Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.075102

ARTICLE

FNRE: A Novel Approach to Heterogeneous Label Noise Rates Estimation in Federated Learning

Qian Rong[1], Lu Zhang[2], Ling Yuan[1,*], Zhong Yang[3] and Guohui Li[3]

[1]School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China
[2]School of Cyber Science and Engineering, Huazhong University of Science and Technology, Wuhan, China
[3]School of Software Engineering, Huazhong University of Science and Technology, Wuhan, China

*Corresponding Author: Ling Yuan. Email: cherryyuanling@hust.edu.cn

Received: 24 October 2025; Accepted: 03 March 2026; Published: 08 May 2026

ABSTRACT: Federated learning (FL) enables collaborative model training across decentralized clients without sharing raw data, thereby preserving privacy. However, in real-world FL deployments—such as sensor-based activity recognition, wearable health monitoring, and industrial Internet of Things, where local training data often suffer from heterogeneous noisy labels due to diverse collection environments, sensor limitations, and labeling errors. These noisy labels, typically distributed unevenly across clients due to differences in client-side annotation, exacerbate Non-Independent and Identically Distributed (non-IID) data issues, leading to biased updates, unstable convergence, and degraded global model performance. Accurate estimation of client-specific noise rates is therefore crucial for adaptive algorithm selection, personalized parameter tuning, noise-aware aggregation, and resource allocation in FL. Existing noise rate estimation methods, primarily developed for centralized settings, require client-specific clean validation sets or prior knowledge of noise, making them impractical in privacy-sensitive federated settings. In this work, we propose a federated noise rate estimation (FNRE) method that eliminates the need for per-client clean datasets or prior knowledge of noise. Our approach requires only a minimal assumption—at least one client with a small clean validation set—and leverages the global model's predictions to estimate local noise rates across all clients. Specifically, the method computes global prediction accuracy using data from the small, clean subset of clients, broadcasts this accuracy to all participants, and enables each client to infer its noise rate using its own annotated labels and the predicted label sequence. We further provide a theoretical analysis with provable error bounds. Extensive experiments on image classification (CIFAR-10, CIFAR-100) and sensor-based activity recognition (Widar, WISDM-W) under various synthetic and real-world noisy label settings demonstrate that our method achieves a noise rate estimation Mean Absolute Error (MAE) of only 0.82%–2.19%, outperforming state-of-the-art baselines by 29.8%–49.9% on average while maintaining practicality in privacy-sensitive federated environments.

KEYWORDS: Noise rate estimation; noisy labels; federated learning; sensor-based activity recognition

1 Introduction

Federated learning (FL) is a distributed machine learning paradigm that enables collaborative model training across multiple devices or organizations without centralizing data [1–3]. This approach inherently protects data privacy by keeping local data on client devices, while only sharing model updates. In many real-world applications, especially those involving large-scale sensor networks—such as sensor-based activity recognition, wearable health monitoring systems, and industrial Internet of Things infrastructures—FL provides an effective framework for processing the massive volume of distributed sensing data while

complying with privacy constraints [4–6]. However, in these practical scenarios, the quality of client-side local data is often difficult to guarantee due to environmental diversity during data collection, sensor limitations or malfunctions, and human or algorithmic errors in the labeling process [7,8]. A particularly prevalent issue among these is 'noisy labels,' where some training samples are assigned incorrect class labels. Studies indicate that real-world datasets can contain label noise rates ranging from 8.0% to 38.5% [9,10], which severely degrade model performance [11].

Moreover, in the federated learning paradigm, this noisy label is often heterogeneously distributed across client devices, leading to varied noise rates and mislabeling types [12]. This heterogeneity in noisy labels exacerbates the already existing Non-Independent and Identically Distributed (non-IID) nature of federated data, leading to amplified divergence in local model updates, biased gradient aggregation, and unstable convergence in the global model training process [13,14]. A one-size-fits-all noise-robust learning strategy often fails to effectively address the diverse noise characteristics across clients. Therefore, accurately estimating label noise rates at each client becomes crucial, as it enables data quality diagnosis, supports the design of client-specific noise-robust learning strategies [12,15,16], personalized hyperparameter tuning [17], noise-aware aggregation mechanisms [18], and resource allocation optimization [19]—ultimately enhancing both robustness and generalization in federated settings.

Table 1: Comparison of Methods in Terms of Complexity and Requirements. **Noise type** represents prior knowledge of the type of noise, **Multi-class** represents Multi-class classification, n_k denotes the number of training samples on client k, d represents the dimensionality of the training data, ✓ indicates support or requirement, while × indicates lack of support or no requirement.

Method	Time Complexity	Validation Set	Noise Type	Multi-class	Federated Setting
ROC [20]	$O(n_k \log n_k)$	✓	×	×	×
TiCE [21]	$O(dn_k)$	✓	×	×	×
KM [22]	$O((d+n_k)^2)$	✓	×	×	×
INCV [23]	$O(n_k)$	×	✓	✓	×
IR [24]	$O(n_k^2)$	×	✓	×	×
MPEIA [25]	$O(dn_k^2)$	✓	×	✓	×
DEDPUL [26]	$O(n_k \log n_k)$	✓	×	×	×
(TED)n [27]	$O(n_k \log n_k)$	✓	×	×	×
SuDPL [28]	$O(n_k)$	✓	×	×	×
Ours	$O(n_k)$	✓	×	✓	✓

Most existing studies on noise rate estimation have been developed for centralized learning settings, leveraging approaches such as **noise transition matrices** [29–31], **cross-validation** [23,32], or **mixture proportion estimation (MPE)** [20,25,28]. As summarized in Table 1, these centralized methods lack the inherent capability for distributed deployment. Consequently, directly transposing them to a federated environment would necessitate treating each client as an isolated estimation task. However, this imposes a strict constraint: every client must possess either prior knowledge of the specific noise type or a local clean auxiliary dataset. In practical federated scenarios involving hundreds of clients with heterogeneous noise, assuming universal access to such prior knowledge or auxiliary data is unrealistic. Therefore, existing centralized noise rate estimation methods are fundamentally incompatible with the constraints of federated learning.

To address the above challenges, we propose a federated noise rate estimation (FNRE) algorithm that requires only a minimal condition: at least one client possessing a small, clean validation dataset. Specifically,

our method first leverages the global model to generate predicted labels for all samples on each client, forming a local sequence of predicted labels, called the auxiliary label sequence. Then, the global model's prediction accuracy is computed using clean validation data from the small subset of participating clients, and the resulting accuracy values are uploaded to the server. Finally, the server broadcasts the global model's prediction accuracy to all clients, enabling each client to infer its local noise rate by combining the auxiliary label sequence, the global prediction accuracy, and its own annotated labels.

In summary, the key contributions of this paper are:

- **Federated Noise Rate Estimation:** We propose a federated noise rate estimation algorithm that achieves privacy-preserving distributed noise rate estimation by transmitting only non-sensitive local accuracy statistics, thereby significantly relaxing constraints on auxiliary data. Specifically, our method enables network-wide estimation provided that at least one client possesses a minimal clean validation set (e.g., as few as 300 samples). This capability substantially reduces barriers to practical deployment in real-world applications.
- **Privacy-Preserving Estimation with Error Bounds:** To the best of our knowledge, the proposed Federated Noise Rate Estimation (FNRE) algorithm is the first framework specifically tailored for estimating heterogeneous label noise rates in federated settings. It inherently supports distributed deployment while strictly preserving privacy. Furthermore, we provide a rigorous theoretical analysis, establishing convergence error bounds to mathematically guarantee the reliability of our estimator.
- **Experimental Validation:** Extensive experiments on computer vision tasks (CIFAR-10, CIFAR-100) and sensor-based activity recognition (Widar, WISDM-W) demonstrate superior estimation accuracy compared to baselines.

2 Related Work

In this section, we primarily focus on introducing noise estimation methods.

Noise transition matrix. The core idea is to model the probability $T_{ij} = P(\tilde{y} = j \mid y = i)$ of a clean label $y = i$ being flipped to a noisy label $\tilde{y} = j$. Once the transition matrix T is estimated, the overall noise rate can be derived directly from its off-diagonal elements. Early works such as [29,30] introduced the use of anchor points—cleanly labeled samples whose predicted posterior is close to one-hot distributions—enabling exact or approximate estimation of certain rows of T. MPEIA [25] proposed an efficient mixture proportion estimation approach based on a linear independence assumption, leveraging kernel mean embeddings and quadratic programming to accurately recover the noise transition matrix. While Noise Transition Matrix based methods allow explicit calculation of noise rates and have shown promise in centralized scenarios, their direct application to federated noise rate estimation faces several critical limitations. First, they typically require clean anchor samples for each class, which is often infeasible in federated environments due to privacy constraints and the heterogeneous nature of client datasets. Second, applying such methods in FL would necessitate estimating a separate transition matrix for each client, implying that each client must independently possess sufficient clean samples—an assumption rarely satisfied in practice. Third, estimation accuracy can degrade significantly as the number of classes increases, particularly when client datasets are small or class-imbalanced, which is common in federated settings.

Cross-validation. Cross-validation-based methods estimate noisy label rates by partitioning the dataset, training models on one subset, and evaluating their performance on the other. For example, INCV [23] proposed randomly splitting a noisy training set into two disjoint subsets, assuming that their noise transition matrices are identical, and then inferring the noise rate from the test accuracy obtained via cross-validation. While conceptually simple, this approach requires prior knowledge of the type of label

noise (e.g., symmetric or asymmetric). Extending this approach to federated learning presents significant challenges, as the estimation process must be executed independently on each client, requiring prior knowledge of the specific type of noisy labels at each client.

