FedLID: Leveraging Dimensionality to Overcome Noisy Labels in Federated Learning

Yufei Kang, Member, IEEE, Chenghao Hu, Member, IEEE, Baochun Li, Fellow, IEEE

Abstract—Federated Learning (FL) facilitates multiple clients to train a shared model on their own data in a privacy-preserving manner. In real-world FL applications, clients may contain label noise due to human annotation errors. Existing methods often struggle in the presence of data heterogeneity, as they operate under the assumption that noisy data tend to produce higher losses. This assumption convolutes the distinction between hard samples and mislabeled samples, particularly in data heterogeneous environments. In this study, we comprehend the mislabeled data within the FL process by examining the dimensionality of their deep representation subspace. Our investigations reveal distinct trends in local intrinsic dimensionality between noisy and clean local samples. On the basis of this finding, we propose FedLID as a practical solution that refurbishes mislabeled data and regularizes noisy data simultaneously. Empirically, we demonstrated that FedLID surpasses current state-of-the-art methods in managing noisy labels in federated learning with data heterogeneity. Furthermore, experiments under symmetric label noise and on the large-scale CIFAR-100 dataset confirm that FedLID generalizes robustly across diverse noise types and dataset scales, showcasing the potential of subspace dimensionality in advancing data quality research.

Index Terms—Federated Learning, Noisy Labels, Local Intrinsic Dimensionality, Data Heterogeneity, Label Correction, Privacy-Preserving Machine Learning

I. Introduction

Over the past decade, machine learning has witnessed unprecedented advancements, driving transformative breakthroughs across diverse fields such as computer vision and natural language processing. This progress has been fueled by several key factors, including improvements in hardware capabilities, the development of more efficient optimization algorithms, and, most notably, the availability of large-scale datasets. The ability of deep learning models to learn complex patterns and generalize effectively is heavily dependent on the quantity and quality of training data. As a result, centralized data collection has traditionally been the dominant paradigm for training high-performance models.

However, the centralization of real-world data introduces significant challenges related to data ownership, security, and privacy. Collecting and aggregating sensitive user data on centralized servers raises concerns about regulatory compliance, potential security breaches, and user trust. Consequently, enabling machine learning models to learn from decentralized, heterogeneous data sources without direct aggregation has become increasingly important. With the rapid proliferation of edge devices generating vast amounts of data and advancements in their computational capabilities, edge computing has emerged as a promising alternative to traditional cloud-based learning. This paradigm shift has led to the rise of federated learning (FL), a decentralized approach to training machine learning models.

Federated Learning [1] has gained significant research attention in recent years due to its ability to enhance privacy while enabling collaborative learning. In FL, a central server orchestrates training across multiple distributed clients, allowing them to collaboratively learn a global model without exposing their local datasets. This decentralized learning framework preserves user privacy and mitigates risks associated with data centralization. Due to these advantages, FL has been widely adopted in real-world applications, including Google's next-word prediction in Gboard and the Melloddy drug discovery project [2], demonstrating its potential to drive innovation across various domains while upholding data privacy.

In practical federated learning systems, clients are responsible for annotating their local datasets, but this process is prone to errors due to various factors, including a lack of domain expertise, limited annotation effort, or inattentiveness. As a result, mislabeled samples inevitably arise, introducing label noise into the training process. Even a small fraction of such mislabeled data can severely degrade model performance, as deep neural networks have a strong tendency to memorize incorrect labels at different levels of representation [3]. This issue is further exacerbated in FL due to the heterogeneity of client data distributions, annotation quality, and computational resources. Therefore, it is crucial to develop a robust learning method that can effectively mitigate the impact of label noise, ensuring reliable model performance in real-world FL deployments across diverse client environments.

A key challenge in mitigating the impact of noisy data in federated learning is accurately identifying mislabeled samples across different clients. Unlike centralized training, where a large and diverse dataset is available for robust noise detection, FL operates under decentralized conditions, where each client has limited data that is often non-independent and non-identically distributed (non-IID). This non-IID nature introduces significant variations

Yufei Kang, Chenghao Hu and Baochun Li are with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada, M5S 3G4.
E-mail: yufei.kang@mail.utoronto.ca, ch.hu@mail.utoronto.ca, bli@ece.toronto.edu.

Manuscript received July 21, 2025; revised May 4, 2026; accepted June 20, 2026

in data distributions across clients, making it difficult to distinguish mislabeled samples from genuine variations in local data patterns. As a result, noise mitigation techniques developed for centralized settings often perform suboptimally in FL, as they fail to account for the heterogeneity and limited sample sizes of individual clients [4]. Overcoming these challenges requires tailored strategies that can adapt to the decentralized and heterogeneous nature of FL while effectively suppressing the impact of mislabeled data.

To address the challenge of noisy labels in FL, researchers have proposed various solutions [5]–[10]. While these methods have demonstrated effectiveness in IID settings, they often struggle in non-IID scenarios due to their reliance on multiple pre-assumptions. A common assumption is that the server holds a clean auxiliary dataset to aid noise detection. However, this contradicts the core privacy principles of federated learning, where the server has no prior knowledge of client data before training begins. Another widely used assumption is the "small-loss" criterion, which suggests that mislabeled samples tend to exhibit higher loss values and that updates from noisy clients behave as statistical outliers. Many existing methods attempt to mitigate label noise by leveraging an auxiliary dataset or employing a warm-up stage to profile the global distribution. However, in heterogeneous FL settings, these assumptions often break down—a higher loss value does not necessarily indicate label noise, as it could instead correspond to an underrepresented class or a genuinely hard-to-learn instance. Consequently, these methods struggle to distinguish between true noise and natural variations in non-IID data, exposing a critical gap in existing research.

In this study, we investigate the dimensionality of deep representation subspaces in training data to gain a deeper understanding of the challenges associated with learning from noisy labels in federated learning. Specifically, we analyze how the intrinsic dimensionality of learned feature representations evolves throughout the FL process and leverage this insight to distinguish between clean and mislabeled samples. Our observations reveal a distinct pattern in the evolution of Local Intrinsic Dimensionality (LID) values for noisy and clean data: mislabeled samples consistently exhibit higher LID scores and undergo a more pronounced rebound compared to their clean counterparts. This behavior suggests that noisy samples occupy a more complex and less stable representation subspace, making them identifiable through dimensionality analysis.

Moreover, unlike conventional loss-based approaches, which often struggle in heterogeneous FL settings due to varying data distributions across clients, LID remains robust and reliable even in the presence of non-IID data. The stability of LID across different local distributions enables it to serve as an effective criterion for identifying mislabeled samples without relying on assumptions that may not hold in federated environments. By leveraging these properties, we introduce a noise-resilient approach that improves the reliability of federated learning in real-world, data-heterogeneous scenarios.

Motivated by these observations, we introduce FedLID as a novel solution to address the noisy data challenge in federated learning. In FedLID, we first associate the credibility of local samples with their Local Intrinsic Dimensionality value. This allows us to classify samples based on their reliability: for samples with high-confidence mislabeling, we correct their labels using predictions from the global model; for suspicious samples, we assign pseudo-labels and regularize their loss with a credibility term to prevent overfitting to potential noise. Importantly, all actions in FedLID are performed locally on the client side, without the need to transmit any additional data to the server, thereby maintaining the local privacy that is fundamental to the principles of FL.

To evaluate the effectiveness of our approach, we conducted comprehensive experiments that demonstrate FedLID's ability to accelerate the convergence of federated learning models in the presence of noisy labels. Our results also show that FedLID is robust against varying levels of label noise and different degrees of local data heterogeneity. By integrating dimensionality-based noise detection with a privacy-preserving mechanism, FedLID offers a promising solution to the persistent challenge of noisy data in federated learning, ensuring both model accuracy and privacy.

Our main contributions are summarized as follows: First, we explore the dimensionality of deep representation subspaces of training data in federated learning and reveal a distinct transformation pattern on noisy data compared to clean data, reflected by the LID value. Additionally, we design an LID-based evaluation method to quantify sample-wise credibility. Furthermore, to alleviate the negative impact of samples with false annotations, we propose the FedLID method, which refurbishes mislabeled data and regularizes noisy data jointly. Empirically, we showed the effectiveness and robustness of our proposed FedLID method in real-world settings with varying noisy and heterogeneous data.

