3: Test Accuracy (%) comparison results on Clothing1M datasets under real-world label noise

Indicator	Mean	EMD	Sign	CIFAR-10	CIFAR-100
$\hat{q}_n$	✗	✗	✗	0.07	0.13
$P_\phi[n]$	✗	✗	✗	0.05	0.11
ρ_n	✓	✗	✗	0.03	0.09
ρ_n	✗	✓	✗	0.02	0.05
ρ_n	✗	✓	✓	**0.01**	**0.02**

Table 4: MSE comparison results of the first stage ablation study in FedES. Settings: CIFAR-10 dataset ($\mu = 0.5$, noise type: asymmetric, data partition: Non-IID) and CIFAR-100 dataset ($\mu = 0.5$ noise type: asymmetric, data partition: Non-IID)



(a) CIFAR-10 (b) CIFAR-100

Figure 3: Ablation study of s_1 for pre-training. Settings: CIFAR-10 dataset ($\mu = 0.5$, noise type: asymmetric, data partition: Non-IID) and CIFAR-100 dataset ($\mu = 0.3$, noise type: asymmetric, data partition: Non-IID)

In the first stage, the global model warms up for s_1 rounds with the FedAvg updating process before federated noise estimation. Due to the lack of prior knowledge on multi-source label noise, it is hard to exactly determine the optimal setting of s_1. To evaluate the effect of s_1, federated noise estimation is conducted under different settings of s_1 on the CIFAR-10 dataset ($\mu = 0.5$, noise type: asymmetric, data partition: Non-IID) and CIFAR-100 dataset ($\mu = 0.3$, noise type: asymmetric, data partition: Non-IID) as shown in Fig. 3. It can be seen that, after a small number of pre-training rounds, federated noise estimation performance becomes stable in a certain range. In other words, the setting of s_1 would not significantly affect FedES's performance.

How SeGD and NaAgg benefit Noise-Robust FL

Table 5 shows FedES test accuracy with different de-noise schemes. When noisy clients are involved in federated learning without any de-noise schemes, there is a significant degradation in classification performance. Although client selection (CS) can increase test accuracy by discarding clients with data quality below 0.5, it still results in information loss for discarded clients and noisy labels for retained clients. A softer solution is to allow all clients to engage in and apply SeGD to perform selective gradient decay on model parameters for each client, resulting in a significant improvement

CS	SeGD	NaAgg	CIFAR-10	CIFAR-100
✗	✗	✗	58.75±0.06	53.01±0.80
✓	✗	✗	67.91±0.15	59.97±0.29
✗	✓	✗	76.15±0.82	62.76±0.64
✗	✗	✓	74.26±0.97	61.17±0.18
✗	✓	✓	**85.74±0.99**	**65.11±0.09**

Table 5: Test Accuracy comparison results of the second stage ablation study in FedES. Settings: CIFAR-10 dataset ($\mu = 0.3$, noise type: symmetric, data partition: Non-IID) and CIFAR-100 dataset ($\mu = 0.3$, noise type: asymmetric, data partition: IID)

in performance. NaAgg assigns updated local models corresponding criticality levels during aggregation, further improving the global model's performance. Using SeGD and NaAgg can help explore valid information in noisy clients meanwhile maintaining the criticality of model parameters and clients, considerably outperforming the classification performance when there are only high data quality clients (i.e., larger than 0.5).

6 Conclusion

In conclusion, this paper introduces a pioneering approach to federated label noise modeling, departing from the binary noisy-vs-clean clients initialization to accommodate the nuanced variations in noise rates among clients. The proposed Federated Early-Stopping (FedES) framework is specifically crafted to address the challenges posed by clients exhibiting diverse noise rates. Our experimental results, conducted on a diverse set of datasets containing both synthetic and real-world label noise, unequivocally demonstrate the superior performance of FedES compared to state-of-the-art federated noisy label learning (FNLL) methods. This empirical evidence substantiates the efficacy of our federated noise model and the FedES framework in achieving noise-robust learning in federated settings. We anticipate that our contributions will not only advance the field of federated learning but also inspire further research in the development of practical and adaptable federated learning frameworks for real-world applications.

Acknowledgments

This work is supported by the National Key Research and Development Plan of China No.2021YFC2501202, National Natural Science Foundation of China No.62202455, 61972383, Beijing Municipal Science & Technology Commission No.Z221100002722009, Beijing Natural Science Foundation No.L222100, and China Scholarship Council No.202204910370.

References

[Arpit *et al.*, 2017] Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017.

[Bai *et al.*, 2021] Yingbin Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, and Tongliang Liu. Understanding and improving early stopping for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:24392–24403, 2021.