Mixture proportion estimation. Mixture proportion estimation (MPE) methods estimate the fraction of examples belonging to a target distribution within a mixture distribution, and have been adapted to estimate noisy label rates. For instance, KM [22] embeds distributions into a reproducing kernel Hilbert space (RKHS) and uses convex optimization to achieve statistically consistent mixture proportion estimates in a non-parametric setting. IR [24] estimates the noise rate by evaluating the conditional probability of noisy samples, but the method's estimation accuracy is limited by model assumptions, extremely small conditional probabilities, and the choice of parameters in density estimation. By strengthening certain distributional assumptions, the ROC [20] introduces a mixture proportion estimation approach that estimates the true mixture proportion, while also providing a practical convergence rate analysis for MPE. Building on this, TiCE [21] employs decision tree induction to identify high-purity positive examples in positive-unlabeled datasets, leveraging lower bound properties of label frequencies. DEDPUL [26] jointly estimates mixture proportions and performs Positive-Unlabeled (PU) classification by modeling density differences and calibrating posterior probabilities. TED^n [27] estimates the noisy label proportion using the Best Bin Estimation method, which leverages the proportion of instances in classifier score histograms corresponding to near-pure negative regions; however, its estimation accuracy degrades significantly when class distributions heavily overlap, pure regions are absent, or binning strategies are suboptimal. SuMPE [28] proposes a novel method for mixture proportion estimation under a relaxed irreducibility assumption, leveraging density ratio or conditional probability estimation to avoid reliance on complete class separability; however, its accuracy may be limited in cases of extreme class distribution overlap or when the estimation model is unstable. Although MPE-based methods can achieve accurate noise rate estimates and are theoretically well-founded, their application to federated learning is hindered by several factors: they often require a certain number of clean samples for each class as positive examples, rely on accurate modeling of class-conditional distributions, and may incur substantial computational cost when applied independently on each client. These requirements are difficult to satisfy in federated environments characterized by strict privacy constraints, heterogeneous and non-IID data, and resource-limited devices, making centralized MPE methods impractical for federated noise rate estimation.

3 Problem Setup

Consider a federated learning system with K clients, where each client k holds a private dataset $\mathcal{D}_k = \{(x_i^k, \tilde{y}_i^k)\}_{i=1}^{n_k}$. Here, $x_i^k \in \mathcal{X}$ denotes the i-th data sample of client k, $\tilde{y}_i^k \in \{1, 2, \ldots, C\}$ is its annotated label, and $y_i^k \in \{1, 2, \ldots, C\}$ represents the corresponding (unknown) true label, with C being the number of classes in the classification task, n_k is the number of training samples on the client k. Let the annotation labels be arranged into the sequence $L_a^k = [\tilde{y}_1^k, \tilde{y}_2^k, \ldots, \tilde{y}_{n_k}^k]$, and the true labels into $L_t^k = [y_1^k, y_2^k, \ldots, y_{n_k}^k]$.

For each client k, the overall noise rate for client k is then given by:

$$\epsilon_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbb{I}(\tilde{y}_i^k \neq y_i^k), \tag{1}$$

where $\mathbb{I}(\cdot)$ is the indicator function. **The goal is to estimate the overall noise rates $\{\epsilon_k\}_{k=1}^K$ for all clients, without accessing the raw data** $\{(x_i^k, \tilde{y}_i^k)\}$ directly.

4 Proposed Method

4.1 Theoretical Framework for Noise Rate Estimation

To estimate the overall noise rates $\{\epsilon_k\}_{k=1}^{K}$ for each client in a federated learning setting, we first introduce the foundational principles underlying our noise rate estimation approach.

4.1.1 Auxiliary Labels Sequence

If the true label sequence L_t^k were available, the overall noise rate ϵ_k could be directly computed. However, L_t^k is typically unavailable in practice. Therefore, we relax the assumptions and consider an auxiliary label sequence $L_{\text{aux}}^k = [\hat{y}_0^k, \hat{y}_1^k, \ldots, \hat{y}_n^k]$, which may also contain erroneous annotations. While the exact positions of the noisy labels in L_{aux}^k are unknown, we assume that the **noise rate of the auxiliary labels** ϵ_{aux} is known a priori. Under this relaxed setting, we aim to utilize L_{aux}^k and its known noise rate ϵ_{aux} to estimate the original noise rates ϵ_k in $\mathcal{D}_k$.

4.1.2 Agreement Probability

We define γ as the probability that the labels at corresponding positions in the auxiliary label sequence $L_{\text{aux}}^k = [\hat{y}_i^k]_{i=1}^{n_k}$ and the annotation label sequence $L_a^k = [\tilde{y}_i^k]_{i=1}^{n_k}$ are identical. Formally, γ is computed as:

$$\gamma = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbb{I}(\hat{y}_i^k = \tilde{y}_i^k). \tag{2}$$

This probability γ quantifies the agreement between the auxiliary and annotation label sequence, providing a measure of their alignment. The agreement between the auxiliary label sequence L_{aux}^k and the annotation label sequence L_a^k arises from two scenarios:

1. **Correct Agreement:** Both labels are correct and match the true label y_i^k in L_t^k.
2. **Coincidental Error:** Both labels are incorrect but erroneously coincide.

Thus, the probability γ can be decomposed into two components governed by the noise rates ϵ_k (of L_a^k) and ϵ_{aux} (of L_{aux}^k):

$$\gamma = \underbrace{(1-\epsilon_k)(1-\epsilon_{\text{aux}})}_{\text{Correct Agreement}} + \underbrace{\epsilon_k \epsilon_{\text{aux}} \cdot \rho}_{\text{Coincidental Error}} . \tag{3}$$

4.1.3 Mathematical Formulation: Deriving Noise Rate from Agreement

Formally, ρ can be expressed as the sum of products of mislabeling probabilities from the auxiliary label sequence and the annotation label sequence. Let T_k and T_{aux} denote the noise transition matrices for L_a^k and L_{aux}^k, respectively, where $T_k^{ij} = P(\tilde{y}_i^k = j \mid y_i^k = i)$ and $T_{\text{aux}}^{ij} = P(\hat{y}_i^k = j \mid y_i^k = i)$. Then, for samples with true label $y_i^k = i$, the probability that both annotation labels $\tilde{y}_i^k$ and auxiliary labels $\hat{y}_i^k$ are incorrectly assigned to the same class $j \neq i$ is $T_k^{ij} \cdot T_{\text{aux}}^{ij}$. Summing over all possible incorrect classes $j \neq i$ and weighted by the class distribution, we have:

$$\rho = \sum_{i=1}^{C} \left(\frac{n_k^i}{n_k} \sum_{j \neq i} T_k^{ij} \cdot T_{\text{aux}}^{ij} \right), \tag{4}$$

where n_k^i is the number of samples with true label i in $\mathcal{D}_k$, and $\frac{n_k^i}{n_k}$ represents the proportion of class i. In practical scenarios, the noise transition matrices T_k^{ij} and T_{aux}^{ij}, as well as the class-wise sample counts n_k^i,

are typically inaccessible. If T_k^{ij} were known, the noise rate ϵ_k could be directly inferred without additional estimation procedures. Therefore, it is impossible to compute ρ exactly.

We interpret ρ as a hyperparameter contingent upon the underlying noise structure. In the absence of prior knowledge regarding this structure, we adopt a uniform label noise assumption in accordance with the Maximum Entropy Principle [33]. This approach enables us to determine ρ in a way that provides a tractable estimate with minimal prior bias. Conversely, when prior information about the noise characteristics is available, ρ can be calibrated to reflect those specific characteristics.

Under the uniform noise assumption, each incorrect label is assigned uniformly at random to one of the other $C-1$ classes, the noise transition probabilities satisfy $T_k^{ij} = T_{\text{aux}}^{ij} = \frac{1}{C-1}$ for $j \neq i$, and ρ reduces to:

$$\rho = \frac{1}{C-1}. \tag{5}$$

This simplification reduces the original expression for γ to:

$$\gamma = (1-\epsilon_k)(1-\epsilon_{\text{aux}}) + \frac{\epsilon_k \epsilon_{\text{aux}}}{C-1}. \tag{6}$$

Based on the preceding analysis, with γ, ϵ_{aux}, and C known, the noise rate ϵ_k for the label sequence L_a can be explicitly derived. We solve for ϵ_k through algebraic manipulation:

$$\epsilon_k = \frac{(1-\epsilon_{\text{aux}}) - \gamma}{(1-\epsilon_{\text{aux}}) - \frac{\epsilon_{\text{aux}}}{C-1}}. \tag{7}$$

This closed-form solution enables efficient estimation of ϵ_k without iterative optimization, leveraging only the observable agreement rate γ, the auxiliary noise rate ϵ_{aux}, and the number of classes C.

4.1.4 Rationality of the Uniform Label Noise Assumption

The assumption of uniform label noise is grounded in the Principle of Maximum Entropy [33]. In the absence of specific prior knowledge about the noise structure (i.e., the noise transition matrix), the uniform distribution is the most conservative choice, as it makes the fewest assumptions and introduces the least bias. In a privacy-preserving federated setting, obtaining client-specific noise transition matrices is often infeasible, making the uniform assumption a practical and principled starting point.