This paper is organized as follows: Section II provides the necessary background, including the problem formulation and key preliminaries. Section III investigates the behavior of subspace dimensionality under noisy labels in federated learning. Section IV details the proposed FedLID framework. Section V presents comprehensive experimental results and analysis. Finally, Section VII concludes the paper.

II. Preliminaries

A. Federated Learning

In synchronous federated learning, a cloud server coordinates multiple clients to iteratively train a global model. In each communication round, each client optimizes a local model based on its own data, which may be non-IID (non-independent and identically distributed) and potentially include noisy labels. Once the client completes local optimization, it sends the model updates to the

TABLE I: Summary of Key Notations

Notation	Description
N, $\mathcal{N}$	Total number and set of clients
i, n_i	Client index; number of local samples
ω, ω^t	Global model; global model at round t
$\omega_i^{t,E}$	Local model of client i after E epochs
p_i	Aggregation weight of client i
K, $\mathcal{K}$	Number and set of selected clients
r, $F(r)$	Distance variable; its CDF
LID_F	LID of sample x
k	Nearest neighbor count for LID
$r_i(x)$	i-th neighbor distance
$\widehat{\mathrm{LID}}(x)$	Estimated LID of x
$g = h^{L-1}$	Penultimate layer output
y, $\widehat{y}$	Raw label; model prediction
$\widehat{\mathrm{LID}}_x$	Batch-estimated LID of x
α_x	Credibility score, $\alpha_x \in [0, 1]$
$\mathcal{L}_{CE}$, $\mathcal{L}(\omega)_x$	Cross-entropy; refined training loss
ρ, M	Noise rate; total number of classes
k_n	Top-k_n classes for label corruption

central server. The server waits for updates from several clients, aggregates them, and updates the global model. This iterative process continues until the global model converges to an optimal solution.

For a multi-class classification task, we consider a federated learning system consisting of a set of $\mathcal{N} = 1, \ldots, N$ clients, all coordinated by a central server. Each client i possesses a local dataset consisting of n_i training samples, $\{(x_{i,k}, y_{i,k})\}_{k=1}^{n_i}$, where $x_{i,k}$ represents the input data point and $y_{i,k}$ denotes the corresponding annotated label. These local datasets are used to train individual models on each client.

The learning process relies on a loss function $f(\cdot)$, which measures the performance of a model parameterized by ω on a given sample. Specifically, $f(\omega; x_{i,j})$ quantifies how well the model parameter ω performs on the j-th sample $x_{i,j}$ with label $y_{i,j}$ from client i. Cross-entropy loss is commonly used in classification tasks. The aggregation weight p_i for each client i reflects the relative importance of that client's local dataset in the global model update and is constrained such that $\sum_{i=1}^{N} p_i = 1$. The global objective of federated learning is to solve the following optimization problem:

$$\min_{\omega} F(\omega) := \sum_{i=1}^{N} p_i F_i(\omega) = \sum_{i=1}^{N} \frac{p_i}{n_i} \sum_{j=1}^{n_i} f(\omega; x_{i,j}), \quad (1)$$

where $F_i(\omega)$ is the local loss function for client i.

Following the Federated Averaging (FedAvg) protocol, at each t-th communication round, the server samples a subset of K clients ($K = |\mathcal{K}|$, with $\mathcal{K} \subseteq \mathcal{N}$) and broadcasts the current global model ω^t to the selected clients. Each client i initializes its local model $\omega_i^{t,0} = \omega^t$ and performs E steps of local optimization to compute the updated local model $\omega_i^{t,E}$. Once all selected clients have completed their local updates, they send their model updates back to the central server. Upon receiving the updates, the server aggregates them using the corresponding weights p_i and computes the new global model as:

$$\omega^{t+1} = \sum_{i \in \mathcal{K}} p_i \omega_i^{t,E}. \quad (2)$$

This process is repeated for multiple communication rounds, enabling the global model to improve progressively.

B. Local Intrinsic Dimensionality

Local Intrinsic Dimensionality (LID) is an expansion-based measure used to estimate the intrinsic dimensionality of the underlying data subspace, which is crucial for understanding the structure of high-dimensional data [11]. The concept of LID is grounded in the idea that the local geometry of data can be captured by how distances and volumes scale in the local neighborhood of a point. To illustrate, in Euclidean space, consider a D-dimensional unit ball. When this ball is scaled by a factor of r, its volume increases in proportion to r^D. This means that the volume of the ball grows at a rate that is determined by the intrinsic dimension D of the space.

This relationship between volume growth and the scaling factor provides a way to deduce the dimensionality of the data space. By observing how the volume changes as the radius is adjusted, we can estimate the intrinsic dimension of the data subspace in the vicinity of a given point. Specifically, if two volume measurements are taken at radii r_1 and r_2, yielding corresponding volumes V_1 and V_2, the ratio of these volumes can be expressed as:

$$\frac{V_2}{V_1} = \left(\frac{r_2}{r_1}\right)^D. \quad (3)$$

From this equation, the dimension D can be inferred by solving for it as follows:

$$D = \frac{\ln(V_2/V_1)}{\ln(r_2/r_1)}. \quad (4)$$

This measurement of intrinsic dimensionality would determine D by estimating the volumes in terms of the numbers of data points captured by the unit balls. Transferring the concept of expansion dimension from the Euclidean space to the statistical setting of continuous distance distributions, the notion of unit ball volume is replaced by the probability measure associated with the unit balls. The key underlying idea is that at each datapoint, the number of neighboring datapoints would grow with the radius of neighborhood, and the corresponding growth rate would then be a proxy for "local" dimension.

Definition 1 (Local Intrinsic Dimensionality). Given a data sample $x \in X$, let $r > 0$ be a random variable denoting the distance from x to other data samples. If the cumulative distribution function $F(r)$ is positive and continuously differentiable at distance $r > 0$, the LID of x at distance r is given by:

$$LID_F(r) := \lim_{\epsilon \to 0} \frac{\ln(F((1+\epsilon)r)/F(r))}{\ln(1+\epsilon)} = \frac{rF'(r)}{F(r)}, \quad (5)$$

whenever the limit exists. The LID at x is in turn defined as the limit of the radius $r \to 0$:

$$LID_F = \lim_{r \to 0} LID_F(r). \tag{6}$$

LID_F describes the relative rate at which its cumulative distance function $F(r)$ increases as the distance r increases. In the ideal case where the data in the vicinity of x are distributed uniformly within a local submanifold, LID_F equals the dimension of the submanifold. Nevertheless, in more general cases, LID also provides a rough indication of the dimension of the submanifold containing the sample x that would best fit the data distribution in the vicinity of x. Intuitively, the LID at x is an approximation of the dimension of a smooth manifold containing x that would "best" fit the distribution D in the vincinity of x.

C. LID Estimation

Given a reference sample point $x \sim P$, where P represents a global data distribution, P induces a distribution of distances relative to x, meaning each sample $x_* \sim P$ has an associated distance value $d(x, x_*)$. When considering a dataset X drawn from P, the smallest k nearest neighbor distances from x can be interpreted as extreme events from the lower tail of the induced distance distribution. According to extreme value theory, the tails of continuous distance distributions tend to converge to the Generalized Pareto Distribution (GPD). This theoretical framework allows us to estimate the Local Intrinsic Dimensionality value of the data around x.

Using maximum likelihood estimation, the LID can be estimated as:

$$\widehat{LID}(x) = -\left(\frac{1}{k} \sum_{i=1}^{k} \log \frac{r_i(x)}{r_{\max}(x)} \right)^{-1}, \tag{7}$$

where $r_i(x)$ denotes the distance between x and its i-th nearest neighbor, and $r_{\max}(x)$ represents the maximum distance among the nearest neighbors. This formulation provides an estimate of the intrinsic dimensionality of the local neighborhood around x by leveraging the scaling behavior of the nearest neighbor distances, making it a useful tool for understanding the local data structure.

D. Noise Model

In this work, we develop a novel noise injection framework designed to more closely mirror the nature of label noise observed in real-world scenarios, particularly in decentralized environments like federated learning. Unlike prior studies that often rely on simplistic noise injection techniques—such as random label flipping across all classes—we argue that such methods lack realism and fail to reflect the semantic structure of natural data. To address this gap, our noise model introduces an instance-dependent mechanism where mislabeling is based on class confusion informed by the predictions of moderately accurate neural networks.