[Bernhardt *et al.*, 2022] Mélanie Bernhardt, Daniel C Castro, Ryutaro Tanno, Anton Schwaighofer, Kerem C Tezcan, Miguel Monteiro, Shruthi Bannur, Matthew P Lungren, Aditya Nori, Ben Glocker, et al. Active label cleaning for improved dataset quality under resource constraints. *Nature communications*, 13(1):1161, 2022.

[Dayan *et al.*, 2021] Ittai Dayan, Holger R Roth, Aoxiao Zhong, Ahmed Harouni, Amilcare Gentili, Anas Z Abidin, Andrew Liu, Anthony Beardsworth Costa, Bradford J Wood, Chien-Sung Tsai, et al. Federated learning for predicting clinical outcomes in patients with covid-19. *Nature medicine*, 27(10):1735–1743, 2021.

[Deng *et al.*, 2021] Yongheng Deng, Feng Lyu, Ju Ren, Yi-Chao Chen, Peng Yang, Yuezhi Zhou, and Yaoxue Zhang. Fair: Quality-aware federated learning with precise user incentive and model aggregation. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2021.

[Fang and Ye, 2022] Xiuwen Fang and Mang Ye. Robust federated learning with noisy and heterogeneous clients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10072–10081, 2022.

[Fischer, 2011] Hans Fischer. *A history of the central limit theorem: from classical to modern probability theory*, volume 4. Springer, 2011.

[Frankle and Carbin, 2018] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

[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. *Advances in neural information processing systems*, 31, 2018.

[Han *et al.*, 2020] Bo Han, Gang Niu, Xingrui Yu, Quanming Yao, Miao Xu, Ivor Tsang, and Masashi Sugiyama. Sigua: Forgetting may make learning with noisy labels more robust. In *International Conference on Machine Learning*, pages 4006–4016. PMLR, 2020.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Ju *et al.*, 2022] Lie Ju, Xin Wang, Lin Wang, Dwarikanath Mahapatra, Xin Zhao, Quan Zhou, Tongliang Liu, and Zongyuan Ge. Improving medical images classification with label noise using dual-uncertainty estimation. *IEEE transactions on medical imaging*, 41(6):1533–1546, 2022.

[Kairouz *et al.*, 2021] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[Karimi *et al.*, 2020] Davood Karimi, Haoran Dou, Simon K Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical image analysis*, 65:101759, 2020.

[Krizhevsky *et al.*, 2009] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[Li *et al.*, 2020] Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *International conference on artificial intelligence and statistics*, pages 4313–4324. PMLR, 2020.

[Li *et al.*, 2021] 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.

[Liang *et al.*, 2023] Siqi Liang, Jintao Huang, Dun Zeng, Junyuan Hong, Jiayu Zhou, and Zenglin Xu. Fednoisy: Federated noisy label learning benchmark. *arXiv preprint arXiv:2306.11650*, 2023.

[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 *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[Nagalapatti and Narayanam, 2021] Lokesh Nagalapatti and Ramasuri Narayanam. Game of gradients: Mitigating irrelevant clients in federated learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9046–9054, May 2021.

[Nguyen *et al.*, 2019] Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, and Thomas Brox. Self: Learning to filter noisy labels with self-ensembling. *arXiv preprint arXiv:1910.01842*, 2019.

[Rieke *et al.*, 2020] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):119, 2020.

[Rolnick *et al.*, 2017] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.

[Rubner *et al.*, 2000] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99, 2000.

[Song *et al.*, 2022] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[Tanaka *et al.*, 2018] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5552–5560, 2018.

[Tuor *et al.*, 2021] Tiffany Tuor, Shiqiang Wang, Bong Jun Ko, Changchang Liu, and Kin K Leung. Overcoming noisy and irrelevant data in federated learning. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5020–5027. IEEE, 2021.

[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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 322–330, 2019.

[Wu *et al.*, 2023] Nannan Wu, Li Yu, Xuefeng Jiang, Kwang-Ting Cheng, and Zengqiang Yan. Fednoro: Towards noise-robust federated learning by addressing class imbalance and label noise heterogeneity. *arXiv preprint arXiv:2305.05230*, 2023.

[Xia *et al.*, 2020] Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Jiankang Deng, Jiatong Li, and Yinian Mao. Extended t: Learning with mixed closed-set and open-set noisy labels. *arXiv preprint arXiv:2012.00932*, 2020.

[Xia *et al.*, 2021] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *International conference on learning representations*, 2021.

[Xiao *et al.*, 2015] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015.

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10184–10193, 2022.

[Zeng *et al.*, 2022] Bixiao Zeng, Xiaodong Yang, Yiqiang Chen, Hanchao Yu, and Yingwei Zhang. Clc: A consensus-based label correction approach in federated learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(5):1–23, 2022.

[Zeng *et al.*, 2023] Bixiao Zeng, Xiaodong Yang, Yiqiang Chen, Hanchao Yu, Chunyu Hu, and Yingwei Zhang. Federated data quality assessment approach: Robust learning with mixed label noise. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

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