Empirically, as detailed in Section 5.7, we verify that our method is robust to violations of this assumption. Even under non-uniform label noise settings (e.g., asymmetric or instance-dependent noise), adopting the uniform label noise assumption yields virtually no degradation in estimation accuracy. This empirical evidence strongly supports adopting this assumption for practicality and efficiency, without sacrificing significant accuracy.

4.1.5 Practical Implementation and Validation Data Handling

A neural network is trained to generate the auxiliary label sequence $L_{\text{aux}}^k = [\hat{y}_i^k]_{i=1}^{n_k}$ based on its classification predictions:

$$\hat{y}_i^k = \text{argmax}(f(x_i)) \tag{8}$$

where $f(\cdot)$ denotes the global model, and $\text{argmax}(\cdot)$ represents the index of the maximum element, which corresponds to the predicted label of the neural network for a given sample.

In practical implementation, we assume access to a small set $S = \{(x_i, y_i)\}_{i=1}^{m}$ of training samples with known true labels. The prediction error of the global model $f(\cdot)$ on this validation set S is assessed. The noise rate ϵ_{aux} of L_{aux} is then equated to this predictive error. We demonstrate that robust performance can be achieved when the clean validation set S has a sample size m greater than 300, irrespective of the training set size. In most tasks, a test set is typically available to evaluate the model's performance. A small validation set can usually be constructed by sampling a few instances from the test set. If access to the test set is unavailable, a small set of samples can be labeled manually (e.g., 300 samples). For tasks with high labeling difficulty, such as medical images, experts can be enlisted to annotate a small number of samples. For high-labeling-difficulty tasks, multimodal large models or human annotators can be used for labeling. In cases where client data is completely inaccessible, the server can be treated as a client with validation data, and a small client can be constructed on the server to simulate the validation process.

Furthermore, in a federated learning setting, it is sufficient for only a single client to possess a small amount of clean validation data, rather than requiring every client to have such a dataset. From a privacy-preserving perspective, no raw data ever leaves local storage; only the global model's prediction accuracy on the validation set is communicated, ensuring that sensitive information remains protected while still enabling reliable noise rate estimation.

4.1.6 Error Bound Convergence Analysis

We establish rigorous theoretical foundations for our noise rate estimation methodology by deriving precise error bounds and statistical guarantees.

The error in estimating ϵ_k arises from two components: the estimation error of γ, denoted as $|\hat{\gamma} - \gamma| \leq \epsilon_\gamma$ and the estimation error of ϵ_{aux}, denoted as $|\hat{\epsilon}_{\text{aux}} - \epsilon_{\text{aux}}| \leq \epsilon^*_{\text{aux}}$. Using a first-order Taylor expansion, the error in ϵ_k can be approximated as:

$$|\hat{\epsilon}_k - \epsilon_k| \approx \left| \frac{\partial \epsilon_k}{\partial \gamma} \cdot \epsilon_\gamma + \frac{\partial \epsilon_k}{\partial \epsilon_{\text{aux}}} \cdot \epsilon^*_{\text{aux}} \right|. \tag{9}$$

The partial derivatives are computed as:

$$\frac{\partial \epsilon_k}{\partial \gamma} = \frac{-1}{(1 - \epsilon_{\text{aux}}) - \frac{\epsilon_{\text{aux}}}{C-1}}, \tag{10}$$

$$\frac{\partial \epsilon_k}{\partial \epsilon_{\text{aux}}} = \frac{-(1-\gamma) - \frac{\gamma}{(C-1)}}{\left[(1 - \epsilon_{\text{aux}}) - \frac{\epsilon_{\text{aux}}}{C-1}\right]^2}. \tag{11}$$

Thus, the error bound is:

$$|\hat{\epsilon}_k - \epsilon_k| \leq \left| \frac{\epsilon_\gamma}{(1 - \epsilon_{\text{aux}}) - \frac{\epsilon_{\text{aux}}}{C-1}} \right| + \left| \frac{(1-\gamma) + \frac{\gamma}{(C-1)}}{\left[(1 - \epsilon_{\text{aux}}) - \frac{\epsilon_{\text{aux}}}{C-1}\right]^2} \cdot \epsilon^*_{\text{aux}} \right|. \tag{12}$$

The estimation errors satisfy:

$$\epsilon_\gamma \leq \sqrt{\frac{\log(2/\delta)}{2n_k}}, \quad \epsilon^*_{\text{aux}} \leq \sqrt{\frac{\log(2/\delta)}{2m}}, \tag{13}$$

where δ is the confidence level, n_k is the number of samples on client k, and m is the size of the clean validation set. Substituting the above results, the error bound becomes:

$$|\hat{\epsilon}_k - \epsilon_k| \leq \left| \frac{\sqrt{\frac{\log(2/\delta)}{2n_k}}}{(1-\epsilon_{\text{aux}}) - \frac{\epsilon_{\text{aux}}}{C-1}} \right| + \left| \frac{(1-\gamma) + \frac{\gamma}{(C-1)}}{\left[(1-\epsilon_{\text{aux}}) - \frac{\epsilon_{\text{aux}}}{C-1}\right]^2} \cdot \sqrt{\frac{\log(2/\delta)}{2m}} \right|. \tag{14}$$

The error bound converges at a rate of $O\left(\frac{1}{\sqrt{n_k}} + \frac{1}{\sqrt{m}}\right)$, demonstrating the reliability and scalability of the proposed method.

4.2 Federated Noise Rate Estimation Framework

In the federated learning paradigm, we establish a privacy-preserving framework for noise rate estimation that operates under stringent data locality constraints. The fundamental challenge lies in accurately estimating label noise rates across heterogeneous clients while maintaining data confidentiality. Fig. 1 shows an overview of the Federated Noise Rate Estimation Framework.

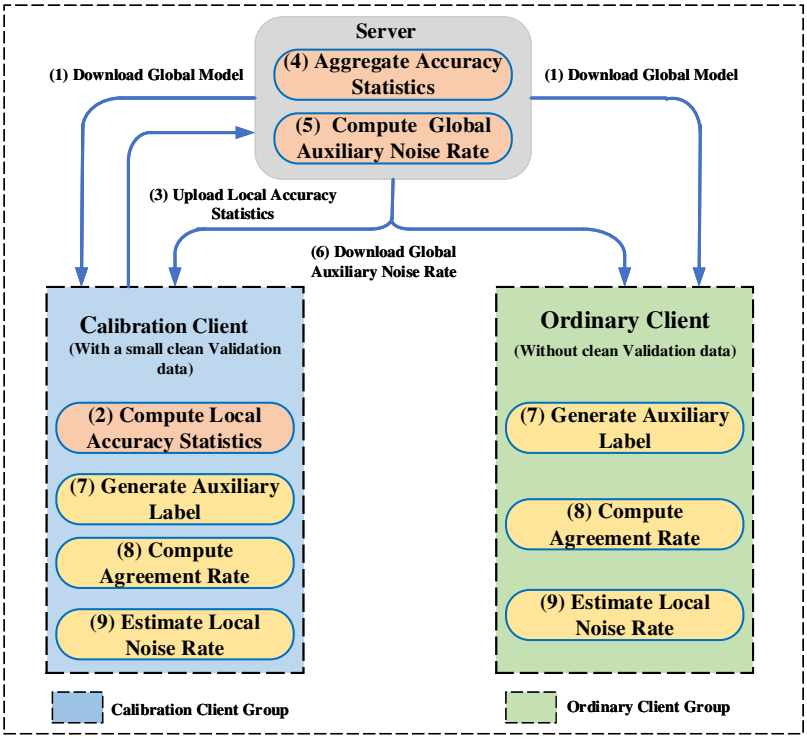


Figure 1: An overview of federated noise rate estimation framework.

4.2.1 Assumptions

The key Assumptions underlying our framework are detailed as follows:

- *Auxiliary Data Scarcity*: $m \ll \sum_{k=1}^{K} n_k$, ensuring the auxiliary dataset remains significantly smaller than the aggregate of all client datasets. We only require a dataset significantly smaller than the aggregate of all client datasets (e.g., a few hundred samples), which is readily achievable in practical applications.
- *Distributed Auxiliary Access*: Only a subset $\mathcal{K}_{\text{cal}} \subseteq \{1, 2, \ldots, K\}$ of clients, termed *calibration clients*, have access to portions of $\mathcal{D}^{\text{aux}}$, $|\mathcal{K}_{\text{cal}}|$ can be equal to 1. This allows the auxiliary data to reside

on any number of clients—minimally just one. This feature provides substantial flexibility within the federated environment, thereby drastically reducing the deployment requirements compared to existing approaches.

4.2.2 Algorithmic Framework

Our federated noise estimation protocol (Algorithm 1) operates through four sequential phases, each designed to preserve privacy while enabling accurate quantification of the noise rate.

Phase I: Model Initialization and Local Training

The server initializes a global model f_θ parameterized by θ and distributes it to all participating clients. Each client k performs local training on their private dataset $\mathcal{D}_k$ using standard optimization procedures, producing a locally adapted model variant f_{θ_k}.

Phase II: Privacy-Preserving Accuracy Aggregation

Calibration clients compute local accuracy statistics on their auxiliary data subsets without exposing raw data:

For each calibration client $k \in \mathcal{K}_{\text{cal}}$:

$$c_k = \sum_{(x_i, y_i) \in \mathcal{D}_k^{\text{aux}}} \mathbb{I}(f_\theta(x_i) = y_i) \tag{15}$$

$$m_k = |\mathcal{D}_k^{\text{aux}}| \tag{16}$$

where $\mathcal{D}_k^{\text{aux}} \subseteq \mathcal{D}^{\text{aux}}$ denotes client k's portion of the auxiliary dataset. Clients transmit only the aggregated counts (c_k, m_k) to the server, ensuring individual sample privacy while enabling global accuracy estimation.