In our framework, we first define a noise rate parameter $\rho \in [0, 1]$, which specifies the fraction of mislabeled data within each client's local dataset. For each client, we uniformly sample a percentage of ρ data points from the local training set to be corrupted. To determine plausible incorrect labels for these selected samples, we train a baseline deep neural network on the clean version of the dataset and use its predictions to identify which classes are most likely to be confused with the true label. For example, we use LeNet-5 with approximately 80% accuracy for MNIST, and VGG-16 with around 60% accuracy for more complex datasets such as CIFAR-10, CIFAR-10N, and CINIC-10. These moderately performing models reflect realistic classification scenarios where mistakes are not arbitrary but tend to cluster around visually or semantically similar classes.

For each sample selected for corruption, we extract the softmax output from the trained reference model and determine its top-k_n most likely, but incorrect, predicted classes. The incorrect label is then randomly selected from this top-k_n subset. This top-k_n strategy allows us to simulate structured, instance-dependent noise where mislabels are biased toward similar or confusable categories, which is more aligned with real-world annotation errors. The parameter k_n controls the severity and diversity of the mislabeling: lower values of k_n simulate more plausible, near-class confusions, while setting $k_n = M - 1$ (where M is the total number of classes) effectively reduces the model to an instance-independent noise scheme, as the new label is then randomly chosen from all remaining classes.

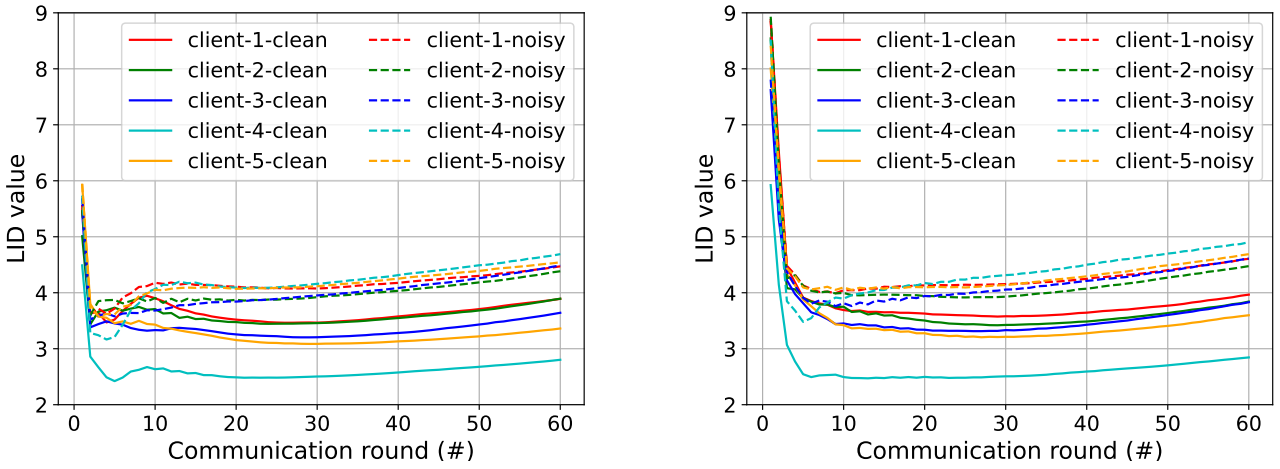
This instance-dependent noise model provides a flexible and principled framework for studying the impact of noisy labels in federated settings. It enables us to benchmark learning algorithms like FedLID under more realistic and challenging noise conditions while preserving a high degree of control over the noise characteristics. By grounding the mislabeling process in actual confusion patterns produced by neural networks, our approach ensures that noisy labels are not only statistically meaningful but also semantically plausible—thereby making the evaluation more robust and practically relevant.

III. Subspace Dimensionality in Federated Learning with Noisy Data

Our objective is to delve into the phenomenon of FL overfitting mislabeled training data and to explore viable solutions. We begin with an example to demonstrate how mislabeling influences the manifold transforms during learning, as signified by the LID value. We then analyze this effect and suggest that LID is a reliable and effective indicator for identifying noisy samples in federated learning, particularly in the context of data heterogeneity.

A. LID Estimation Within Batches

Since computing neighborhoods with respect to the entire dataset X is expensive, we estimate the LID value of a training example x from its k-nearest neighbor set



(a) IID (b) Non-IID

Fig. 1: Comparison of Subspace Dimensionality (LID scores) between clean and mislabeled samples, evaluated using a LeNet-5 model on both IID (left) and non-IID (right) MNIST datasets. Across all clients, the LID scores for noisy samples consistently exceed those of clean samples, with a noticeable rebound trend observed in noisy samples after reaching their lowest point. In contrast, the LID scores for clean samples remain relatively stable. These trends are evident in both the IID and non-IID settings, highlighting the robustness of LID as an indicator for detecting mislabeled data across different data distributions.

within a local batch randomly selected from X during local training [12]. Consider a L-layer neural network $h : P \to \mathbb{R}_c$, where h^i is the intermediate transformation of the i-th layer, and c is a positive number indicating the number of classes. Given a batch of local samples $X_B \subseteq X$, and a sample x, we estimate the LID score of x as follows:

$$\widehat{LID}(x) = -\left(\frac{1}{k}\sum_{i=1}^{k}\log\frac{r_i(g(x), g(X_B))}{r_{max}(g(x), g(X_B))}\right)^{-1}, \quad (8)$$

where $g = h^{L-1}$ is the output of the second-to-last layer of the network, $r_i(g(x), g(X_B))$ is the distance of $g(x)$ to its i-th nearest neighbor in the transformed set $g(X_B)$, and r_{max} represents the radius of the neighborhood. $\widehat{LID}(x, X_B)$ reveals the dimensional complexity of the local subspace in the vicinity of x, taken after transformation by g.

With a constant batch size m and a typical dimensionality d for gradients of a layer, the primary computational burden lies in the distance calculations. Given these constraints, the distance calculation involves $O(md)$ operations, and sorting or retrieving specific distances has a complexity of $O(m \log m)$. These operations are efficiently manageable within our server and empirical tests further substantiate our computational analysis. The use of vectorized operations and optimized data structures further reduces overhead, enabling efficient distance calculations and minimizing latency. Thus, the LID estimation remains computationally feasible, making it well-suited for real-world scenarios. In this study, we employ a local batch size of 128 and set k to 40, ensuring that the estimation within the batch closely approximates that within the full dataset, while minimizing computational costs.

B. Subspace Dimensionality and Label Quality

We now demonstrate, using the example below, how changes in the dimensionality of a training example's subspace are a crucial indicator of label quality in the training data as federated learning progresses. In our case, we trained a LeNet-5 model on the MNIST dataset. Specifically, we set up a federated learning scenario with five clients, all of whom are selected at each round. Furthermore, we constructed a noise setting where 40% of the training samples are wrongly labeled using the top-k_n ($k_n = 5$) most similar but incorrect classes, following the noise model described in Section II-D.

In addition, to create a more realistic federated learning scenario, we ran experiments on both an IID and a non-IID setting. In the non-IID setting, the local data are sampled following Dirichlet distribution with the concentration parameter of 0.5, representing a mildly non-IID class distribution. For clarity, we averaged the LID values at the representation layers separately for correctly labeled samples and mislabeled samples for each client.

The resulting LID scores on IID and non-IID settings are shown in Fig. 1. The LID value calculated on the same client are in the same color where the solid line represents the clean samples and the dashed line represents the noisy ones. On the clean samples, we observe that the LID scores decrease initially and keep stable at a very low level as learning proceeds. However, for the noisy samples, we observe an initial decrease in LID score but a rebound following up. Quantitatively, the average LID scores of the

noisy samples are always higher than that of the clean ones.

Significantly, the data presented in Fig. 1b from a non-IID setting show that while the absolute values of LID scores slightly increase, the trends in LID changes for both clean and noisy data remain consistent. This pivotal observation—that mislabeling influences manifold transformations—emphasizes the essential role of Local Intrinsic Dimensionality evaluation. LID emerges as a vital tool for pinpointing mislabeled samples, particularly given its robust performance in heterogeneous federated learning scenarios.

C. Potential Explanation

One way to understand this phenomenon is by looking at how the manifold changes as federated learning moves forward. When we have a training sample $x \in X$, its starting position is linked to a lower-dimensional local subspace defined by the underlying manifold A. Notably, LID exposes the growth patterns of the distance distribution from x, which is influenced by the dimension of the manifold that best associates with x.

