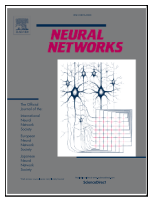
Neural Networks 201 (2026) 108889

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



Review

Federated learning with noisy labels: A comprehensive and concise review of current methodologies and future directions

Jia Dong [a,*], Rui Zhu [b], Xinyi Shang [a], Jing-Hao Xue [a]

[a] Department of Statistical Science, University College London, London, WC1E 6BT, UK
[b] Bayes Business School, City St George's, University of London, London, EC1Y 8TZ, UK

ARTICLE INFO

Keywords:
Federated learning
Noisy label learning
Federated learning with label noise
Comprehensive review

ABSTRACT

Federated learning, a vital paradigm in modern machine learning, enables private and decentralised training of models that is crucial for learning from sensitive data. Noisy label learning, another vital paradigm in modern machine learning, addresses the training of models from the data with potentially incorrect labels. Their integration, namely *federated learning with noisy labels* (FLNL), is an emerging but challenging topic arising from the practice of machine learning, which, however, still lacks a review of its research progress. The aim of this paper is to fill in this gap. We first summarise four core challenges to FLNL: localised label noise, across-client heterogeneity of label noise, localised overfitting to label noise, and inadequate benchmarking. We then propose a taxonomy to categorise current FLNL studies into four types that address the four challenges correspondingly: sample-wise methods, client-wise methods, model-wise methods, and benchmark-wise studies. This work offers the first comprehensive and concise review dedicated to FLNL; moreover, we also provide future research directions for this rapidly evolving and practically significant field.

1. Introduction

Federated learning is a decentralised paradigm in modern machine learning, in which multiple clients collaboratively train a global model under the coordination of a central server, while not sharing their local data. This privacy-preserving paradigm is particularly crucial in domains such as healthcare (Li et al., 2025), finance (Kang et al., 2024), mobile applications (Gecer & Garbinato, 2024), and smart cities (Al-Huthaifi et al., 2023), where data privacy and security are paramount.

In traditional machine learning, the labels of training data are usually assumed to be correct. However, in real-world practice of machine learning, noisy labels are not uncommon, due to, for example, annotation errors and crowdsourced labelling, especially as datasets continue to scale and diversify (Johnson & Khoshgoftaar, 2022; Shi et al., 2024; Zhang et al., 2025). Unfortunately, the models learnt from the data with noisy labels can overfit to those noisy labels, leading to poor generalisation (Song et al., 2025, 2022). Hence in modern machine learning, there is another paradigm termed noisy label learning (or label-noise learning), which focuses on reliably training models from the data with potentially incorrect labels.

Compared with the situation in traditional (usually centralised) machine learning, the problem of model learning from the data with noisy labels is even more challenging in federated learning. The combination of decentralised data ownership, privacy constraints and noisy labels introduces the complexities that are rarely encountered in either federated learning or noisy label learning alone. In particular, federated learning with noisy labels (FLNL) must simultaneously ensure privacy preservation, handle data and noise heterogeneity and maintain robust model convergence, which makes it much more challenging than standard noisy label learning. We first summarise four core challenges faced by FLNL as follows.

- First, due to privacy constraints in federated learning, data with noisy labels are held locally within individual clients and cannot be shared and centrally identified and corrected at the global server. For example, in a federated healthcare system, each hospital can annotate their medical images locally, but incorrect labels cannot be corrected centrally due to privacy constraints. Hence, unlike centralised settings where data cleansing can be performed globally, the first challenge to FLNL is how to develop effective local mechanisms to identify and correct localised noisy labels at individual clients (Ji et al., 2024; Jiang et al., 2024; Jiang & Zhang, 2025).
- Second, the severity of label noise in federated learning is heterogeneous across local clients, as each client holds their private and

* Corresponding author.
E-mail address: jia.dong.23@ucl.ac.uk (J. Dong).

https://doi.org/10.1016/j.neunet.2026.108889
Received 29 August 2025; Received in revised form 24 February 2026; Accepted 21 March 2026
Available online 22 March 2026

decentralised datasets that are usually collected and labelled in uncontrolled local environments. For example, in a federated system of e-commerce platforms with annotators of quite different qualities, the global server needs to prevent the noisy updates provided by low-quality platforms from compromising the high-quality contributions of high-quality platforms. Hence, the second challenge to FLNL is how to tackle across-client heterogeneity of label noise at the global server side (Ali & Arafa, 2025; Morafah et al., 2025; Zeng et al., 2024a).

- Third, deep neural networks are known to be able to overfit noisy labels, which can severely impair generalisation. In the federated setting, such an overfitting is exacerbated by the lack of centralised control, the limited local data, and the heterogeneity across local clients. For example, a mobile-device model trained on a small, noisy dataset may overfit to wrong patterns, and such a local model can degrade the performance of the global model after aggregation. Hence, the third challenge is how to design tailored optimisation schemes for FLNL to mitigate the overfitting to those heterogeneous noisy labels (Lu et al., 2024; Pu et al., 2025; Yu et al., 2025).
- Fourth, the current benchmarking protocols for federated learning or centralised noisy label learning are inadequate in the FLNL setting. The benchmarking protocols of centralised noisy label learning fail to capture the heterogeneity and privacy constraints inherent in federated learning, while the benchmarking protocols for federated learning often assume that all client data are clean. Hence, the fourth challenge is to develop an adequate benchmarking protocol for accurate and fair comparison of FLNL methods (Jiang et al., 2025; Liang et al., 2023).

While federated learning and noisy label learning, as two separate paradigms in machine learning, have each garnered substantial attention, FLNL as their integration remains underexplored in surveys.

On the one side, existing surveys of federated learning focus on different aspects of this framework. For instance, Ye et al. (2023) analyse heterogeneous federated learning, identifying challenges in statistical, model, communication and device heterogeneity. Jia et al. (2025) emphasise communication-efficient federated learning, reviewing compression, client selection and over-the-air aggregation strategies for mobile edge environments. Sabah et al. (2024) review personalised federated learning, analysing optimisation strategies that balance global generalisation and local adaptation. Kim et al. (2025) examine fair federated learning, presenting a taxonomy of fairness-enhancing mechanisms and their trade-offs with privacy. Hu et al. (2024) focus on security and privacy threats in federated learning, covering adversarial attacks and defences across secure aggregation and differential privacy frameworks. However, these surveys on federated learning usually assume that client data are clean and thus do not consider the impact of label noise in federated learning, for example, the impact of noisy local updates from clients on the aggregation into the global model at the server. Hence, the findings from these surveys are insufficient for the research of FLNL.