Phase III: Global Auxiliary Noise Rate Computation

The server aggregates received accuracy statistics to derive the global auxiliary noise rate:

$$\epsilon_{\text{aux}} = 1 - \frac{\sum_{k \in \mathcal{K}_{\text{cal}}} c_k}{\sum_{k \in \mathcal{K}_{\text{cal}}} m_k}. \tag{17}$$

This quantity represents the model's error rate on clean auxiliary data and serves as a calibration reference for local noise estimation. The server broadcasts ϵ_{aux} to all clients.

Phase IV: Local Noise Rate Inference

Each client k leverages the global auxiliary noise rate to estimate its local noise rate through the following procedure:

- **Auxiliary Label Generation:** Client k applies the trained global model f_θ to its local dataset, generating auxiliary predictions:

$$L_{\text{aux}}^k = [\hat{y}_i^k]_i^{n_k}, \quad \hat{y}_i^k = argmax(f_\theta(x_i^k)). \tag{18}$$

- **Agreement Rate Computation:** The agreement rate between model predictions and observed labels is calculated as:

$$\gamma_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbb{I}(\hat{y}_i^k = \tilde{y}_i^k). \tag{19}$$

- **Local Noise Rate Estimation:** The local noise rate is computed as:

$$\epsilon_k = \frac{(1-\epsilon_{\text{aux}}) - \gamma_k}{(1-\epsilon_{\text{aux}}) - \frac{\epsilon_{\text{aux}}}{C-1}}, \tag{20}$$

where C denotes the number of classes in the classification task.

Algorithm 1: Federated noise rate estimation protocol

Require: Number of clients K, Global model f_θ, local datasets $\mathcal{D}_k$, calibration clients $\mathcal{K}_{\text{cal}}$, auxiliary dataset $\mathcal{D}^{\text{aux}}$, number of classes C, selection rate P_{select}.
Ensure: Estimated local noise rates $\{\epsilon_k\}_{k=1}^K$.

Phase I: Model Initialization and Local Training

Server initializes a global model f_θ.
Server randomly selects participating clients $\mathcal{S}_t$ based on selection rate P_{select}.
Server distributes f_θ to all participating clients.
for each client $k \in \mathcal{S}_t$ in parallel **do**
    Performs local training, upload local model f_{θ_k}.
end for
Server aggregates local models {using} FedAvg.

Phase II: Privacy-Preserving Accuracy Aggregation

Calibration client downloads the global model
for each calibration client $k \in \mathcal{K}_{\text{cal}}$ in parallel **do**
    Compute correct predictions c_k using Eq. (15).
    Compute $m_k = |\mathcal{D}_k^{\text{aux}}|$.
    Transmit aggregated counts (c_k, m_k) to the server.
end for

Phase III: Global Auxiliary Noise Rate Computation

Server aggregates all accuracy statistics (c_k, m_k).
Compute global auxiliary noise rate ϵ_{aux} using Eq. (17).
Server broadcasts ϵ_{aux} to all clients.

Phase IV: Local Noise Rate Inference

for each client $k \in \{1, \ldots, K\}$ in parallel **do**
    Generate auxiliary prediction sequence L_{aux}^k.
    Calculate agreement rate γ_k using Eq. (19).
    Compute local noise rate ϵ_k using Equation Eq. (20).
end for
Return $\{\epsilon_k\}_{k=1}^K$

4.2.3 Communication Efficiency and Algorithmic Complexity

Our framework achieves exceptional privacy preservation and communication efficiency through a streamlined two-phase protocol. Privacy is guaranteed via strict data locality (raw datasets remain exclusively on client devices), and statistical aggregation (clients transmit only aggregated counts (c_k, m_k), insufficient for individual sample reconstruction). The communication architecture requires merely two rounds—calibration clients upload accuracy statistics to the server, followed by broadcast of the global auxiliary noise rate—resulting in $\mathcal{O}(|\mathcal{K}_{\text{cal}}|)$ communication complexity independent of dataset sizes or model dimensions.

Computationally, the framework scales linearly with local dataset sizes ($\mathcal{O}(n_k)$ client-side, $\mathcal{O}(|\mathcal{K}_{\text{cal}}|)$ server-side) while remaining independent of total network size, ensuring practical scalability for large-scale federated deployments with thousands of participating clients.

4.2.4 Extension to Class-Specific Noise Rate Estimation

The proposed federated noise rate estimation framework can be readily extended to estimate **class-specific noise rates** $\{\epsilon_k^c\}_{c=1}^C$ for each client. The core idea is to perform all aggregations and computations in a class-stratified manner.

Specifically, in **Phase II (Privacy-Preserving Accuracy Aggregation)**, calibration clients would compute and transmit accuracy statistics (c_k^c, m_k^c) for each individual class c present in their auxiliary data. The server, in **Phase III (Global Auxiliary Noise Rate Computation)**, aggregates these class-specific statistics to derive a global auxiliary noise rate ϵ_{aux}^c for each class. Finally, in **Phase IV (Local Noise Rate Inference)**, each client k would first stratify its local dataset based on observed noisy labels. Then, for each class c, it would compute a class-specific agreement rate γ_k^c between the global model's predictions and its local observed labels for that class. Utilizing the broadcasted ϵ_{aux}^c and the computed γ_k^c, each client can then infer its class-specific noise rate ϵ_k^c, providing a fine-grained characterization of label noise across different categories. This class-aware extension increases communication complexity linearly with the number of classes ($\mathcal{O}(C \cdot |\mathcal{K}_{\text{cal}}|)$) but remains highly efficient and privacy-preserving.

5 Experimental

5.1 Datasets

This section details the datasets utilized to evaluate the proposed methodologies. The selected datasets encompass computer vision tasks with natural images (CIFAR-10 [34], CIFAR-100 [34], CIFAR-10N [35], CIFAR-100N [35], ILSVRC 2012 [36]) and human activity recognition tasks using sensor data (Widar [37], WISDM-W [38]), providing a diverse evaluation framework that demonstrates the generalizability of our approach across different modalities and problem domains.

- **CIFAR-10 and CIFAR-100** [34] These are standard benchmark datasets for image classification. CIFAR-10 comprises 10 distinct classes, while CIFAR-100 extends this with 100 finer-grained categories. Both consist of 32 × 32 color images, representing a variety of common objects and animals, making them widely used for evaluating image recognition algorithms.
- **CIFAR-10N and CIFAR-100N** [35] These are noisy variants of CIFAR-10 and CIFAR-100, specifically designed to emulate real-world label noise challenges. Their labels contain various types of human-induced errors, such as aggregation errors and spurious correlations, which commonly arise from crowdsourcing or imperfect annotation processes. This characteristic makes them invaluable for research focused on robust learning in the presence of realistic label noise.
- **ILSVRC 2012** [36] The ILSVRC 2012 dataset, a subset of the ImageNet Large Scale Visual Recognition Challenge, consists of approximately 1.2 million training images, 50,000 validation images, and 100,000 test images, spanning 1000 object categories. Each image is labeled with a single class from a diverse set of object types, covering animals, vehicles, everyday objects, and scenes.
- **Widar** [37] The Widar dataset is designed for **contactless gesture recognition** based on Wi-Fi signals. The system collects fine-grained *Channel State Information (CSI)* using an Intel 5300 network interface card in a **3 × 3 MIMO** antenna configuration (three transmit, and three receive antennas). Wi-Fi access points and receiving antennas are fixed, and 17 participants perform 22 different gestures (e.g., pushing, pulling, sweeping, clapping) within a predefined area.

- **WISDM-W** [38] The WISDM (*Wireless Sensor Data Mining*) dataset is a widely used benchmark for **human activity recognition (HAR)** using accelerometer and gyroscope data collected from smartphones and smartwatches. A total of 51 participants performed **18 daily activities** (such as walking, jogging, sitting, standing, climbing stairs, opening doors, eating) for approximately 3 min each. To enhance its utility for machine learning tasks, similar activities were combined (e.g., various eating activities merged into "eating"), and uncommon or problematic activities were removed. Due to differences in data collection methodologies (e.g., smartwatch and smartphone data not always collected simultaneously or precisely synchronized), WISDM-W is treated as an independent dataset from its smartphone counterpart (WISDM-P).

Table 2 provides a comprehensive overview of these datasets, including their key characteristics and statistics.

Table 2: Summary of experimental datasets used in this study.

Dataset	Samples	Classes	Modality	Data Dimension	Characteristics
CIFAR-10	60,000	10	RGB Images	3 × 32 × 32	Natural object recognition
CIFAR-100	60,000	100	RGB Images	3 × 32 × 32	Fine-grained object classification
CIFAR-10N	60,000	10	RGB Images	3 × 32 × 32	Real human annotation noise
CIFAR-100N	60,000	100	RGB Images	3 × 32 × 32	Real human annotation noise
ILSVRC 2012	1,281,167	1000	RGB Images	3 × 227 × 227	Large-scale labeled datasets
Widar	16,594	22	Wireless Signal	22 × 20 × 20	Gesture recognition
WISDM-W	20,672	6	Accelerometer Gyroscope	200 × 6	Activity recognition, wearable sensors

5.2 Federated Learning Settings

This section details the specific configurations adopted for the federated learning environment, encompassing data distribution, noise injection mechanisms, validation data setup, and other key training parameters.