As the learning process progresses, the manifold transforms, constantly improving its fit to the training data. If x is labeled correctly and many of its neighbors also have accurate labels, we can expect the learning process to converge towards a local subspace with a relatively low intrinsic dimensionality. However, it should be noted that the learning process still risks overfitting to the clean data if we carry on with the training for too long. Overfitting could lead to an increase in the dimensionality of the local manifold eventually, which aligns with Fig. 1, where the LID values on clean samples slightly increase towards the end.

If x is mislabeled, it will move to a different subspace, A', as learning progresses. This subspace is associated with other points that share the incorrect label applied to x. During this shift, the neighborhood of the transformed x, or x', tends to include more points from A' that share x's label, and fewer points from the original neighborhood in A. In relation to the points of A', the incorrectly labeled x' is an outlier spatially, as its coordinates are linked to A, not A'. Therefore, the presence of x' compels the local subspace around it to increase its dimensionality to accommodate it.

This also explains why the LID value remains steady against non-IID distributions, while existing methods based on distance or loss struggle. In such distributions, the original distribution of x is uneven across classes. This leads to a final manifold that is uneven in terms of class or samples. This irregularity can arise from both the original uneven distribution and incorrect labeling. However, solely measuring these distances cannot differentiate between these two sources of irregularity. Existing measurements based on distance excel in IID settings but falter in non-IID ones for this reason. Conversely, LID reflects the growth characteristics of the distance distribution, making it an effective and robust identifier in non-IID federated settings.

IV. FedLID: A Robust Federated Learning Framework with Noisy Labels

In the previous section, we showed that mislabeling distorts manifold structures and that Local Intrinsic Dimensionality value serves as a reliable signal for detecting noisy samples. Building on this insight, we propose FedLID, a novel framework for mitigating the effects of noisy labels in federated learning. FedLID uses LID-based evaluations to estimate the credibility of local training samples, enabling clients to better manage mislabeled data. Detected noisy samples are corrected using predictions from the global model. To further enhance robustness, a credibility-aware regularization term is incorporated into the local loss function, which discourages overfitting and encourages the model to prioritize clean samples during training.

A. Credibility Estimation via LID Dynamics

A core challenge in learning with noisy labels is to effectively assess the likelihood of local samples being mislabeled. This problem becomes even more pronounced in federated learning, where the decentralized nature and non-IID data distribution exacerbate label noise. Existing approaches often rely on heuristics derived from local losses or gradient statistics to flag anomalous behavior. For instance, some studies consider samples with consistently high loss as suspicious or rely on the inconsistency of local client updates to infer noise. However, these indicators are sensitive to model initialization, optimization dynamics, and data heterogeneity, often failing to generalize across diverse FL settings.

In contrast, our method FedLID introduces a geometry-aware strategy by linking the reliability of local samples to their Local Intrinsic Dimensionality value. LID measures the intrinsic dimensional complexity around a sample in the representation space. Intuitively, clean samples tend to reside in low-dimensional manifolds that capture the underlying data distribution, while noisy samples often lie in more chaotic, high-dimensional regions.

Consider a client in federated learning holding a local dataset X and a local deep model $h(\cdot)$ with L layers. Let $g = h^{L-1}$ denote the function mapping input samples to the output of the penultimate layer—often referred to as the representation layer. For example, in the case of the LeNet-5 model, the feature representations are extracted from the output of the fourth ReLU layer (referred to as relu4), which precedes the first fully connected layer. Similarly, for the VGG-16 model, the feature representations are extracted from the batch normalization layer immediately preceding the classifier head (referred to as bn). These representation layers capture high-level semantics, which are essential for computing LID.

Given a mini-batch $X_B \subseteq X$ and a sample $x \in X_B$, we estimate its LID value, denoted $\widehat{\mathrm{LID}}_x$, based on its neighborhood in the feature space using the method

described in ??. As demonstrated in Section III, noisy samples tend to exhibit higher absolute LID values and sharper rebounds in their LID trajectories, indicating unstable and complex local structures. In contrast, clean samples often converge to lower and more stable LID values. These observations suggest that the LID dynamics of a sample can serve as an indicator of its potential label quality.

To quantify this distinction, we define a credibility score α_x for each training sample x, which reflects the likelihood that the sample is correctly labeled. The first step in calculating this score involves tracking the ratio between the current LID value and the minimum LID value observed for x throughout training. Formally, we define the unnormalized credibility score as:

$$\alpha_x = -\ln\left(\text{clip}\left(\frac{\widehat{\text{LID}}_x}{\min \widehat{\text{LID}}_x}, -10^4, 10^4\right)\right), \tag{9}$$

where $\min \widehat{\text{LID}}_x$ is the minimum LID value observed for sample x up to the current round. The logarithm compresses extreme values and emphasizes relative increases, while the clipping bounds help mitigate the influence of outliers due to stochastic training noise or model instability.

Notably, because the scale of α_x may vary across clients due to differences in data distributions and model behaviors, we apply min-max normalization to calibrate the score across each client. This ensures the credibility scores are comparable and bounded within a consistent range. The normalized credibility score is computed as:

$$\alpha_x = \frac{\alpha_x - \alpha_{\min}}{\alpha_{\max} - \alpha_{\min}}, \tag{10}$$

where $\alpha_{\min}$ and $\alpha_{\max}$ are the minimum and maximum unnormalized scores over the batch or local dataset. After normalization, $\alpha_x \in [0, 1]$, with higher values indicating greater confidence that the sample is clean. These credibility scores are later integrated into the loss function to suppress the influence of suspected noisy labels during training, as detailed in the following subsections.

B. Label Correction via Credibility Scores

With the credibility score α_x, we gain valuable insights into the trustworthiness of local samples on individual clients. Instead of discarding potentially mislabeled data, as many existing approaches do, we propose a more constructive approach: correcting the mislabeled data using predictions from the global model. This strategy leverages the strengths of the federated architecture by integrating the global model's knowledge to improve the local dataset.

Let (x, y) represent a local training data pair on a local client, where x is the input sample and y is the raw label. In real-world scenarios, the client-annotated label y may not always be correct. Therefore, we refer to y as the raw label rather than the true label. During each round of federated learning, the client downloads the global model and generates a predicted label $\widehat{y}$ for the sample x. This predicted label serves as a potential correction to the raw label when necessary.

During local training, the client calculates the credibility score α_x for each sample x to assess the likelihood that the raw label y is correct. This calculation is performed using the method outlined in Eq. (9). A lower value of α_x suggests a higher likelihood of incorrect labeling. To identify mislabeled samples, we set a threshold on α_x. If α_x is smaller than the threshold, which can be defined as $\alpha_{min} < \alpha_{\text{avg}} - 3 \cdot \alpha_{\text{var}}$, where α_{avg} and α_{var} are the mean and variance of the credibility scores across a given batch, the sample is considered mislabeled. This threshold is easily established after a few initial training rounds and does not impose any significant computational overhead.

Once a sample is identified as potentially mislabeled, its raw label y is replaced by the federated model's predicted label $\widehat{y}$. Since the global model is the result of aggregation across multiple clients, its prediction provides a more reliable correction, especially in cases where individual client data may be noisy. This label correction process increases the number of clean samples, which, in turn, improves the quality of subsequent learning rounds.

By integrating this correction strategy, we take advantage of the global model's collective knowledge while avoiding the drawbacks of simply discarding potentially noisy data. This approach not only enhances the quality of training data but also accelerates the convergence of the federated learning process, leading to a more robust and accurate global model.

C. Regularization on Potential Noisy Samples

When training deep neural networks on datasets with noisy labels, a well-known phenomenon is the "early learning" phase, as described in [13], where the model initially fits the clean samples more accurately. During this phase, the model focuses on learning the clean data, but as training progresses, it starts to memorize the mislabeled samples. This occurs because, over time, the cross-entropy loss associated with clean samples diminishes as they are correctly classified, while the loss for noisy samples persists. Consequently, the optimization process shifts its focus to fitting the noisy samples, which can degrade the model's performance.

To mitigate the influence of noisy samples, we propose compensating the loss associated with clean (or potentially clean) samples by introducing an α_x term. This term modulates the contribution of each sample to the overall loss based on its credibility. A sample with a higher credibility score, indicated by a larger α_x, receives a greater weight in the loss, while a sample with a lower credibility score, indicating a higher likelihood of being noisy, contributes less.