On the other side, existing surveys of noisy label learning are confined to centralised learning. For instance, Song et al. (2022) group the methods for learning with noisy labels into five categories: robust architecture, robust regularisation, robust loss function, loss adjustment and sample selection. Li and Zhu (2024) analyse noisy label learning for classification, highlighting semi-supervised and contrastive learning as state-of-the-art approaches. Shi et al. (2024) focus on medical image analysis, summarising strategies designed to handle annotation inconsistency and domain specific noise. Song et al. (2025) review deep learning under noisy supervision, covering loss regularisation, robust regression, sample reweighting, noise transition modelling and semi-supervised learning. However, these surveys assume centralised full access to the entire dataset, which does not hold in federated learning, where data are distributed and cannot be centrally accessed. Consequently, they do not consider any label noise heterogeneity across local clients. Therefore, their findings cannot be readily generalised to FLNL, where label noise is both decentralised and heterogeneous.

That is, up to present, there is no survey of FLNL. Hence, by filling in this gap, this paper aims to present the first comprehensive and concise review dedicated to FLNL, as well as to offer a roadmap to future research in this rapidly evolving and practically significant field.

In this survey, as illustrated in Fig. 1, we categorise existing FLNL methods into four groups. This taxonomy is derived from a comprehensive literature analysis, revealing that current research has made efforts to address either of the four core challenges discussed above. The first group is for *sample-wise methods* (Section 2), which address noisy labels mainly locally at individual client sides. The second group is for *client-wise methods* (Section 3), which focus on mitigating the impact of across-client heterogeneity of label noise at global server side. The third group is for *model-wise methods* (Section 4), which design tailored optimisation schemes to mitigate the overfitting to heterogeneous noisy labels. The fourth group concerns *benchmark-wise studies* (Section 5) that are specifically conducted for FLNL. It is important to note that, while these categories represent the primary focus of the methods, some overlap is inevitable. Our classification of each work is therefore based on its principal novel contribution and which challenge it is primarily designed to address. Beyond our comprehensive reviews of FLNL methods, we also discuss limitations and future research directions (Section 6) for each group of FLNL methods.

2. Sample-wise methods

Sample-wise methods focus on instance-level solutions to the issue of label noise mainly locally at individual clients. Within this category, three types of approaches have emerged, namely *sample selection*, *sample weighting* and *label correction*, as illustrated in Fig. 1 and Fig. 2. Sample selection (Section 2.1) aims to identify samples with clean labels and remove samples that are likely to be with noisy labels. Sample weighting (Section 2.2) adjusts the relative contribution of samples during training, reducing the influence of noisy ones without discarding them. Label correction (Section 2.3) goes a step further by attempting to correct unreliable labels.

2.1. Sample selection

Sample-selection methods primarily aim at identifying and preserving clean samples while filtering out label-noisy samples from local datasets before training. These methods provide mainly local mechanisms for improving the label quality of the training data.

One of the earliest sample-selection methods is proposed by Tuor et al. (2021). In this method, a benchmark model is first trained on a small, clean benchmark dataset; using this benchmark model, each client then evaluates the loss values of its local samples; by statistically comparing the distributions of client losses with that of the benchmark dataset, a global filtering threshold is determined; and finally, each client discards the data samples whose losses exceed the threshold and retains only the low-loss samples for subsequent federated training. This method ensures that only reliable data are used in training. However, it relies on a clean dataset external to each client, which increases deployment complexity.

To address this limitation, Fed-DR-Filter Duan et al. (2022) constructs class-specific k-nearest-neighbour subgraphs, where samples within the largest connected component of each subgraph are initially identified as clean; subsequently, each retained sample is revalidated by the proportion of same-class neighbours among its k nearest neighbours, and only those samples surpassing a predefined neighbour-similarity threshold are retained. This method does not require clean data.

Another limitation of Tuor et al. (2021) is its tendency to discard samples with high training losses indiscriminately, inadvertently excluding representative, yet challenging, samples. To address this issue, FedSPL (Wang & Zhou, 2022) employs two homogeneous models to jointly

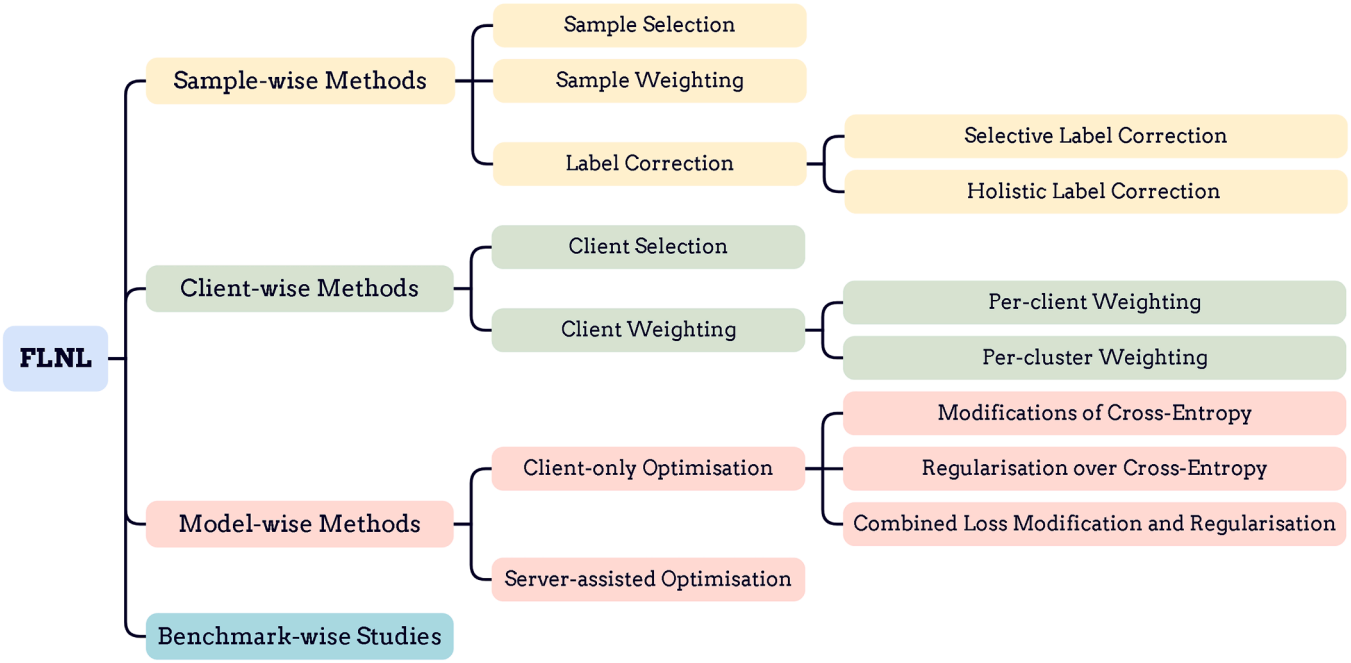


Fig. 1. The taxonomy of FLNL methods. On the second column, each of the four groups of methods corresponds to one of the four core challenges: sample-wise methods handle local label noise, client-wise methods mitigate client noise heterogeneity, model-wise methods develop robust optimisation strategies, and benchmark-wise studies design evaluation protocols and datasets.