5.2.1 Data Distribution Configuration

We adopt a non-independent and identically distributed (non-IID) data partitioning strategy following the methodology established by [4]. Specifically, we employ the Dirichlet-based partitioning scheme to create realistic statistical heterogeneity across participating clients.

For classification tasks on datasets with C classes, we generate client-specific data distributions by sampling from a Dirichlet distribution $\mathrm{Dir}(\alpha)$, where the concentration parameter α controls the degree of non-IIDness. Lower values of α result in highly skewed class distributions, with each client predominantly containing samples from a few classes, while higher values of α produce more balanced class distributions

across clients. Additionally, we utilize the same Dirichlet parameter α to determine sample allocation, ensuring that both class distribution and sample quantity exhibit realistic heterogeneity patterns observed in practical federated environments.

5.2.2 Label Noise Configuration

We implement a comprehensive noise simulation strategy that encompasses three distinct noise paradigms across participating clients. Specifically, symmetric noise [39] is applied to one-third of clients, with label corruption occurring uniformly across all classes with equal probability. Asymmetric noise [39] is applied to another one-third of clients, featuring class-dependent noise patterns that reflect realistic annotation biases commonly observed in practical scenarios. The remaining one-third of clients are subject to instance-dependent noise [40], where the noise probability varies with sample characteristics and proximity to decision boundaries, thereby simulating the inherent uncertainty in borderline cases.

To simulate realistic noise rate distributions across the federated system, we employ three distinct probabilistic sampling strategies. The first strategy uses a Gaussian distribution $\mathcal{N}(0.2, 0.1)$ with a mean noise rate of 0.2, corresponding to moderate noise conditions. The second approach employs a Gaussian distribution $\mathcal{N}(0.5, 0.2)$ with a mean noise rate of 0.5, simulating high noise environments. The third strategy leverages a Beta distribution $\text{Beta}(0.5, 0.5)$, which provides U-shaped noise rate sampling across the entire range.

Each client randomly selects one noise rate from these distributions and is subsequently assigned a noise type through random allocation. This dual randomization process ensures diverse noise characteristics across the federated system, creating a heterogeneous noise environment that closely mirrors real-world federated learning deployments. This systematic configuration enables comprehensive evaluation of algorithm robustness under varying noise intensities and patterns, providing a rigorous assessment of the proposed methodology's effectiveness across diverse, challenging scenarios.

5.2.3 Validation Data Configuration

To maintain realistic evaluation conditions while ensuring fair performance assessment, we implement a limited clean validation setup. Specifically, only five randomly selected clients have access to clean validation data, with a total validation set size of 300 samples distributed across them. This configuration reflects practical federated scenarios where clean, high-quality data is scarce and unevenly distributed among participants.

5.2.4 System and Communication Parameters

The federated learning system parameters are configured according to dataset-specific requirements to ensure optimal convergence and fair comparison across different experimental conditions. Table 3 presents the detailed parameter configurations for each dataset.

Table 3: Federated learning system parameters for different datasets.

Parameter	CIFAR-10	CIFAR-100	Widar	WISDM-W
Number of clients	100	100	40	80
Communication rounds	100	100	80	80
Local epochs	5	5	3	3
Client participation rate	0.1	0.1	0.15	0.15

(Continued)

Table 3 (continued)

Parameter	CIFAR-10	CIFAR-100	Widar	WISDM-W
Local batch size	32	32	16	16
Learning rate	0.01	0.01	0.005	0.005
Dirichlet parameter α	0.4	0.4	0.4	0.4
Model	ResNet18	ResNet18	ResNet18	LSTM
Aggregation algorithm	FedAvg	FedAvg	FedAvg	FedAvg
Local optimization algorithm	SGD	SGD	SGD	SGD
Auxiliary dataset size m	300	300	300	300
Calibration clients	5	5	5	5

5.3 Baselines

We select MPEIA [25], INCV [23], KM [22], ROC [20], TiCE [21], IR [24], DEDPUL [26], $(\text{TED})^n$ [27], and SuDPL [28] as our baseline methods, which are introduced in detail in the *Related Work* section. Each baseline method is executed independently on every client. For methods originally designed for binary classification, including ROC, IR, TiCE, DEDPUL, $(\text{TED})^n$, KM, and SuMPE, we treat samples with the true label of a given class as positive examples and those with noisy labels as negative examples. We then estimate the proportions of positive and negative samples in each class and aggregate the results across all classes to obtain the overall noisy label rate. For MPEIA, KM, ROC, TiCE, DEDPUL, $(\text{TED})^n$, and SuDPL, we provide each client with a small clean validation set. For INCV, we assume prior knowledge of the type of label noise on each client.

5.4 Comparison of Baseline Methods for Noise Rate Estimation Error

We conduct comprehensive comparative experiments to evaluate the effectiveness of our proposed federated noise rate estimation method against nine state-of-the-art baseline approaches across four benchmark datasets encompassing diverse data modalities: computer vision tasks (CIFAR-10, CIFAR-100) and sensor-based activity recognition (Widar, WISDM-W), as shown in Tables 4 and 5. The experimental framework evaluates performance under three distinct noise rate distribution scenarios: Gaussian distributions with low variance $\mathcal{N}(0.2, 0.1)$, high variance $\mathcal{N}(0.5, 0.2)$, and Beta distributions $\text{Beta}(0.5, 0.5)$. Our evaluation methodology operates as follows: during the 10th to 15th communication rounds of federated training, we compute noise rate estimates for each participating client at each round, then average these five estimates to obtain the client's final noise rate estimate. Subsequently, we calculate the error between each client's estimated and ground-truth noise rates, with the reported metrics representing the Mean Absolute Error (MAE) and the standard deviation of the error values for all participating clients within each experimental configuration. This rigorous evaluation framework ensures a comprehensive assessment of both estimation accuracy and cross-client robustness under realistic federated learning conditions.

Table 4: Mean Absolute Error (MAE) and standard deviation of error for client-level noise rate estimation across all participating clients in federated learning environments on CIFAR-10 and CIFAR-100 datasets under three noise distribution configurations. Lower values indicate better performance.

Dataset	CIFAR10						CIFAR100					
Noise Distribution	Gaussian (0.2,0.1)		Gaussian (0.5,0.2)		Beta (0.5,0.5)		Gaussian (0.2,0.1)		Gaussian (0.5,0.2)		Beta (0.5,0.5)	
Method\Error (%)	MAE	Std	MAE	Std	MAE	Std	MAE	Std	MAE	Std	MAE	Std
ROC	10.90	2.30	13.50	2.76	16.10	3.21	16.60	3.12	18.05	3.42	21.32	3.52
TiCE	3.68	1.12	4.09	1.34	5.12	1.62	4.73	1.24	5.53	1.33	6.11	1.46
KM	3.55	1.24	4.17	1.56	4.56	1.71	4.61	1.33	5.14	1.41	5.56	1.52
INCV	4.60	1.12	4.56	1.23	4.92	1.45	6.29	2.31	6.58	2.74	6.44	2.65
IR	6.25	3.04	4.24	1.82	5.57	3.66	5.15	2.56	7.14	3.36	6.12	2.96
MPEIA	4.64	2.21	6.04	2.02	6.95	2.84	6.14	2.52	10.86	3.77	9.84	4.21
DEDPUL	5.11	2.31	7.85	3.26	8.11	2.98	6.17	2.49	8.96	3.05	7.64	2.65
$(TED)^n$	2.63	1.21	3.12	1.44	3.55	1.42	4.22	2.23	4.65	2.31	5.14	2.86
SuDPL	1.30	0.56	1.66	0.71	2.04	0.86	1.65	0.86	1.98	1.12	2.34	1.42
Ours	**0.82**	**0.48**	**0.97**	**0.65**	**1.32**	**0.98**	**1.46**	**0.96**	**1.51**	**0.97**	**1.65**	**1.01**

Bold values represent the best performance, highlighting the superiority of our proposed method.

Table 5: Mean Absolute Error (MAE) and standard deviation of error for client-level noise rate estimation across all participating clients in federated learning environments on Widar and WISDM-W datasets under three noise distribution configurations. Lower values indicate better performance.