This adjustment leads to a refined loss function that compensates the clean samples by increasing their contribution and decreasing the impact of potentially noisy samples. The adjusted training loss for each sample x is then expressed as:

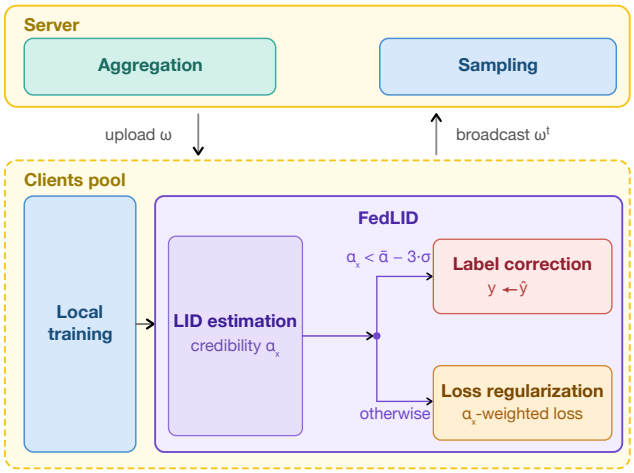


Fig. 2: Overview of the FedLID framework.

$$\mathcal{L}(\omega)_x := \mathcal{L}_{CE}(y, y^*) + \alpha_x, \tag{11}$$

where ω represents the local model weights, $\mathcal{L}_{CE}$ is the cross-entropy loss, and y^* is the predicted label. By incorporating the α_x term, the model places more emphasis on clean samples, thereby mitigating the negative effects of noisy labels and improving the robustness of the learning process.

D. Overview

FedLID is a flexible, plug-and-play module designed for seamless integration into existing federated learning frameworks, requiring minimal changes to the system architecture. Its adaptability allows it to work with various aggregation methods and client selection strategies, making it a versatile solution for a wide range of federated learning applications.

A key advantage of FedLID is that it operates entirely on the client side, ensuring that no raw data or additional metadata is transmitted to the central server. Clients only send model updates, maintaining the confidentiality of private client data throughout the federated learning process. This privacy-preserving approach enables collaborative learning while safeguarding the security of local datasets.

The core concept behind FedLID is to assess the reliability of local samples using their Local Intrinsic Dimensionality value and adjust their influence during training based on a computed credibility score. Clean samples are assigned higher weights, while noisy or mislabeled samples, indicated by lower credibility scores, have reduced influence. This mechanism prevents noisy data from dominating the training process, thereby improving model robustness. By operating entirely on the client side, FedLID preserves data privacy and integrates seamlessly into existing federated learning frameworks with minimal computational overhead, as detailed in Fig. 2.

V. Experimental Results

In this section, we empirically evaluate the performance of FedLID through a series of image classification tasks designed to reflect practical federated learning scenarios. We consider both IID and non-IID data distributions, and introduce varying levels of label noise to assess the FedLID's robustness to local data quality discrepancies. Our evaluation spans five benchmark datasets—CIFAR-10, CIFAR-10N, CINIC-10, CIFAR-100 and MNIST. Throughout all experimental settings, FedLID consistently demonstrates superior performance compared to baseline methods and recent noise-robust federated learning algorithms, validating its effectiveness as a reliable and generalizable solution.

A. Experimental Setup

All experiments are conducted using the open-source federated learning framework Plato, which provides a modular and extensible infrastructure for simulating various federated learning scenarios. Our experiments are executed on a server equipped with an NVIDIA GeForce RTX 4090 GPU, offering 24 GB of CUDA memory to support the training of deep convolutional neural networks with large batch sizes and high-resolution datasets.

a) Data Heterogeneity: To emulate the non-IID characteristics of real-world federated data, we partition each dataset across clients using a Dirichlet distribution with concentration parameter η. This factor controls the degree of heterogeneity in the local class distributions. Specifically, a lower η value, such as $\eta = 0.1$, results in highly skewed label distributions on each client, with most data belonging to only a few classes. In contrast, a higher η (e.g., $\eta = 10$) yields more balanced class distributions, approaching the IID setting. We set $\eta = 0.5$ by default in our main experiments to simulate realistic heterogeneity.

b) Data Noise: To assess the impact of noisy labels under heterogeneous data distributions, we introduce instance-dependent label noise into several clean datasets, including MNIST, CIFAR-10, CIFAR-100, and CINIC-10, following the procedure detailed in Section II-D. Specifically, for each client, 20% of local training samples are randomly selected for mislabeling. The incorrect labels are then drawn uniformly from the top-k_n ($k_n = 5$) most confusable classes, as determined by a reference model trained on clean data. This setup enables a controlled yet realistic simulation of local annotation errors that often arise in practical scenarios.

In addition to these synthetically corrupted datasets, we include the naturally noisy CIFAR-10N benchmark, which contains human-annotated labels with inherent variability. We evaluate performance under two representative noise conditions: the "aggregate" setting, with an estimated noise rate of 9.03%, and the "worst" setting, where the noise rate increases to 40.21% due to a higher degree of semantic confusion. These settings collectively provide a diverse and practical testbed for evaluating the robustness of federated noise learning methods.

Method	The Achieved Accuracy(%)			
	CIFAR-10	CIFAR-10N	CINIC-10	MNIST
FedAvg	65.2 ± 1.5	72.7 ± 1.3	52.3 ± 0.6	84.9 ± 1.2
FedProx	66.1 ± 0.6	$\mathbf{74.5 \pm 0.3}$	51.7 ± 0.7	84.7 ± 1.0
FedDiv	36.2 ± 15.6	13.3 ± 3.4	50.8 ± 1.8	79.2 ± 0.7
FedCorr	36.9 ± 10.1	43.0 ± 5.1	50.5 ± 0.7	75.0 ± 4.7
FedLID	$\mathbf{66.4 \pm 0.8}$	73.0 ± 0.8	$\mathbf{52.6 \pm 0.3}$	$\mathbf{86.8 \pm 1.1}$

TABLE II: Average and standard deviation of the final test accuracies achieved by various methods on CIFAR-10, CIFAR-10N, CINIC-10, and MNIST under mildly non-IID data distributions with injected label noise. Results in bold indicate the highest accuracy for each dataset. FedLID consistently outperforms all baselines on most cases.

Method	The Achieved Accuracy(%)			
	CIFAR-10 ($\rho = 0.4$)	CIFAR-10N ($worst$)	CINIC-10 ($\rho = 0.4$)	CIFAR-100 Symmetric Noise
FedAvg	56.8 ± 1.3	55.0 ± 2.3	52.6 ± 0.2	68.7 ± 0.9
FedProx	55.8 ± 0.3	54.5 ± 2.0	53.2 ± 0.8	67.5 ± 0.7
FedDiv	12.7 ± 4.6	24.9 ± 21.1	50.8 ± 0.9	47.3 ± 2.5
FedCorr	26.2 ± 5.1	31.9 ± 3.0	52.1 ± 0.2	30.9 ± 2.3
FedLID	$\mathbf{56.9 \pm 2.1}$	$\mathbf{56.1 \pm 0.8}$	$\mathbf{53.6 \pm 0.2}$	$\mathbf{69.2 \pm 0.6}$

TABLE III: Average and standard deviation of the converged test accuracies of various methods on CIFAR-10, CIFAR-10N, CINIC-10, CIFAR-100 at deteriorated noise levels and challenging noise type.

c) Datasets and Models: We evaluate our method on five image classification benchmarks. For synthetic noisy settings, we use MNIST, which consists of $60,000$ grayscale images of handwritten digits in 10 classes; CIFAR-10, which contains $60,000$ color images in 10 classes; CIFAR-100, which contains $60,000$ color images across 100 fine-grained classes; and CINIC-10, which extends CIFAR-10 to $270,000$ images by including additional samples from downsampled ImageNet, offering more visual diversity. For real-world noise, we use CIFAR-10N [14], where images are labeled by multiple annotators on Amazon Mechanical Turk, producing realistic noise.