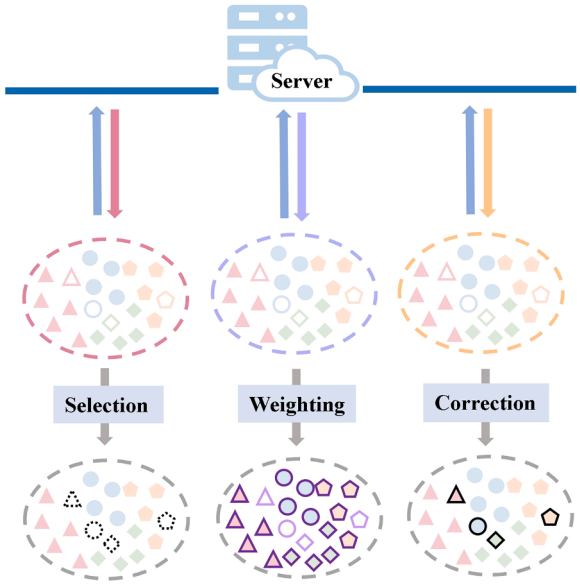


Fig. 2. Illustration of three types of sample-wise approaches (from left to right: sample selection, sample weighting, and label correction) for handling noisy labels at a local client. Different coloured shapes denote different labels, while unfilled shapes indicate noisy labels. After sample selection, dashed shapes represent the noisy samples filtered out. After sample weighting, coloured outlines indicate that different weights are assigned to samples. After label correction, black outlined shapes represent the noisy labels that have been corrected.

predict labels for each batch of data, and then dynamically sets adaptive thresholds based on the intra-class mean and variance of these predictions, selectively identifying easy-to-learn samples for initial training phases. Then, through a self-paced learning strategy, FedSPL (Wang & Zhou, 2022) progressively incorporates more difficult yet correctly labelled samples into training, rather than prematurely discarding them all, thus effectively mitigating the limitation of neglecting challenging but informative samples.

Moreover, in order to consider the significant heterogeneity in label noise commonly found in federated learning scenarios, FedFixer (Ji et al., 2024) emphasises mutual collaboration between a local personalised model and a global model within each client, alternately selecting low-loss samples from the counterpart model's predictions. FedCNL (Sun et al., 2024) proposes a curriculum-based federated learning approach to tackle heterogeneity, which first employs a noise modelling module to adaptively distinguish clean clients from noisy clients and further identifies clean and noisy samples within each client in an unsupervised manner, and then progressively incorporates data from clean to increasingly noisy samples, ensuring stable training under varying noise conditions. PPASNL (Han & Yan, 2023) addresses the heterogeneity in label noise by employing a supervised contrastive learning model at the client side to calculate feature-difference scores, where a high score indicates that the sample is more likely to be noisy.

2.2. *Sample weighting*

Sample-selection approaches typically rely on hard filtering strategies such as removing high-loss samples or assume that clean and noisy samples can be reliably separated by using fixed criteria, which can be impractical in real-world settings. In contrast, sample-weighting methods aim to improve generalisation by assigning adaptive weights to training samples, allowing the model to down-weight unreliable examples rather than discarding them.

To instantiate this idea, Comm-FedBiO (Li et al., 2022) formulates the noisy label problem as a federated bi-level optimisation problem: The inner problem minimises the weighted training loss over all clients, while the outer problem minimises the validation loss at the server. Therefore, instead of explicitly identifying noisy labels, Comm-FedBiO (Li et al., 2022) assigns each sample a continuous weight that reflects its contribution to generalisation. Fed-NL (Mishra & Gupta, 2025) assigns each training sample a confidence-based weight derived from its prediction reliability, where samples with higher confidence receive

larger weights and suspected noisy samples are down-weighted. This weighting strategy allowing clients to mitigate the impact of noisy data.

2.3. Label correction

Label correction refers to strategies that aim to mitigate the adverse impact of noisy labels by refining or replacing them with more reliable ones. In FLNL, label-correction methods can be primarily classified into two categories: Selective label correction (Section 2.3.1) involves initially selecting samples by distinguishing between clean and noisy ones, preserving labels of clean samples while assigning pseudo-labels to noisy samples for subsequent learning; holistic label correction (Section 2.3.2) corrects labels of all samples irrespective of their original cleanliness.

2.3.1. Selective label correction

Sample-selection strategies often assume that high-loss samples are likely to be with noisy labels. However, high-loss samples may also include hard but correctly labelled instances that are critical for generalisation, and discarding such samples may lead to biased learning and degraded performance. Sample-weighting methods can make use of noisy samples, but when the noise ratio is high, simply reducing their weights is still insufficient to prevent the propagation of incorrect supervision signals. Therefore, selective label-correction methods aim to first distinguish clean and noisy samples based on certain confidence or structure-based metrics, and noisy ones are then relabelled using pseudo-labels generated by the model or aggregated knowledge. This process ensures that the model training can still exploit noisy samples in an informed way.

In this direction, FedNoiL (Wang et al., 2022) uses the global model to compute confidence scores for local samples, classifying them into clean and noisy subsets based on confidence scores, with noisy samples assigned pseudo-labels derived from the global model's predictions. A similar strategy is adopted by RobustFL (Yang et al., 2022) but using local and global class-wise centroids to evaluate the confidence of samples. FedCorr (Xu et al., 2022) classifies each client's data into clean and noisy samples through Gaussian mixture modelling of sample losses. FedSTSS (Rong et al., 2023) uses both cross-entropy loss and class-probability entropy as metrics to classify samples into clean and noisy subsets.

As FedCorr (Xu et al., 2022) relies on a single local model to identify noisy samples, which may lead to local confirmation bias, FedCoop (Tam et al., 2023) mitigates this by collaboratively computing comprehensive reliability scores by using both local models and global features aggregated by the server. Moreover, for samples identified as noisy, their pseudo-labels are the labels of the most similar global class prototypes, hence FedCoop (Tam et al., 2023) effectively mitigates local confirmation bias by leveraging cross-client feature alignment.

In parallel to confirmation bias, another limitation of FedCorr (Xu et al., 2022) is that its Gaussian mixture models (GMMs) of losses are fitted independently on each client's limited data, which may result in inaccurate or unstable estimates of the noise distribution, especially in small-sample regimes. To overcome this, FedDiv (Li et al., 2024b) proposes a collaborative noise filtering mechanism that aggregates GMM parameters across clients to form a more accurate global noise model.

An assumption made by some label-correction methods is that clean clients exist (Jiang et al., 2024; Xu et al., 2022), which does not hold in many real-world scenarios, such as crowdsourcing or adversarial attacks, where all clients may suffer from noisy labels. To address this, FedClean (Jiang & Zhang, 2025) introduces a two-stage collaborative correction mechanism, in which label correction is collaboratively guided by local noisy label learning and the global model, ensuring robust performance even when all clients contain noisy labels.