Dataset	Widar						WISDM-W					
Noise Distribution	Gaussian (0.2,0.1)		Gaussian (0.5,0.2)		Beta (0.5,0.5)		Gaussian (0.2,0.1)		Gaussian (0.5,0.2)		Beta (0.5,0.5)	
Method\Error (%)	MAE	Std	MAE	Std	MAE	Std	MAE	Std	MAE	Std	MAE	Std
ROC	14.38	5.17	18.72	6.58	21.46	7.25	17.39	3.92	19.51	4.56	24.16	4.67
TiCE	5.29	2.08	6.12	2.46	6.75	2.35	6.14	1.68	7.73	2.42	8.25	3.12
KM	5.18	2.51	5.92	2.69	6.58	2.81	5.95	1.79	7.38	2.53	8.22	3.18
INCV	5.27	1.52	5.82	1.76	6.18	2.38	7.12	3.15	7.91	3.41	7.58	3.84
IR	6.61	2.89	4.93	2.35	6.57	2.97	5.78	2.52	7.03	3.47	6.97	3.25
MPEIA	6.15	3.18	7.52	3.55	8.94	4.35	7.46	3.17	11.68	4.73	12.38	4.61
DEDPUL	6.47	3.28	8.52	3.91	10.46	4.68	7.54	3.35	10.52	4.79	9.78	4.59
$(TED)^n$	4.62	1.74	5.36	2.09	6.23	3.47	5.76	1.93	6.51	2.87	7.18	3.36
SuDPL	2.68	1.28	3.12	1.38	3.42	1.54	3.06	1.61	3.73	1.57	3.88	1.91
Ours	**1.59**	**1.06**	**1.72**	**1.45**	**2.19**	**1.36**	**1.18**	**1.07**	**1.66**	**1.32**	**1.52**	**1.11**

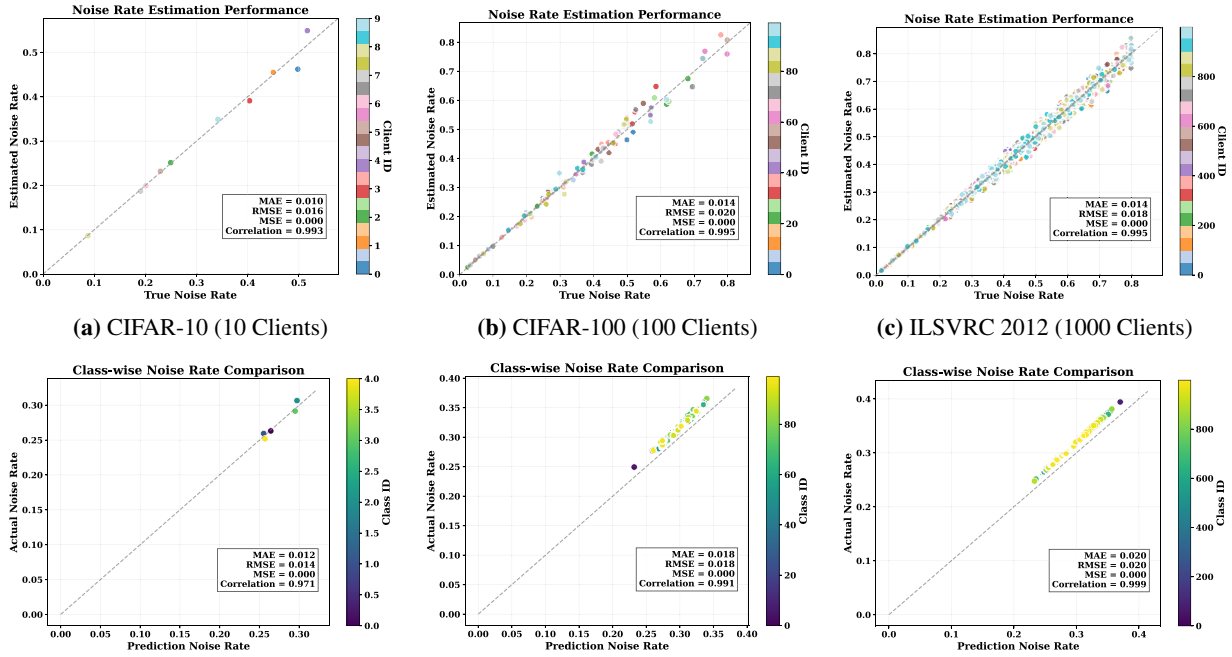
Bold values represent the best performance, highlighting the superiority of our proposed method.

The experimental results demonstrate substantial performance advantages for our proposed methodology across all evaluated configurations, as shown in Tables 4 and 5. On computer vision datasets (CIFAR-10 and CIFAR-100), our approach achieves MAE values ranging from 0.82% to 1.65%, representing significant improvements over the best-performing baseline method SuDPL, which records MAE values between 1.30% and 2.34%. The performance differential becomes particularly pronounced under high-noise scenarios, where our method maintains consistent accuracy while baseline methods exhibit substantial degradation. On sensor-based datasets (Widar and WISDM-W), our superiority becomes even more evident, achieving exceptional MAE values of 1.18%–2.19%, substantially outperforming all baseline approaches. Traditional methods such as ROC demonstrate particularly poor adaptation to sensor data characteristics, with MAE values exceeding 17% in multiple configurations. Beyond accuracy improvements, our method exhibits superior stability, as evidenced by consistently lower standard deviations ranging from 0.48 to 1.25 across all experimental configurations, substantially lower than those of most baseline approaches. This enhanced stability is

particularly critical in federated environments, where client heterogeneity can introduce performance variance. The consistent performance advantages across heterogeneous datasets—from low-resolution natural images to high-dimensional sensor signals—demonstrate the universal applicability and cross-domain generalization capability of our approach.

5.5 Noise Rate Estimation under Federated Configurations

The comprehensive experimental evaluation provides substantive evidence on the performance characteristics of the proposed noise rate estimation algorithm across diverse federated learning paradigms, as shown in Fig. 2. The systematic assessment comprises six distinct experimental configurations, designed to examine algorithmic scalability and estimation precision across varying client participation and categorical scales. Due to computational resource constraints, we were unable to fully train complete training on the large-scale classification dataset ILSVRC 2012. Therefore, we used the publicly available pre-trained ResNet50 model as the trained model to evaluate the noise rate. We employed non-independent and identically distributed (non-IID) data partitioning with a Dirichlet parameter of 0.4 to allocate samples from the ILSVRC 2012 dataset across all participating clients. The experimental configurations for CIFAR-10, CIFAR-10 subset, and CIFAR-100 datasets maintain consistency with the parameters specified in Table 2. In subfigures (a), (b), and (c), client noise rates follow a normal distribution with mean 0.5 and variance 0.2. In subfigures (d), (e), and (f), categorical noise rates follow a uniform distribution ranging from 0.2 to 0.6.



(a) CIFAR-10 (10 Clients) **(b)** CIFAR-100 (100 Clients) **(c)** ILSVRC 2012 (1000 Clients)

(d) CIFAR-10 Subset (5 Classes) **(e)** CIFAR-100 (100 Classes) **(f)** ILSVRC 2012 (1000 Classes)

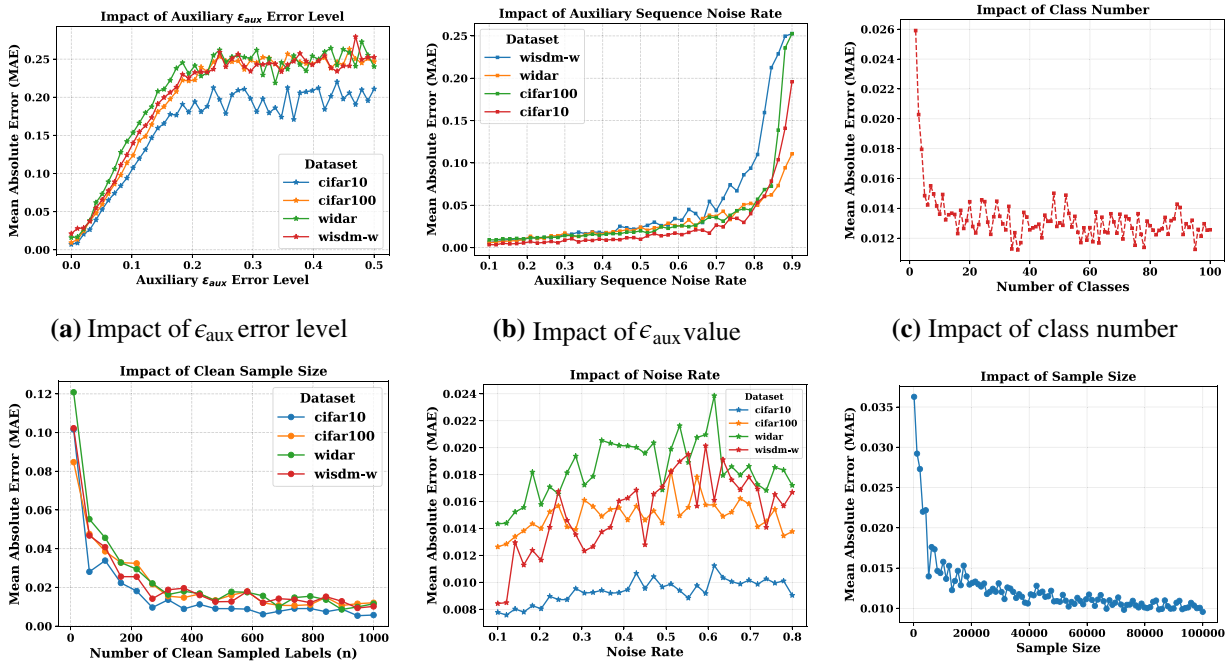
Figure 2: Performance evaluation of the proposed noise rate estimation algorithm across different federated learning scenarios. The top row demonstrates client-wise noise rate estimation performance under varying numbers of participating clients, while the bottom row illustrates class-wise estimation accuracy across different numbers of classes. All experiments show a strong correlation between true and estimated noise rates with consistently low estimation errors.

The empirical findings demonstrate notable consistency in estimation accuracy across different federated learning architectures. Specifically, estimation precision remains remarkably stable regardless of participant scale, with Mean Absolute Error (MAE) values ranging from 0.010 to 0.014 across configurations spanning 10 to 1000 clients. The observed correlation coefficients, consistently exceeding 0.993, suggest

strong linear relationships between true and estimated noise rates. This pattern persists across heterogeneous dataset characteristics, from small-scale CIFAR-10 deployments to large-scale classification scenarios comprising 1000 clients and 1.2 million samples. The experimental progression from 5-class to 1000-class scenarios indicates maintained estimation accuracy, with MAE values demonstrating controlled variance (0.010–0.020) despite significant increases in the number of categories. Collectively, this experimental evidence robustly indicates that our federated noise rate estimation method is applicable not only to small and medium-sized federated classification tasks but also to large-scale scenarios, thereby demonstrating substantial practical value.