We use LeNet-5 as the local model for MNIST due to its suitability for small grayscale inputs, and VGG-16 for CIFAR-10, CIFAR-100 CIFAR-10N, and CINIC-10, given its deep architecture and strong performance on natural images. This choice ensures consistency with prior work and allows us to assess scalability to deeper models.

d) Federated Configuration: We simulate a federated system with a total of 50 clients. In each communication round, 20 clients (40%) are randomly selected to train VGG-16 models, while 10 clients (20%) are selected to train LeNet-5 models. Clients perform local training on their non-overlapping partitions and send updates to the server for aggregation. We use the Adam optimizer with a learning rate of 1×10^{-3} and a batch size of 128. Each selected client performs 1 local epochs before synchronization. For LeNet-5, we train the model for 100 communication rounds, and for VGG-16, we train for 120 rounds to ensure convergence on more complex datasets.

e) Evaluation Protocol: To obtain reliable results, each experiment is repeated three times with different random seeds. We report the mean and standard deviation of the converged accuracies. For CIFAR-10N, all evaluations are performed using the clean labels in the original CIFAR-10 test set to ensure a fair assessment of the generalization performance under label noise.

f) Competing Strategies: We compare FedLID with two types of baseline methods. The first group includes standard federated learning algorithms that are designed to address statistical heterogeneity, such as FedAvg [1] and FedProx [15]. The second group consists of recent methods explicitly designed for handling noisy labels in federated learning, including FedCorr [4] and FedDiv [10], which are shown to outperform many prior approaches across a range of noise settings. All competing methods are implemented in the Plato framework using their official codebases and reported hyperparameter settings to ensure a fair and consistent comparison.

B. Main Results

To evaluate the effectiveness of FedLID, we conducted a federated learning environment with both data heterogeneity and noisy local labels. The local datasets were partitioned using a Dirichlet distribution with a concentration parameter of $\eta = 0.5$, simulating a mildly non-IID data distribution typical in real-world FL settings. For noise injection, we utilized the CIFAR-10N dataset with the "aggregation" mislabeling type, where labels are randomly altered in a way that mimics label noise in practice. Additionally, we artificially corrupted 20% of the training labels in CIFAR-10 and CINIC-10, and 40% in MNIST, creating a diverse range of noisy training conditions that reflect real-world challenges in federated learning scenarios.

Table II summarizes the converged test accuracies and standard deviations for all five evaluated methods. FedLID consistently achieves the highest performance across most datasets, demonstrating strong robustness against both label noise and data heterogeneity. Specifically, it attains 66.4% on CIFAR-10, 52.6% on CINIC-10, and 86.8% on MNIST, outperforming all other baselines under the same noisy, non-IID conditions. Although FedProx achieves the best accuracy on CIFAR-10N with 74.5%, FedLID closely follows with 73.0%, further demonstrating its reliability in realistic federated learning scenarios. These results highlight the effectiveness of FedLID's dimensionality-based design in mitigating the adverse impact of corrupted labels and heterogeneous data distributions.

In contrast, noise-robust FL methods such as FedCorr and FedDiv perform significantly worse, especially on the more challenging datasets. On CIFAR-10, FedDiv achieves a mere 36.2%, and FedCorr achieves 36.9%. On CIFAR-10N, these methods show even more dramatic performance degradation, with FedDiv reaching 13.3% and FedCorr

reaching 43.0%. These methods rely on loss-based heuristics to identify potentially mislabeled samples by detecting outliers in the loss distribution or its variants. However, under non-IID conditions, the substantial variance in loss values across clients renders these heuristics unreliable, as the local data distributions vary widely. As a result, these methods frequently misidentify clean samples as noisy ones and vice versa, leading to incorrect corrections that accumulate over time, resulting in poor performance.

C. Robustness against Noise Types and Levels

We evaluate FedLID's robustness under two complementary dimensions of noise variation: noise severity under our instance-dependent noise model, and noise type by including symmetric label noise—a classical benchmark widely used in the literature.

a) Instance-Dependent Noise at High Severity: To further assess the robustness of FedLID under more challenging noise conditions, we increased the label corruption severity across datasets to simulate realistic scenarios with significant noise. Specifically, we introduced 40% label noise into the CIFAR-10 and CINIC-10 datasets, representing substantial label corruption. For the CIFAR-10N dataset, we applied the "worst" noise setting, which causes a high degree of semantic confusion between classes, making it difficult for models to recover the correct labels.

As before, all experiments adopted a mildly non-IID local data distribution using a Dirichlet parameter of $\eta = 0.5$, reflecting typical heterogeneity in federated learning where each client holds a distinct data subset, resembling real-world scenarios.

The results, shown in Table III, indicate that while the overall accuracy of all methods declined due to the increased noise severity, FedLID consistently outperformed all other methods. Specifically, for CIFAR-10 with 40% label noise, FedLID achieved a test accuracy of 56.9%, outperforming FedAvg (56.8%), FedProx (55.8%), and other noise-robust methods such as FedCorr (26.2%) and FedDiv (12.7%). Similarly, on the CIFAR-10N dataset under the "worst" noise setting, FedLID achieved the highest accuracy of 56.1%, far exceeding FedAvg (55.0%), FedProx (54.5%), and the noise mitigation methods FedCorr (31.9%) and FedDiv (24.9%). For CINIC-10 with 40% noise, FedLID achieved 53.6%, again outperforming FedAvg (52.6%), FedProx (53.2%), FedCorr (52.1%), and FedDiv (50.8%). These results demonstrate FedLID's resilience, even when facing extreme levels of label noise.

FedLID consistently outperformed other methods across all datasets, maintaining high accuracy even under severe noise. Its advantage was most evident in challenging settings, highlighting the robustness of its dimensionality-based approach. In contrast, loss-based methods like FedCorr and FedDiv, which rely on heuristics to detect noisy labels via loss outliers, were unreliable under high noise and non-IID conditions. Variability in local loss distributions led to frequent misidentification of clean and noisy samples, resulting in poor convergence and significantly lower performance.

Method	The Achieved Accuracy(%) ($\eta = 0.1$)	($\eta = 0.3$)	($\eta = 1.0$)
FedAvg	36.8 ± 7.6	64.3 ± 1.8	65.6 ± 0.4
FedProx	39.2 ± 5.0	64.6 ± 3.5	66.0 ± 1.7
FedDiv	10.0 ± 0.0	19.3 ± 15.5	43.7 ± 12.8
FedCorr	10.1 ± 0.1	20.3 ± 3.2	45.6 ± 18.5
FedLID	$\mathbf{39.5 \pm 2.1}$	$\mathbf{65.2 \pm 0.6}$	$\mathbf{66.1 \pm 1.2}$

TABLE IV: Average and standard deviation of the converged test accuracies of various methods on CIFAR-10 with noise at different non-IID levels.

Method	The Achieved Accuracy(%) ($\eta = 0.1$)	($\eta = 0.3$)	($\eta = 1.0$)
FedAvg	$\mathbf{42.8 \pm 2.7}$	70.7 ± 1.8	74.1 ± 1.4
FedProx	42.1 ± 2.5	70.2 ± 2.4	74.3 ± 0.6
FedDiv	10 ± 0.0	12.0 ± 2.8	27.7 ± 14.2
FedCorr	14.2 ± 0.9	23.3 ± 2.3	55.6 ± 10.6
FedLID	39.0 ± 4.3	$\mathbf{72.3 \pm 2.1}$	$\mathbf{74.7 \pm 0.8}$

TABLE V: Average and standard deviation of the converged test accuracies of various methods on CIFAR-10N at different non-IID levels.

b) Symmetric Noise: Symmetric label noise, where each label is independently flipped to any other class with equal probability, is a classical noise type widely studied in the literature. Our noise model subsumes it as a special case: when $k_n = M - 1$, the incorrect label is drawn uniformly from all remaining classes, recovering precisely the symmetric noise setting. This makes symmetric noise a natural and principled benchmark within our framework.

The rightmost column of Table III confirms that FedLID consistently outperforms or matches the best competing method under symmetric noise. Notably, FedCorr and FedDiv exhibit substantially higher variance than FedLID, revealing their instability in this setting. FedLID's LID-based approach remains effective because the geometric inconsistency of mislabeled samples persists regardless of noise structure.

D. Robustness against Data Heterogeneity

We evaluate FedLID's robustness under varying degrees of data heterogeneity in this part. We vary the Dirichlet concentration parameter η across three representative settings: $\eta = 0.1$ (high heterogeneity), $\eta = 0.3$ (moderate), and $\eta = 1.0$ (near-IID). To isolate the effect of heterogeneity, label noise is held constant: we apply a 20% corruption rate to CIFAR-10, CINIC-10 and CIFAR-100, use the "aggregation" noise from CIFAR-10N.