Most methods aforementioned assume that label noise is either closed-set noise (where true labels are in a predefined set) or open-set noise (where true labels are outside a predefined set). One step further, FedMIN (Zeng et al., 2024b) proposes to handle mixed label noise by modelling both closed-set and open-set noises, with pseudo-labels assigned to closed-set noisy samples only, enabling more accurate treatment.

2.3.2. Holistic label correction

As in practice all samples may carry uncertain or incorrect labels, holistic label-correction methods are proposed, which treat all samples as potentially noisy, improving the robustness against label noise without relying on explicit sample separation.

To instantiate this idea, FedELC (Jiang et al., 2024) introduces a trainable label correction mechanism, in which each sample's label is modelled as a latent random variable rather than a fixed input, where label distributions are updated and labels are refined during local training. DFLMV (Huang & Shu, 2024) proposes that clients collaboratively denoise their data with neighbouring clients' models in a peer-to-peer manner: For each local sample, the client gathers predictions from its neighbours and updates the label based on the majority voting. FedFDC (Ma et al., 2025) corrects labels through a temporal dual-view consistency mechanism, where labels are updated only if predictions from strong and weak augmentations remain consistent across multiple communication rounds, ensuring reliable label refinement.

3. Client-wise methods

Client-wise methods address noisy labels at the client level rather than the sample level. We summarise the client-wise methods into two types of approaches, namely *client selection* and *client weighting*, as illustrated in Fig. 1 and Fig. 3.

Client selection (Section 3.1) aims to identify reliable clients and exclude those with high levels of noise. Client weighting (Section 3.2), instead of discarding noisy clients, adjusts each client's contribution to the global model according to its estimated reliability, either without or with client clustering first (Fig. 3; also Section 3.2.1 and Section 3.2.2, respectively).

3.1. Client selection

Client selection methods primarily aim at choosing clients with fewer noisy labels in each communication round while excluding those clients whose updates are persistently harmful. These methods provide a global mechanism for improving the robustness of the aggregated global model under client-heterogeneous label noise.

Identifying noisy samples within each client can introduce substantial computational overhead, posing a challenge for edge devices with limited local resources. To reduce this burden, Yang et al. (2021) propose a server-driven client selection algorithm that uses a small clean validation dataset maintained by the server to evaluate uploaded local models, rank client reliability, and prioritise clients with lower noise, thus ensuring client selection without any additional computation on the client side.

Another limitation of sample-wise strategies is that label correction assumes a reasonably well-trained global model, which may not hold under high noise and non-independent and identically distributed (non-IID) data. To address this, ClipFL (Morafah et al., 2025) shifts the focus from correcting noisy labels to removing persistently unreliable clients based on performance trends, introducing a noise candidacy score that tracks how often a client performs poorly on a clean validation set across rounds and excluding those with the highest scores, hence mitigating noise propagation without relying on a strong global model.

Moreover, most existing methods assume a fixed set of clients, making it difficult to maintain performance when new clients join during training. FedDC (Giap et al., 2025) addresses this by extending FedCorr (Xu et al., 2022) to support dynamic clients, initially treating newcomers as noisy and gradually refining their estimates. This design allows the model to continue training without disruption, ensuring steady performance in real-time systems.

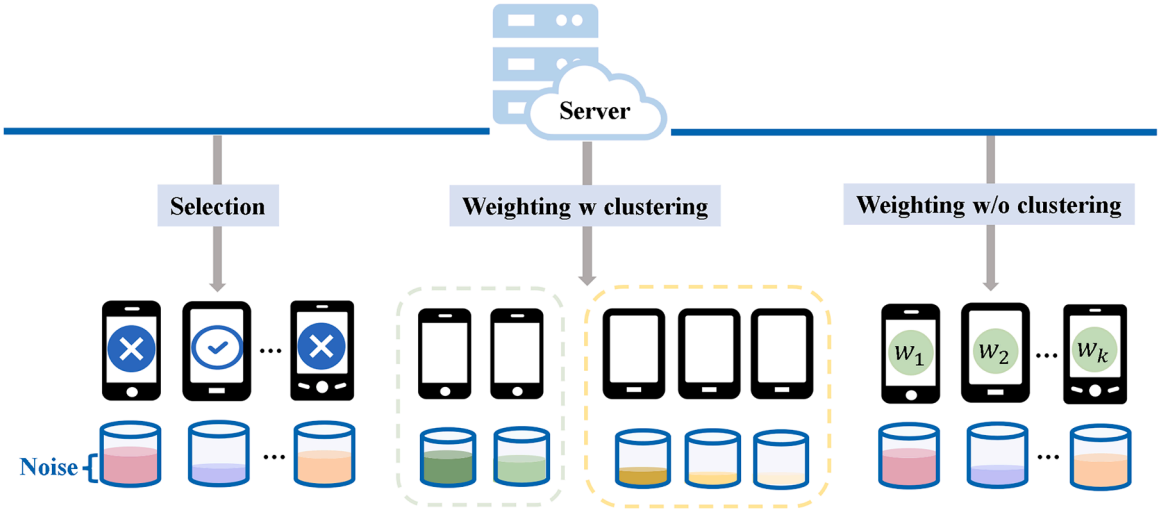


Fig. 3. Illustration of three types of client-wise approaches (from left to right: client selection, client weighting with clustering first and client weighting without clustering first) for handling noisy labels at the client level. Each transparent cylinder represents the client's local dataset, while the coloured inner cylinder represents the portion of data with noisy labels. Different colours indicate different levels of label noise across clients. The larger the coloured portion within a cylinder, the higher the proportion of noisy samples in that client. After client selection, unreliable clients are excluded from global aggregation. With clustering first, clients with similar data distributions are grouped for joint training and client weighting. After weighting, clients are assigned different weights for global aggregation according to their estimated reliability.

3.2. Client weighting

Client weighting refers to the strategies that reduce the influence of unreliable clients while still retaining their contributions to the global model. Unlike client selection, which completely excludes clients with high noise ratios, client-weighting schemes assign different aggregation weights so that reliable clients have a stronger influence while noisy clients are assigned smaller weights rather than being discarded. In FLNL, client-weighting methods can be primarily classified into two categories: Per-client weighting (Section 3.2.1) directly assigns weights to individual clients based on their estimated reliability; per-cluster weighting (Section 3.2.2) first clusters clients into groups and then assigns weights at the group level.

3.2.1. Per-client weighting

Per-client weighting methods assign aggregation weights to individual clients according to the estimates of their reliability. Unlike per-cluster weighting methods in Section 3.2.2, Per-client methods operate at the client level without requiring prior clustering of clients.