5.6 Parameter Impact Analysis

This section provides a detailed analysis of the experimental results, elucidating the performance characteristics of the proposed federated noise rate evaluation algorithm under various conditions, as shown in Fig. 3. The Mean Absolute Error (MAE) serves as the primary metric for quantifying the algorithm's accuracy in noise rate estimation.



(a) Impact of ϵ_{aux} error level **(b)** Impact of ϵ_{aux} value **(c)** Impact of class number

(d) Impact of validation set size **(e)** Impact of overall noise level **(f)** Impact of training sample size

Figure 3: Comprehensive experimental evaluation of the proposed federated noise rate estimation algorithm across multiple parameter configurations. The experiments demonstrate the algorithm's performance sensitivity to various factors including the error level of evaluating ϵ_{aux}, the value of ϵ_{aux}, the number of categories, clean validation set size, noise rate, and training set size across four benchmark datasets (CIFAR-10, CIFAR-100, Widar, and WISDM-W).

Fig. 3a investigates the impact of estimation error in the noise rate of auxiliary label sequence $L_{\mathbf{aux}}$ (denoted as $|e - \hat{e}|$, where e is the true auxiliary noise rate and $\hat{e}$ is its estimated value) on the performance of the federated noise rate evaluation algorithm. The increase in $|e - \hat{e}|$ shows a positive correlation with Mean Absolute Error (MAE), highlighting the critical importance of accurately estimating the noise rate of auxiliary sequences. Fig. 3b demonstrates a clear declining trend in performance as the noise rate of auxiliary label sequences increases from 0.1 to 0.9. It is evident that our federated noise rate estimation algorithm achieves optimal performance at lower auxiliary label sequence noise levels (0.1–0.3), with performance

subsequently deteriorating as noise intensity increases. Fig. 3c presents the scalability characteristics of our federated noise rate estimation algorithm across different numbers of categories (2–100) on the CIFAR-100 dataset. The results demonstrate the algorithm's exceptional stability, with MAE values maintained within a narrow range (0.012–0.026). This stability reflects robust scalability, indicating that the algorithm maintains consistent performance regardless of classification complexity—a crucial factor for the practical deployment of federated learning across diverse task domains. Fig. 3d illustrates the effect of clean validation set size on our federated noise rate evaluation algorithm. It can be clearly observed that stable performance is achieved across multiple datasets once the clean validation set exceeds 300 samples. Fig. 3e shows the impact of different overall noise rate on our federated noise rate estimation algorithm in federated learning, where client noise rates follow a normal distribution with a mean and variance of 0.1. Although the Mean Absolute Error increases somewhat with rising overall noise rates, the MAE values remain within a narrow range (0.008–0.024), demonstrating good stability. Fig. 3f presents noise rate evaluation results using different numbers of training samples sampled from the LSVRC 2012 dataset. The results show that our federated noise rate evaluation algorithm performs better as the training data size increases.

5.7 Validating the Non-Uniform Label Noise for Simplifying ρ Computation

In this section, we experimentally validate that simplifying the computation of ρ is reasonable. We constructed two types of non-uniform noise—asymmetric noise and sample-dependent noise—on four datasets (CIFAR-10, CIFAR-100, Widar, and WISDM-W) under federated settings. The noise rate on each client follows a Gaussian distribution, with noise types being either asymmetric or sample-dependent. FNRE* represents noise rate estimation using Eq. (4) to compute ρ, while FNRE uses the simplified Eq. (5).

5.7.1 Real-World Datasets

The experimental results are shown in Tables 6 and 7. Across all datasets (CIFAR-10, CIFAR-100, Widar, WISDM-W), noise types (asymmetric and sample-dependent), and noise rate distributions, FNRE* and FNRE demonstrate consistently similar MAE and standard deviation values with minimal differences. This indicates that introducing the uniform label noise assumption for simplified computation of ρ does not significantly degrade performance. We further validated this on real-world noisy label datasets CIFAR-10N and CIFAR-100N, as shown in Table 7, where FNRE* and FNRE again exhibit comparable MAE and standard deviation values with consistently small differences.

Table 6: Performance comparison between FNRE* and FNRE on synthetic noise.

Dataset	Method	Asymmetric Noise				Sample-Dependent Noise			
		Gaussian (0.2,0.1)		Gaussian (0.3,0.1)		Gaussian (0.2,0.1)		Gaussian (0.3,0.1)	
		MAE	Std	MAE	Std	MAE	Std	MAE	Std
CIFAR-10	FNRE*	0.87	0.45	0.82	0.44	0.84	0.51	0.92	0.54
	FNRE	0.86	0.47	0.92	0.48	0.92	0.52	1.03	0.56
CIFAR-100	FNRE*	1.36	0.88	1.41	0.87	1.51	0.94	1.65	0.97
	FNRE	1.49	0.91	1.55	0.96	1.54	0.93	1.63	0.99
Widar	FNRE*	1.66	0.97	1.71	1.02	1.51	0.99	1.64	1.04
	FNRE	1.32	1.01	1.43	1.09	1.63	1.06	1.78	1.12

(Continued)

Table 6 (continued)

Dataset	Method	Asymmetric Noise				Sample-Dependent Noise			
		Gaussian (0.2,0.1)		Gaussian (0.3,0.1)		Gaussian (0.2,0.1)		Gaussian (0.3,0.1)	
		MAE	Std	MAE	Std	MAE	Std	MAE	Std
WISDM-W	FNRE*	1.36	0.91	1.42	0.92	1.28	0.93	1.41	0.97
	FNRE	1.44	0.89	1.59	0.96	1.39	0.94	1.52	1.02

Table 7: Performance comparison between FNRE* and FNRE on real-world noise.

Method	CIFAR-10N						CIFAR-100N					
	Aggregate		Random 1		Random 2		Worst		Coarse		Fine	
	9.03%		17.23%		18.12%		40.21%		25.60%		40.20%	
	MAE	Std	MAE	Std	MAE	Std	MAE	Std	MAE	Std	MAE	Std
FNRE*	0.88	0.45	0.92	0.44	0.89	0.41	0.98	0.48	1.59	0.87	1.82	0.94
FNRE	0.91	0.43	0.95	0.49	0.93	0.48	1.06	0.59	1.63	0.91	1.85	0.93

5.7.2 Synthetic Dataset

Additionally, we generated 500 synthetic classification datasets using sklearn's `make_classification` function with random parameters. The number of samples follows a uniform distribution from 2000 to 1,000,000, the number of classes from 2 to 1000, feature dimensions from 4 to 256, and class separation from 1 to 5. We computed the absolute error between FNRE* and FNRE estimates, with results shown in Fig. 4. The absolute errors remain within a very small range.

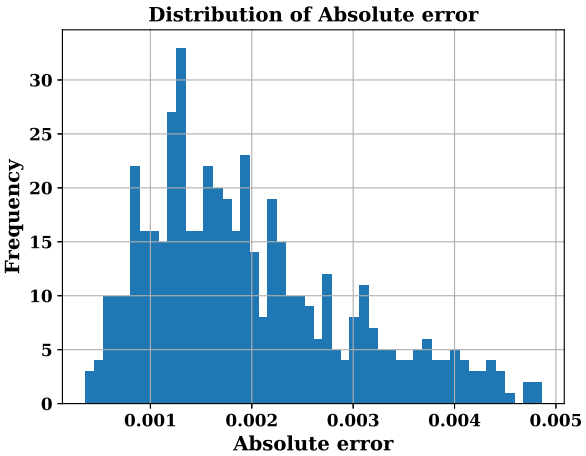


Figure 4: Distribution of absolute error between FNRE* and FNRE on 500 synthetic classification datasets.

Overall, our adoption of the uniform label noise assumption to simplify the computation of ρ does not substantially impact the accuracy of noise rate estimation.

6 Conclusion

We presented Federated Noise Rate Estimation (FNRE), a practical and privacy-preserving method for estimating client-specific label noise rates in federated learning. Unlike conventional centralized approaches, FNRE requires only a small clean validation set from a few clients and avoids per-client clean data or prior noise information. Our approach offers theoretical error guarantees and achieves superior accuracy across diverse datasets and noise conditions. These results demonstrate FNRE's effectiveness in enabling adaptive.

Limitations and Future Work

The proposed Federated Noise Rate Estimation (FNRE) method significantly improves label noise rate estimation in federated learning, but there are several limitations to address in future work:

Dependency on Clean Validation Data: Our approach relies on the prerequisite that at least one client possesses a small, clean validation set to guide the noise estimation.

Assumptions behind the Noise Model: We employ a uniform label noise assumption to simplify the configuration of the parameter ρ. While empirical results (Section 5.7) demonstrate that our method maintains robustness across most non-uniform noise settings, a theoretical analysis of Eq. (3) indicates potential performance degradation under specific conditions. Specifically, the method's effectiveness may diminish when the noise transition pattern is non-uniform and aligns closely with the model's prediction error pattern, particularly in regimes characterized by both high prediction errors and high noise rates.

Practical Considerations: From a practical deployment perspective, in cases where such complex noise-error correlations exist, the parameter ρ cannot be easily simplified and must be treated as a hyperparameter requiring extensive tuning.