The results in Table IV, Table V, Table VI, and Table VII show a consistent trend across all four datasets: test accuracy improves with lower data heterogeneity (i.e., higher η), peaking under the near-IID setting ($\eta = 1.0$). As heterogeneity increases, performance degrades, reflecting

Method	The Achieved Accuracy(%) ($\eta = 0.1$)	($\eta = 0.3$)	($\eta = 1.0$)
FedAvg	46.7 ± 1.5	51.0 ± 0.9	50.7 ± 0.2
FedProx	46.9 ± 3.0	50.4 ± 0.7	$\mathbf{51.0 \pm 0.6}$
FedDiv	26.8 ± 2.9	35.9 ± 8.7	49.3 ± 1.2
FedCorr	$\mathbf{47.2 \pm 3.2}$	50.3 ± 0.5	49.6 ± 0.5
FedLID	47.1 ± 1.7	$\mathbf{51.2 \pm 0.3}$	50.3 ± 1.3

TABLE VI: Average and standard deviation of the converged test accuracies of various methods on CINIC-10 with noise at different non-IID levels.

Method	The Achieved Accuracy(%) ($\eta = 0.1$)	($\eta = 0.3$)	($\eta = 1.0$)
FedAvg	39.9 ± 1.4	42.2 ± 1.0	44.7 ± 0.8
FedProx	42.6 ± 1.0	44.0 ± 0.7	$\mathbf{46.5 \pm 0.6}$
FedDiv	41.3 ± 1.7	42.7 ± 1.2	44.8 ± 0.9
FedCorr	$\mathbf{43.5 \pm 1.2}$	44.4 ± 0.8	46.4 ± 0.7
FedLID	43.2 ± 0.9	$\mathbf{45.2 \pm 0.6}$	46.1 ± 0.5

TABLE VII: Average and standard deviation of the converged test accuracies of various methods on CIFAR-100 with instance-dependent noise ($\rho = 0.2$) at different non-IID levels.

the reduced representativeness of each client's local data—an outcome consistent with real-world observations in federated learning.

Despite the challenges posed by simultaneous label noise and non-IID data, FedLID demonstrates consistently strong performance across all evaluated settings. On CIFAR-10, FedLID achieves the highest accuracy across all levels of data heterogeneity. For CIFAR-10N, FedLID leads in the moderate and near-IID settings with 72.3% and 74.7% accuracy, and ranks second under high heterogeneity. Similarly, on CINIC-10, FedLID performs best in the moderate setting (51.2%), closely following the leading method elsewhere. The stable performance across these three datasets highlights FedLID's adaptability to diverse federated environments where both label noise and non-IID data co-exist.

On CIFAR-100, FedLID achieves the highest accuracy at $\eta = 0.3$ (45.2%), outperforming all baselines, and remains competitive at other heterogeneity levels with consistently the lowest standard deviation. With 100 fine-grained classes, the increased inter-class semantic similarity makes label noise substantially harder to detect—a mislabeled sample is more likely to receive a visually similar incorrect label, resulting in subtler LID distortions. Despite this added difficulty, FedLID's geometry-aware approach remains effective: mislabeled samples continue to exhibit elevated and rebounding LID trajectories regardless of the number of classes, confirming that the key geometric signal is class-count agnostic. The consistently low variance across all η settings further validates FedLID's reliability and scalability to large-scale, fine-grained classification tasks.

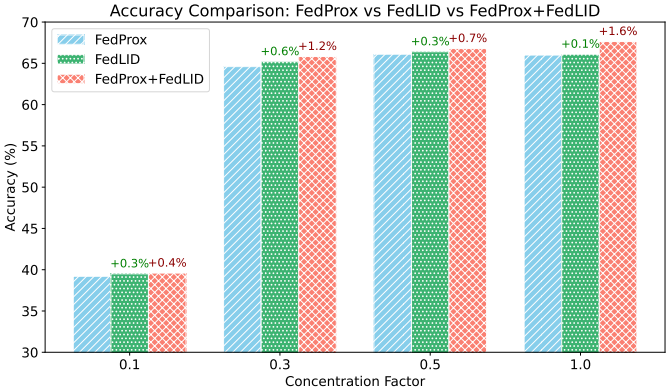


Fig. 3: Test accuracy of FedProx (blue), FedLID (green), and FedProx+FedLID (red) on CIFAR-10 with VGG-16 under varying data heterogeneity levels ($\eta \in 0.1, 0.3, 0.5, 1.0$). FedLID consistently improves performance when integrated into FedProx, demonstrating its plug-and-play compatibility.

E. FedLID As an Add-On Module

FedLID is designed as a flexible, plug-and-play module that can be seamlessly integrated into existing federated learning frameworks with minimal architectural modifications. Its modular nature allows for easy adoption across various aggregation schemes and client selection strategies, making it a broadly applicable solution for enhancing robustness in FL settings.

To demonstrate the compatibility and performance benefits of FedLID when used as an auxiliary component, we conducted experiments incorporating FedLID into the FedProx algorithm. Specifically, we trained a VGG-16 model on the CIFAR-10 dataset under different levels of data heterogeneity, characterized by Dirichlet distribution parameters $\eta \in \{0.1, 0.3, 0.5, 1.0\}$. We compared the final test accuracy of the combined method (FedProx+FedLID) against the standalone FedProx and FedLID methods.

As shown in Figure 3, the red bar corresponds to the combined method FedProx+FedLID, which consistently achieves superior performance compared to the baseline across various settings. This indicates that incorporating FedLID as an auxiliary module into FedProx significantly improves the algorithm's robustness to both label noise and data heterogeneity—two major challenges in federated learning. By leveraging representation-level characteristics to identify and mitigate the impact of noisy samples, FedLID complements the regularization mechanisms of FedProx, leading to more stable and reliable training dynamics.

F. Hyperparameter Analysis

FedLID has two key hyperparameters: the nearest neighbor count k for LID estimation, and the mislabeling threshold multiplier c in $\alpha_{\text{avg}} - c \cdot \alpha_{\text{var}}$.

a) Nearest Neighbor Count k: k governs a bias-variance trade-off in the MLE estimator [12]: larger k reduces variance but risks violating locality, while smaller k preserves local geometry at the cost of higher variance. k must also remain well below the batch size to avoid spanning cross-class neighborhoods. With batch size 128, we set $k = 40$ (31% of the batch), following Ma et al. [12], who validated this configuration across MNIST, CIFAR-10, and CIFAR-100. This choice is further motivated by the non-IID federated setting, where a compact neighborhood reduces cross-class contamination and preserves the geometric signal distinguishing mislabeled from clean samples.

b) Sensitivity to Mislabeling Threshold c: We set $c = 3$, following the standard 3σ convention for outlier detection under approximately unimodal distributions—a reasonable assumption for within-batch credibility scores derived from normalized LID ratios. Crucially, the threshold is fully adaptive: recomputed from batch statistics each round, it automatically tracks the evolving score distribution as training progresses and noise is corrected. This adaptivity reduces sensitivity to the specific value of c and enables generalization across clients with heterogeneous noise rates without per-client tuning.

VI. Related Work

Federated learning with noisy labels has garnered increasing attention and can be broadly categorized into three primary approaches based on differing design principles.

One straightforward approach involves reserving clean data on the server to identify noisy clients and mitigate label noise through techniques such as knowledge distillation (FedNed [16]), adaptive weighting (FOCUS [5]), and local regularization (RHFL [6]). However, in practical scenarios, access to such auxiliary public datasets is often limited or entirely unavailable for both the server and clients.

More advanced approaches aim to relax strong assumptions when estimating noise rates. One promising research direction involves multi-stage frameworks that incorporate a warm-up phase before addressing noisy label learning. For instance, FedNoRo [17] begins by identifying noisy clients using normalized local losses from a warm-up model trained with FedAvg, followed by distance-aware model aggregation in the second stage. IDNTM [18] introduces a three-stage pipeline: federated data extraction, transition matrix estimation—where clients collaboratively train an IDNTM estimation network on the extracted data—and a final federated classifier correction step. Similarly, FedCorr [4] identifies clean clients during the initial stage, uses only the clean ones for fine-tuning in the second stage, and involves all clients in the final training phase.

To further reduce complexity and computational costs, several methods have been proposed to integrate noise-robust mechanisms directly into the FedAvg framework. RoFL [19] employs similarity-based confident example sampling and global-guided pseudo-labeling. FedCNI [20] identifies noisy samples using prototypical similarity and assigns labels through curriculum pseudo-labeling with dynamic confidence. FedFixer [7] defines noisy samples as those whose cross-entropy loss significantly deviates from the mean prediction distribution. Comm-FedBio [21] approximates Shapley values to identify outliers as noisy samples. FedRN [9] leverages k-reliable neighbors—samples with high data expertise or similarity—to evaluate the credibility of local data. FedDiv [10] introduces a global noise filter based on a Gaussian Mixture Model (GMM) to detect noisy samples. LSR [8] applies a local self-regularizer based on self-distillation, while FedLNL [22] introduces a Local Diversity Product Regularization to address label noise.