Li et al. (2023) weight clients by learning efficiency, where a larger reduction in local training loss indicates higher data quality, and aggregation weights are assigned proportionally to this reduction with softmax normalisation ensuring adaptivity across clients. FedLN (Tsouvalas et al., 2024) uses energy scores derived from model outputs to estimate client noise levels, and clients with a higher proportion of samples exceeding a confidence threshold are assigned lower aggregation weights. DETECTION (Wu et al., 2024) adopts local intrinsic dimensionality to assess the client reliability without requiring label information, where higher values indicate noisy data and thus lower aggregation weights to such clients. FedES (Zeng et al., 2024a) evaluates the discrepancy between the local and global parameter importance distributions to estimate the reliability of a client. Aorta (Xu et al., 2024) adopts a reference-guided strategy that weights clients by the cosine similarity between their local models and a clean reference model, rewarding updates that align with the desired learning direction.

A limitation of the above per-client weighting methods is their reliance on fixed scoring metrics that are directly mapped to aggregation weights, which lacks a global optimisation perspective and may lead to suboptimal performance. To address this, FedDPSO (Ouyang et al., 2025) employs a particle swarm optimisation framework that searches over candidate weight vectors to minimise the prediction error of the aggregated model on a validation dataset.

3.2.2. Per-cluster weighting

Per-cluster weighting methods address the difficulty of estimating client reliability from limited local data, by clustering clients into clusters based on properties such as data similarity or estimated noise levels. Aggregation weights are then assigned at the cluster level, which can stabilise reliability estimation under data scarcity and imbalance.

In this direction, FedNoRo (Wu et al., 2023) first uses GMMs of normalised loss vectors to cluster clients into noisy and clean groups. Then, clean clients retain full weight, while noisy clients receive reduced weights that decay with their model distance from clean clients. However, due to the use of loss vectors, the clustering of FedNoRo can be unstable under severe label noise. RCC-PFL (Ali & Arafa, 2025) circumvents this issue by performing label-agnostic clustering based on structural data representations extracted by clients.

4. Model-wise methods

In FLNL, deep neural networks are prone to localised overfitting, where each client model tends to memorise its own noisy labels due to limited and non-IID data. This phenomenon amplifies the impact of noise during aggregation, leading to unstable or biased global updates. Model-wise methods enhance the model robustness against label noise by modifying optimisation objectives of training. We classify the methods within this category into two types, namely *client-only optimisation* and *server-assisted optimisation*, as illustrated in Fig. 1 and Fig. 4. Client-only optimisation (Section 4.1) enhances local training stability through loss-function modification, regularisation or noise-aware objectives that prevent overfitting to noisy labels. Loss modifications down-weight gradients from uncertain samples, regularisation constrains model complexity and smooths predictions, and noise-aware objectives distinguish clean from noisy information. Server-assisted optimisation (Section 4.2) introduces global knowledge to guide local updates and thus improve the robustness of the global model.

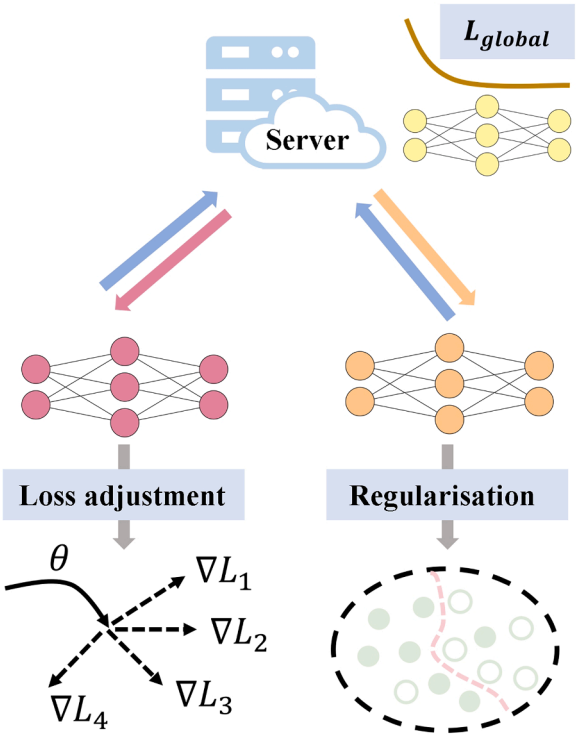


Fig. 4. Illustration of two types of model-wise methods for handling noisy labels at client and server (bottom-up: client-only optimisation and server-assisted optimisation). With client-only optimisation, techniques such as loss adjustment and regularisation modify the local objective to reduce the effect of noisy supervision and suppress overfitting. With server-assisted optimisation, global knowledge is introduced at the server side to guide local updates and thus improve the robustness of the global model.

4.1. Client-only optimisation

Client-only optimisation refers to the strategies that improve robustness against noisy labels by refining local training objectives without relying on server-side supervision. In FLNL, these methods can be further classified into three categories: The first category (Section 4.1.1) modifies the cross-entropy loss to reduce its sensitivity to noisy supervision; the second category (Section 4.1.2) introduces regularisation over the standard loss to enhance stability; the third category (Section 4.1.3) combines loss modification with regularisation to jointly suppress overfitting and mitigate the effects of label noise.

4.1.1. Modifications of cross-Entropy

Methods of loss-function modification aim to mitigate the sensitivity of standard cross-entropy loss to noisy labels. As cross-entropy tends to memorise mislabelled samples (Arpit et al., 2017), leading to biased gradients and degraded global performance, loss-function modification strategies redesign the objective so that optimisation approximates training on clean data. Such a strategy enables clients to use all local data for training, without discarding samples or explicitly identifying noisy labels.

Under this umbrella, noise-resilient federated learning (Mishra et al., 2022) replaces cross-entropy with a weighted objective that gradually reduces reliance on label supervision. Dynamic weights start high and decay over time, shifting learning from label signals toward model predictions, and reducing the risk of memorising mislabelled ones while retaining all samples. FedEFC (Yu et al., 2025) replaces cross-entropy with an enhanced forward correction loss that incorporates a noise transition matrix to adjust predictions before computing the loss. This allows clients to align with the clean label distribution despite noisy supervision without external filtering.

Most FLNL methods focus on class-conditional noise with fixed mislabelling probabilities per class, whereas in practice the label noise is also often instance-dependent. FedBeat (Wang et al., 2024) addresses this issue by integrating an instance-dependent noise transition matrix into cross-entropy, where a dedicated network learns instance-dependent noise transitions from pseudo-labels generated by Bayesian ensemble, and model predictions are corrected through this matrix to align with the clean label distribution.

4.1.2. Regularisation over cross-Entropy

Methods introducing regularisation over cross-entropy aim to enhance local training stability while preserving the simplicity of cross-entropy supervision. They retain the cross-entropy but add auxiliary terms to suppress overfitting to noisy labels. These regularisers depend only on model outputs rather than prior knowledge of label noise, making them suitable for heterogeneous federated settings.