In future work, we plan to reference anchor-based noise matrix estimation methods to construct a non-uniform correction term for the calculation of ρ. This extension would allow the model to adaptively handle complex asymmetric noise distributions.

Acknowledgement: Not applicable.

Funding Statement: This research was funded by the National Natural Science Foundation of China under Grant Numbers 62272180 (for Ling Yuan) and 62272176 (for Guohui Li).

Author Contributions: Qian Rong was responsible for the conception and design of the study and for drafting the paper. Lu Zhang contributed to the drafting of the paper. Ling Yuan was responsible for revising the manuscript critically for intellectual content and for the final approval of the version to be published. Zhong Yang contributed to the drafting of the paper. Guohui Li was involved in revising the manuscript critically for intellectual content and in the final approval of the version to be published. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: All datasets supporting the results or analyses of this study are publicly available in the following repositories:

- **CIFAR-10 and CIFAR-100 datasets:** Available at https://www.cs.toronto.edu/~kriz/cifar.html (No DOI or Accession Number).
- **CIFAR-10N and CIFAR-100N datasets:** Available at http://competition.noisylabels.com/ (No DOI or Accession Number).
- **ILSVRC 2012 datasets:** Available at https://image-net.org/challenges/LSVRC/2012/index.php (No DOI or Accession Number).
- **Widar datasets:** Available at https://tns.thss.tsinghua.edu.cn/widar3.0/index.html with DOI: https://doi.org/10.21227/7znf-qp86.
- **WISDM datasets:** Available at https://www.cis.fordham.edu/wisdm/dataset.php (No DOI or Accession Number).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Bhagoji AN, et al. Advances and open problems in federated learning. Found Trends Mach Learn. 2021;14(1–2):1–210.
2. Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. IEEE Signal Process Magaz. 2020;37(3):50–60.
3. Li G, Cai J, Lu J, Chen H. Incentive mechanism design for cross-device federated learning: a reinforcement auction approach. IEEE Trans Mobile Comput. 2025;24(4):3059–75. doi:10.1109/tmc.2024.3508260.
4. Alam S, Zhang T, Feng T, Shen H, Cao Z, Zhao D, et al. FedAIoT: a federated learning benchmark for artificial intelligence of things. arXiv:2310.00109. 2023.
5. Nguyen DC, Ding M, Pathirana PN, Seneviratne A, Li J, Poor HV. Federated learning for internet of things: a comprehensive survey. IEEE Commun Surv Tut. 2021;23(3):1622–58. doi:10.1109/comst.2021.3075439.
6. Li G, Cai J, He C, Zhang X, Chen H. Online incentive mechanism designs for asynchronous federated learning in edge computing. IEEE Internet Things J. 2024;11(5):7787–804. doi:10.1109/jiot.2023.3316470.
7. Wang H, Jiang T, Guo Y, Guo F, Bie R, Jia X. Label noise correction for federated learning: a secure, efficient and reliable realization. In: 2024 IEEE 40th International Conference on Data Engineering (ICDE). Piscataway, NJ, USA: IEEE; 2024. p. 3600–12.
8. Liang S, Huang J, Hong J, Zeng D, Zhou J, Xu Z. FedNoisy: federated noisy label learning benchmark. arXiv:2306.11650. 2023.
9. Xiao T, Xia T, Yang Y, Huang C, Wang X. Learning from massive noisy labeled data for image classification. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2015. p. 2691–9.
10. Li W, Wang L, Li W, Agustsson E, Van Gool L. Webvision database: visual learning and understanding from web data. arXiv:1708.02862. 2017.
11. Giap T-T, Kieu T-D, Le T-L, Tran T-H. FedDC: label noise correction with dynamic clients for federated learning. IEEE Internet Things J. 2025;12(8):10266–77.
12. Ji X, Zhu Z, Xi W, Gadyatskaya O, Song Z, Cai Y, et al. FedFixer: mitigating heterogeneous label noise in federated learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI Press; 2024. Vol. 38, p. 12830–8.
13. Ke S, Huang C, Liu X. Quantifying the impact of label noise on federated learning. arXiv:2211.07816. 2022.
14. Wu N, Yu L, Jiang X, Cheng K-T, Yan Z. FedNoRo: towards noise-robust federated learning by addressing class imbalance and label noise heterogeneity. arXiv:2305.05230. 2023.
15. Zeng B, Yang X, Chen Y, Shen Z, Yu H, Zhang Y. FedES: federated early-stopping for hindering memorizing heterogeneous label noise. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence; 2024 Aug 3–9; Jeju, Republic of Korea. p. 5416–24.
16. Chen L, Ang F, Chen Y, Wang W. Robust federated learning with noisy labeled data through loss function correction. IEEE Trans Netw Sci Eng. 2022;10(3):1501–11. doi:10.1109/tnse.2022.3227287.
17. Cheng A, Wang Z, Li Y, Cheng J. HPN: Personalized federated hyperparameter optimization. arXiv:2304.05195. 2023.
18. Jiang X, Wen T, Yang Z, Wu L, Chen Y, Sun S, et al. Robust federated learning against noisy clients via masked optimization. arXiv:2506.02079. 2025.
19. Lari E, Arablouei R, Gogineni VC, Werner S. Noise-robust and resource-efficient ADMM-based federated learning for WLS regression. Signal Process. 2025;241(10):110387. doi:10.2139/ssrn.5256354.
20. Scott C. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In: Artificial Intelligence and Statistics. London, UK: PMLR; 2015. p. 838–46. doi:10.3233/978-1-61499-672-9-1618.

21. Bekker J, Davis J. Estimating the class prior in positive and unlabeled data through decision tree induction. In: AAAI'18/IAAI'18/EAAI'18: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. Palo Alto, CA, USA: AAAI Press; 2018. Vol. 32, p. 2712–9.
22. Ramaswamy H, Scott C, Tewari A. Mixture proportion estimation via kernel embeddings of distributions. In: ICML'16: Proceedings of the 33rd International Conference on Machine Learning. London, UK: PMLR; 2016. p. 2052–60.
23. Chen P, Liao BB, Chen G, Zhang S. Understanding and utilizing deep neural networks trained with noisy labels. In: Proceedings of the 36 th International Conference on Machine Learning. London, UK: PMLR; 2019. p. 1062–70.
24. Liu T, Tao D. Classification with noisy labels by importance reweighting. IEEE Trans Pattern Anal Mach Intell. 2016;38(3):447–61. doi:10.1109/tpami.2015.2456899.
25. Yu X, Liu T, Gong M, Batmanghelich K, Tao D. An efficient and provable approach for mixture proportion estimation using linear independence assumption. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2018. p. 4480–9.
26. Ivanov D. Dedpul: difference-of-estimated-densities-based positive-unlabeled learning. In: 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA). Piscataway, NJ, USA: IEEE; 2020. p. 782–90.
27. Garg S, Wu Y, Smola AJ, Balakrishnan S, Lipton Z. Mixture proportion estimation and PU learning: a modern approach. Adv Neural Inf Process Systs. 2021;34:8532–44.
28. Zhu Y, Fjeldsted A, Holland D, Landon G, Lintereur A, Scott C. Mixture proportion estimation beyond irreducibility. In: ICML'23: Proceedings of the 40th International Conference on Machine Learning. London, UK: PMLR; 2023. p. 42962–82.
29. Xia X, Liu T, Wang N, Han B, Gong C, Niu G, et al. Are anchor points really indispensable in label-noise learning? In: Advances in neural information processing systems. Red Hook, NY, USA: Curran Associates, Inc.; 2019.
30. Yao Y, Liu T, Han B, Gong M, Deng J, Niu G, et al. Dual t: reducing estimation error for transition matrix in label-noise learning. Adv Neural Infn Process Syst. 2020;33:7260–71.
31. Li X, Liu T, Han B, Niu G, Sugiyama M. Provably end-to-end label-noise learning without anchor points. In: Proceedings of the 38 th International Conference on Machine Learning. London, UK: PMLR; 2021. p. 6403–13.
32. Menon A, Van Rooyen B, Ong CS, Williamson B. Learning from corrupted binary labels via class-probability estimation. In: Proceedings of the 32nd International Conference on Machine Learning. London, UK: PMLR; 2015. p. 125–34.
33. Jaynes ET. Probability theory: the logic of science. Cambridge, UK: Cambridge University Press; 2003.
34. Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images [master's thesis]. Toronto, ON, Canada: University of Toronto; 2009.
35. Wei J, Zhu Z, Cheng H, Liu T, Niu G, Liu Y. Learning with noisy labels revisited: a study using real-world human annotations. arXiv:2110.12088. 2021.
36. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. Intl J Comput Vision. 2015;115(3):211–52. doi:10.1007/s11263-015-0816-y.
37. Zheng Y, Zhang Y, Qian K, Zhang G, Liu Y, Wu C, et al. Zero-effort cross-domain gesture recognition with Wi-Fi. In: Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services. New York, NY, USA: ACM; 2019. p. 313–25.
38. Kwapisz JR, Weiss GM, Moore SA. Activity recognition using cell phone accelerometers. ACM SigKDD Explor Newsletter. 2011;12(2):74–82.
39. Song H, Kim M, Park D, Shin Y, Lee JG. Learning from noisy labels with deep neural networks: a survey. IEEE Trans Neural Netw Learn Syst. 2022;34(11):8135–53. doi:10.1109/tnnls.2022.3152527.
40. Cheng H, Zhu Z, Li X, Gong Y, Sun X, Liu Y. Learning with instance-dependent label noise: a sample sieve approach. arXiv:2010.02347. 2020.