Notably, FedCorr employs a cumulative local intrinsic dimensionality score to distinguish between clean and noisy clients, supplemented by a loss-based mechanism to identify noisy samples. Building upon this concept, our work extends the use of LID by examining the subspace transformations of noisy samples at a finer granularity. This enables a quantitative assessment of training data reliability and facilitates robust federated learning under data heterogeneity. The effectiveness of our approach is validated through comprehensive empirical evaluations.

VII. Conclusion

In this study, we investigate mislabeled data in federated learning by analyzing the dimensionality of deep representation subspaces. We demonstrate that label noise disrupts manifold transformations and identify local intrinsic dimensionality value as a reliable indicator of noisy samples. Building on this insight, we propose FedLID, a method designed to mitigate label noise in federated learning with heterogeneous data. Specifically, FedLID jointly refurbishes mislabeled samples and regularizes noisy data representations within a unified framework. Empirical results show that FedLID is both effective and robust across varying degrees of data heterogeneity and noise levels. Moreover, FedLID can further enhance the performance of existing noise-robust methods when used in combination. Beyond instance-dependent noise, additional experiments under symmetric label noise confirm that FedLID's LID-based approach generalizes across different noise structures, as the geometric inconsistency of mislabeled samples in the representation subspace persists regardless of how incorrect labels are assigned. Experiments on CIFAR-100 further validate the scalability of FedLID to large-scale, fine-grained classification settings with 100 classes. These findings underscore the promise of subspace dimensionality analysis for improving data quality in federated learning.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Artificial intelligence and statistics. PMLR, 2017, pp. 1273–1282.

[2] M. Oldenhof, G. Ács, B. Pejó, A. Schuffenhauer, N. Holway, N. Sturm, A. Dieckmann, O. Fortmeier, E. Boniface, C. Mayer et al., "Industry-scale orchestrated federated learning for drug discovery," in Proc. the AAAI Conference on Artificial Intelligence (AAAI), vol. 37, no. 13, 2023, pp. 15 576–15 584.
[3] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," Communications of the ACM, vol. 64, no. 3, pp. 107–115, 2021.
[4] J. Xu, Z. Chen, T. Q. Quek, and K. F. E. Chong, "FedCorr: Multi-stage federated learning for label noise correction," in Proc. the IEEE/CVF conference on computer vision and pattern recognition (CVPR), 2022, pp. 10 184–10 193.
[5] Y. Chen, X. Yang, X. Qin, H. Yu, P. Chan, and Z. Shen, "Dealing with label quality disparity in federated learning," Federated Learning: Privacy and Incentive, pp. 108–121, 2020.
[6] X. Fang and M. Ye, "Robust federated learning with noisy and heterogeneous clients," in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10 072–10 081.
[7] X. Ji, Z. Zhu, W. Xi, O. Gadyatskaya, Z. Song, Y. Cai, and Y. Liu, "FedFixer: Mitigating heterogeneous label noise in federated learning," in Proc. the AAAI Conference on Artificial Intelligence, vol. 38, no. 11, 2024, pp. 12 830–12 838.
[8] X. Jiang, S. Sun, Y. Wang, and M. Liu, "Towards federated learning against noisy labels via local self-regularization," in Proc. the 31st ACM International Conference on Information & Knowledge Management (CIKM), 2022, pp. 862–873.
[9] S. Kim, W. Shin, S. Jang, H. Song, and S.-Y. Yun, "FedRN: Exploiting k-reliable neighbors towards robust federated learning," in Proc. the 31st ACM International Conference on Information & Knowledge Management (CIKM), 2022, pp. 972–981.
[10] J. Li, G. Li, H. Cheng, Z. Liao, and Y. Yu, "FedDiv: Collaborative noise filtering for federated learning with noisy labels," in Proc. the AAAI Conference on Artificial Intelligence, vol. 38, no. 4, 2024, pp. 3118–3126.
[11] M. E. Houle, "Local intrinsic dimensionality I: An extreme-value-theoretic foundation for similarity applications," in Similarity Search and Applications (SIASP). Springer International Publishing, 2017, pp. 64–79.
[12] X. Ma, Y. Wang, M. E. Houle, S. Zhou, S. Erfani, S. Xia, S. Wijewickrema, and J. Bailey, "Dimensionality-driven learning with noisy labels," in International Conference on Machine Learning. PMLR, 2018, pp. 3355–3364.
[13] S. Liu, J. Niles-Weed, N. Razavian, and C. Fernandez-Granda, "Early-learning regularization prevents memorization of noisy labels," Proc. Advances in neural information processing systems (NeurIPS), vol. 33, pp. 20 331–20 342, 2020.
[14] J. Wei, Z. Zhu, H. Cheng, T. Liu, G. Niu, and Y. Liu, "Learning with noisy labels revisited: A study using real-world human annotations," in Proc. International Conference on Learning Representations (ICLR), 2022. [Online]. Available: https://openreview.net/forum?id=TBWA6PLJZQm
[15] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," Proc. Machine learning and systems (MLSys), vol. 2, pp. 429–450, 2020.
[16] Y. Lu, L. Chen, Y. Zhang, Y. Zhang, B. Han, Y.-m. Cheung, and H. Wang, "Federated learning with extremely noisy clients via negative distillation," in Proc. the AAAI Conference on Artificial Intelligence, vol. 38, no. 13, 2024, pp. 14 184–14 192.
[17] N. Wu, L. Yu, X. Jiang, K.-T. Cheng, and Z. Yan, "FedNoRo: Towards noise-robust federated learning by addressing class imbalance and label noise heterogeneity," in Proc. International Joint Conferences on Artificial Intelligence (IJCAI.
[18] L. Wang, J. Bian, and J. Xu, "Federated learning with instance-dependent noisy label," in Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024, pp. 8916–8920.
[19] S. Yang, H. Park, J. Byun, and C. Kim, "Robust federated learning with noisy labels," IEEE Intelligent Systems, vol. 37, no. 2, pp. 35–43, 2022.
[20] C. Wu, Z. Li, F. Wang, and C. Wu, "Learning cautiously in federated learning with noisy and heterogeneous clients," in 2023 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2023, pp. 660–665.
[21] J. Li, J. Pei, and H. Huang, "Communication-efficient robust federated learning with noisy labels," in Proc. the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 914–924.
[22] X. Zhou and X. Wang, "Federated label-noise learning with local diversity product regularization," in Proc. the AAAI Conference on Artificial Intelligence, vol. 38, no. 15, 2024, pp. 17 141–17 149.

Yufei Kang received her B.Engr. degree in Telecommunication Engineering from Xidian University, China, in 2019. She is currently working toward the Ph.D. degree in the iQua group at the Department of Electrical and Computer Engineering, University of Toronto, Canada. Her research interests include cloud computing, edge computing, and federated learning.

Chenghao Hu is currently a Ph.D. student in the Department of Electrical and Computer Engineering, University of Toronto, Canada. He received his B.Engr. degree from South China University of Technology, China, and his M.S. degree from Tsinghua University, China. His research interests include distributed machine learning, efficient training, and model deployment.

Baochun Li received his B.Engr. degree from the Department of Computer Science and Technology, Tsinghua University, China, in 1995, and his M.S. and Ph.D. degrees from the Department of Computer Science, University of Illinois at Urbana-Champaign, in 1997 and 2000, respectively. Since 2000, he has been with the Department of Electrical and Computer Engineering, University of Toronto, where he is currently a Professor. He holds the Bell Canada Endowed Chair in Computer Engineering since August 2005. His research interests include cloud computing, security and privacy, distributed machine learning, federated learning, and networking. Dr. Li has co-authored more than 490 research papers, with over 28,700 citations, an H-index of 91, and an i10-index of 366 (Google Scholar). He was the recipient of the IEEE Communications Society Leonard G. Abraham Award in 2000, the Multimedia Communications Best Paper Award from the IEEE Communications Society in 2009, the University of Toronto McLean Award in 2009, the Best Paper Award from IEEE INFOCOM in 2023, and the IEEE INFOCOM Achievement Award in 2024. He is a Fellow of the Canadian Academy of Engineering, a Fellow of the Engineering Institute of Canada, and a Fellow of IEEE.