Cross-entropy tends to make local models memorise noisy labels before global convergence, producing biased updates that are further worsened by client heterogeneity and misaligned early training. To address this, FedELR (Pu et al., 2025) introduces an early-learning regularisation that penalises deviations from a smoothed history of past predictions, delaying memorisation of noisy samples and preserving clean patterns. By aligning local learning stages and relying only on local outputs, FedELR (Pu et al., 2025) enhances the quality of client updates sent to the server. However, FedELR (Pu et al., 2025) relies only on local historical predictions, which can be unreliable under severe noise or unstable training. To overcome this, FLR (Kim et al., 2024) combines local and global exponential moving-average predictions into soft pseudo-labels that serve as stable targets for regularisation, linking current predictions to both local history and global consensus. This design not only mitigates overfitting to noisy labels but also promotes consistency across clients.

Another limitation of cross-entropy training is that it overlooks that structural relationships among samples that may help identify noisy labels. To address this, FedGP (Chen et al., 2023) augments the local objective with a graph-based regularisation, where each client constructs a similarity graph over its local data and jointly optimises both the prediction loss and the graph structure. By enforcing sparsity and neighbourhood consistency of the graph, FedGP (Chen et al., 2023) reduces the influence of samples that are inconsistent with their local neighbourhood. In federated graph learning, cross-entropy training is sensitive to noisy nodes, a problem worsened by structural heterogeneity and class imbalance across clients that hinder their detection. FedRGL (Li et al., 2024a) addresses this by augmenting cross-entropy with several regularisations designed for noisy graphs.

Moreover, cross-entropy training is vulnerable under open-set noisy labels, where clients observe only partial and corrupted sub-spaces of the global label space, leading to poor generalisation to unseen classes. FedDPCont (Di et al., 2024) tackles this by augmenting cross-entropy with a differentially private contrastive regularisation that penalises disagreement with sampled labels from a global distribution, slowing down memorisation of noisy labels and improving the alignment of client updates with the global label distribution. On the other hand, label noise also makes cross-entropy training vulnerable when adapting the global model to local data for personalisation. To address this, FedLTF (Zhan et al., 2025) combines cross-entropy with a distillation loss from a pre-trained teacher and a variance-based penalty that discourages uncertain outputs, retaining the benefits of pre-training while reducing the influence of noisy labels during adaptation.

Finally, we also note that data augmentation offers extra supervision but augmented samples inherit noisy labels. To address this, LSR (Jiang et al., 2022) adds consistency self-regularisation that penalises the discrepancy between model outputs on original and augmented samples, to reduce the influence of noisy supervision as model predictions are often closer to the ground truth.

4.1.3. Combined loss modification and regularisation

Methods to combine both loss modification and regularisation in FLNL aim to simultaneously address bias and variance under extreme conditions such as high noise rates, limited client data, or severe non-IID distributions. Robust loss functions mitigate bias by weakening the impact of noisy labels, while auxiliary regularisers help control model variance. This dual design achieves stronger robustness against label noise than using either component alone and provides flexibility due to being able to exploit the strengths of both loss modification and regularisation.

Many FLNL methods discard or underutilise noisy label data instead of extracting their potential supervision, leading to insufficient use of data. LSG (Bai et al., 2023) addresses this with a local objective combining a sharpened cross-entropy with a Kullback-Leibler divergence regulariser, enabling more effective utilisation of label-noisy data.

On the other hand, when clients have only limited local data, the optimisation of noise transition matrix (NTM) becomes unreliable due to unstable gradient updates and large estimation errors. To address this, FedLNL (Zhou & Wang, 2024) combines a noise-aware loss that incorporates sampled NTMs into the prediction and a diversity regulariser to encourage varied outputs, without direct optimisation of the NTM. As mentioned in Section 4.1.1, FedBeat (Wang et al., 2024) integrates an instance-dependent NTM into cross-entropy, enabling instance-aware correction. FedIDN (Yang et al., 2025) enhances this direction by first detecting noisy clients and then within these clients separating clean and noisy samples. Clean samples are updated with a confidence-regularised loss, noisy samples with a negative cross-entropy loss, and the two losses are combined to balance the exploitation of informative data and the suppression of noise.

A similar case of data scarcity is that, when clients are few, heterogeneous and label-noisy, techniques like contrastive learning, which require many clients to form reliable pairs, may collapse and amplify noise. FedRelaxedCL (Ejigu et al., 2025) mitigates this by combining a label-smoothed symmetric cross-entropy loss with a relaxed contrastive regulariser, which includes background-based contrastive to avoid hard negatives, feature diversity to prevent class collapse, and local-global alignment via a server prototype buffer. This combined objective improves feature quality under small-client, small-batch settings.

4.2. Server-assisted optimisation

Server-assisted optimisation methods address the limitations of purely local training under severe label noise and client heterogeneity. Without global guidance, local models may overfit noise and propagate biased updates, while the lack of global coordination hinders alignment across clients. To address these issues, server-assisted methods often introduce global mechanisms to guide local training with server knowledge, reducing label-noise overfitting and improving client consistency.

As the global model is relatively robust against label noise, FedGR (Tian et al., 2024) uses the global model as a supervisory signal for local training: The global model identifies noisy samples, clients relabel these samples with pseudo-labels, and local models are further regularised through distillation and feature-level consistency with the global model.

An issue with client-weighting methods in Section 3.2 is that they often assume moderate noise, leading to the down-weighting or exclusion of severely noisy clients and the loss of useful information. To address this, FedNed (Lu et al., 2024) introduces a negative distillation framework that leverages signals from highly noisy clients by constructing negative gradients from low-confidence updates and incorporating them into a server-side loss. This penalises the alignment with unreliable behaviour and achieves robust optimisation under severe noise without discarding clients.

Finally, a relevant topic is the extension of federated semi-supervised learning to FLNL. In federated semi-supervised learning, the server maintains a small labelled set, which is often assumed clean. In its extension to FLNL, FedCR (Mao et al., 2025) embeds noise robust objectives into the server-side training, where the server jointly optimises a generalised cross-entropy and a semantic consistency regularisation. In this way, the server can still extract denoised supervisory signals from its noisy label data to guide optimisation.

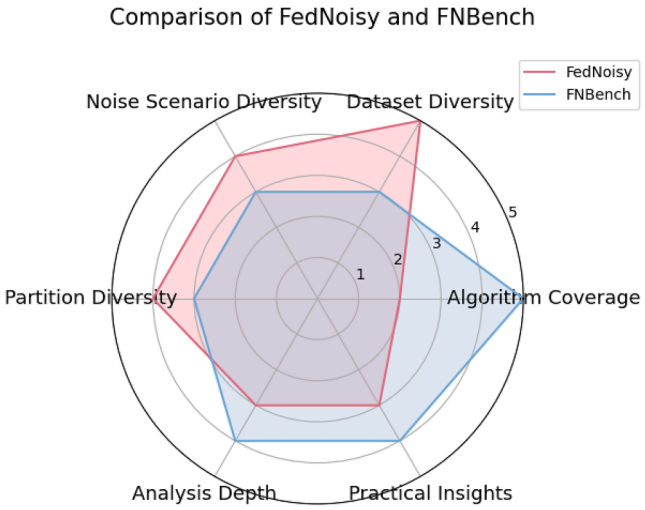


Fig. 5. Illustrative comparison of two benchmarks: FedNoisy (Liang et al., 2023) and FNBench (Jiang et al., 2025). Six aspects are scored on a 0-5 scale (higher is better). (a) Algorithm coverage refers to the range of methods covered. (b) Dataset diversity indicates the diversity of datasets considered, including those with naturally noisy labels. (c) Noise-scenario diversity shows how comprehensively different noise types are modelled (symmetric noise where labels flip uniformly, asymmetric noise where flips are class-dependent, and mixed noise). (d) Partition diversity reflects the richness of data-splitting strategies across clients (IID, sharding, quantity skew, Dirichlet label skew, and label quantity skew). (e) Analysis depth denotes the extent of studies beyond results reporting, such as mechanism explanation and ablation check. (f) Practical insights summarise the guidance offered, such as recommended baselines, regularisation tricks, or cost considerations.

5. Benchmark-wise studies

Unlike methodological proposals, benchmark-wise studies in FLNL aim to establish controlled environments with multiple settings of noise distribution and data heterogeneity, uncovering and comparing the strengths and weaknesses of FLNL methods under varied conditions.

Effective FLNL benchmarks rest on two core elements: noise modelling and data partitioning protocols. Label noise plays a crucial role, and benchmarks typically simulate several types. Symmetric noise assumes a label has an equal probability of flipping to any other incorrect class. Asymmetric noise represents more realistic error patterns where mislabelling is class-dependent, such as a “cat” likely being mislabelled as a “dog”. The most complex setting also involves client heterogeneous noise, where both the type and rate of noise differ across clients. For instance, FedNoisy (Liang et al., 2023) models this through a localised noise scene, assigning each client a noise ratio drawn from a uniform distribution, while FNBench (Jiang et al., 2025) varies noise linearly across clients. To better reflect real federated systems, modern benchmarks also emphasise non-IID data distributions to capture heterogeneity. Common approaches include sharding, which distributes class-sorted data segments; quantity skew, where clients hold differing dataset sizes; and Dirichlet label skew, which produces varied class proportions across clients.

While many studies propose novel FLNL algorithms, only a few have contributed to improving the benchmarking itself. These works are critical as they provide reusable frameworks, datasets, and protocols. FedNoisy (Liang et al., 2023) introduces the first comprehensive FLNL benchmark, supporting 20 experimental settings across six datasets that

Table 1
Comparison of the sample-wise, client-wise, and model-wise methods along three dimensions: noise type (open-set or instance-dependent noise is covered), external reliance (a clean validation set or reference model is required), and theoretical guarantee (formal robustness or convergence analysis is provided).

Category	Noise Type		External Reliance	Theoretical Guarantee
	Open-Set	Instance-Dependent		
Sample-wise	Tuor et al. (2021) FedMIN (Zeng et al., 2024b)	FedCoop Tam et al. (2023)	Tuor et al. (2021) PPASNL Han and Yan (2023) Comm-FedBio Li et al. (2022)	FedFixer Ji et al. (2024) FedClean Jiang and Zhang (2025) Fed-DR-Filter Duan et al. (2022) Comm-FedBiO Li et al. (2022) DFLMV Huang and Shu (2024)
Client-wise	/	FedNoRo Wu et al. (2023)	ClipFL Morafah et al. (2025) (Yang et al., 2021) Aorta Xu et al. (2024) FedDPSO Ouyang et al. (2025)	FedES Zeng et al. (2024a)
Model-wise	FedDPCont Di et al. (2024)	FedBeat Wang et al. (2024) FedIDN Yang et al. (2025)	FedNed Lu et al. (2024) FedCR Mao et al. (2025)	FedEFC Yu et al. (2025) FedELR Pu et al. (2025) FedLNL Zhou and Wang (2024)

Table 2
Summary of strengths, weaknesses, and future research directions of the four categories of FLNL methods, highlighting key challenges and insights from Section 6.

Category	Strengths	Weaknesses	Future Directions
Sample-wise	Identify and correct noisy samples locally. Easy to integrate with local training.	Hard samples may be mistaken as noisy. Pseudo-label errors accumulate.	Use multi-criteria to distinguish samples. Explore soft pseudo-labelling.
Client-wise	Downweight unreliable clients in aggregation. Adaptable to heterogeneous settings.	Low interpretability of reliability scores. Need clean validation data or reference models.	Enhance interpretability using explainable methods. Develop self-supervised reliability estimation.
Model-wise	Improve robustness through loss and regularisation. Reduce overfitting to noisy labels during optimisation.	Local biases amplify in aggregation. Dependence on clean data, ignoring value of noisy samples.	Mitigate bias through cross-client collaboration. Leverage noise via negative learning.
Benchmark	Provide fair comparison among FLNL methods. Ensure reproducible large-scale evaluation.	Limited to noise types. Limited to datasets and tasks. Limited to evaluation metrics.	Build datasets with diverse noise types. Benchmark across diverse datasets, tasks, and models. Include metrics such as speed and fairness.

encompass synthetic globalised and localised noise as well as real-world noise. It provides a unified pipeline and integrates nine baselines from centralised noisy label learning into the setting, enabling controlled and reproducible experimentation. FNBench (Jiang et al., 2025) focuses on realistic and heterogeneous client-level label noise and systematically benchmarks 12 representative methods under these scenarios. It shows that algorithms robust in uniform settings often degrade under more skewed distributions, and diagnoses memorisation failure and representation collapse as the chief causes of performance degradation. Both studies also aim to promote the development of FLNL algorithms that are robust across diverse noise scenarios, by exposing the performance limitations under inconsistent evaluation protocols. A comparison of these two studies is illustrated in Fig. 5. Both studies provide deep analysis and practical insights beyond simple performance comparisons. By releasing open-source frameworks and exposing the limitations of existing methods under diverse conditions, they provide the groundwork for developing future FLNL methods.

6. Future research directions

6.1. Sample-wise methods

Current sample-wise methods still suffer from hard samples. Hard samples are correctly labelled but difficult to classify and frequently exhibit high losses, especially during early or unstable training. As a result, they are easily mistaken as noisy samples and excluded from subsequent training, leading to models underfitting them and generalising poorly. Hence, a future research direction is to better distinguish hard samples from label-noisy samples in FLNL. A concrete approach is to develop multi-criteria decision frameworks that dynamically assign weights or training priorities to ambiguous samples by aggregating evidence from multiple indicators.

Moreover, once the samples with noisy labels are identified, many sample-wise methods perform pseudo-label correction by directly replacing the original labels with the model's current predictions. While this strategy is widely adopted, it is highly sensitive to prediction errors or uncertainty of the model, again especially when the model is still in an early or unstable stage of training, leading to potential error aggregation or confirmation bias. Therefore, a future research direction is to monitor and progressively improve the quality (including both accuracy and uncertainty) of pseudo-labels during training. For example, a potential solution meriting investigation is to leverage soft pseudo-labelling, which uses probability distributions rather than hard decisions of labels, allowing the model to incorporate uncertainty into a reliable learning of noisy labels.

6.2. Client-wise methods

For client-wise methods, although various sophisticated strategies have been proposed to assess client reliability, the interpretability of these reliability scores remains limited and the lack of transparency hinders the practical deployment of these strategies. Hence, a future research direction is to improve the interpretability of client selection/weighting mechanisms in FLNL. In this direction, interpretable methods such as Local Interpretable Model-agnostic Explanations and Shapley values merit consideration for explaining the reliable contribution of individual clients to the global model under label noise.

Another common limitation of many current FLNL methods is their reliance on clean validation datasets or reference models trained on centralised data on the server, which are often unavailable in real-world scenarios of federated learning, where privacy, heterogeneity or scarcity can make centralised clean supervision infeasible. Therefore, a more practically promising future direction is to develop self-supervised schemes for estimating client reliability at local client sides, without requiring any centralised data on the server, or to develop generative techniques for synthesising clean data at the server side.

6.3. Model-wise methods

In FLNL, local models tend to develop biased representations when trained on noisy samples. These local biases are often amplified through model aggregation, ultimately degrading the quality of the global model. In scenarios with severe noise and highly skewed data distributions, biased updates from even a few clients can significantly distort the global optimisation process. Hence, a future direction is to effectively reduce local model biases, for example, by leveraging cross-client collaboration.

Moreover, current model-wise methods largely rely on presumed clean data to guide optimisation, overlooking the complementary potential of noisy or low-quality samples for improving robustness. While negative learning has shown promise in centralised contexts by explicitly teaching models to avoid overfitting incorrect patterns, its application in federated settings remains largely unexplored. Hence, it would be an interesting future direction to leverage the complementary potential of negative learning for FLNL, especially as clean data can be scarce or unreliable in FLNL.

6.4. Benchmark-wise studies

Existing benchmarks primarily employ synthetic label noise for evaluation. While FNBench takes a step forward by incorporating some human annotation errors, the exploration of noise types remains rather limited. Hence, a pressing future direction is to develop richer and more realistic label-noisy datasets, even beyond existing datasets such as Clothing1M and CIFAR-N, for evaluation on more complex noise types, such as instance-dependent noise and open-set noise.

Moreover, nearly all existing benchmark studies concentrate on convolutional neural network-based image classification tasks, restricting our ability to assess the generalisability of FLNL methods. Hence, a future direction is to benchmark FLNL methods across a broader spectrum of data types such as text, graphs and sequential data, tasks such as natural language processing, recommender systems and time-series analysis, and model architectures such as Transformers and graph neural networks.

Finally, we note that future benchmarks should also include a broader set of evaluation metrics beyond accuracy and F1 score; some examples of such metrics are communication and computation cost, convergence speed and fairness.

7. Discussion

In Table 1, each category of existing FLNL methods is summarised along three additional dimensions meriting further investigation. In terms of noise type, most methods assume class-dependent and closed-set label noise, while relatively few works consider instance-dependent or open-set noise. This reflects the lack of noise diversity in existing methods, whereas instance-level or out-of-distribution noise remains underexplored. Regarding external reliance, only a small subset of methods depend on a global validation dataset, whereas the others are designed to operate locally without such supervision. Finally, only a limited number of methods provide formal theoretical guarantees of robustness or convergence, indicating that theoretical understanding of noisy label robustness in FLNL remains an open challenge.

Table 2 summarises the strengths, weaknesses, and future directions of the FLNL methods, showing the trade-offs among them. Sample-wise methods are simple and effective for local noise handling, but may mistake hard samples for noisy ones and accumulate pseudo-label errors. Client-wise methods strengthen global learning by reducing the impact of unreliable clients, but their reliability measures are often hard to interpret and sometimes depend on clean validation data. Model-wise methods strengthen stability through loss design and regularisation, but inconsistencies across clients can still accumulate, especially with limited clean data.

In addition to these design considerations, the choice of methods should also depend on the level of label noise and the degree of data heterogeneity. Methods such as FedEFC (Yu et al., 2025), FedNed (Lu et al., 2024), FedDPSO (Ouyang et al., 2025), and FedDPCont (Di et al., 2024) demonstrate robustness under high or extreme noise, while Aorta (Xu et al., 2024) and FedNoRo (Wu et al., 2023) are better suited for highly non-IID settings.

8. Conclusion

This paper presents the first comprehensive and concise review of research progress of FLNL. We identify four core challenges to FLNL: localised label noise, across-client label-noise heterogeneity, localised overfitting to label noise, and inadequate benchmarking. According to their efforts to tackle the corresponding challenges, we divide existing FLNL studies into four categories: sample-wise methods, client-wise methods, model-wise methods, and benchmark-wise studies. Sample-wise methods mitigate instance-level noise through selection, weighting or label correction, but may discard hard samples or propagate incorrect pseudo-labels. Client-wise methods handle client-level noise via selection or reliability weighting, but sometimes rely on clean data or uninterpretable metrics. Model-wise methods enhance robustness through local objective changes or server guidance, but can introduce local bias under heterogeneity. Benchmark-wise studies provide valuable evaluation frameworks, but remain limited in scope. Future research should better distinguish hard from noisy samples, improve client reliability measures, mitigate model bias, and develop more realistic benchmarks. This comprehensive review fills in a significant gap in the literature, offering valuable insights for interested researchers to advance FLNL, a vital but challenging field in modern machine learning.

CRediT authorship contribution statement

Jia Dong: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization; **Rui Zhu:** Writing – review & editing, Validation, Supervision; **Xinyi Shang:** Writing – review & editing, Validation; **Jing-Hao Xue:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Al-Huthaifi, R., Li, T., Huang, W., Gu, J., & Li, C. (2023). Federated learning in smart cities: Privacy and security survey. *Information Sciences*, *632*, 833–857.

Ali, A., & Arafa, A. (2025). Rcc-pfl: Robust client clustering under noisy labels in personalized federated learning. arXiv:2503.19886.

Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y. et al. (2017). A closer look at memorization in deep networks. In *International conference on machine learning* (pp. 233–242). PMLR.

Bai, D., Wang, S., Wang, W., Wang, H., Zhao, C., Yuan, P., & Chen, Z. (2023). Overcoming noisy labels in federated learning through local self-guiding. In *2023 IEEE/ACM 23Rd international symposium on cluster, cloud and internet computing (CCGrid)* (pp. 367–376). IEEE.

Chen, Z., Li, W., Xing, X., & Yuan, Y. (2023). Medical federated learning with joint graph purification for noisy label learning. *Medical Image Analysis*, *90*, 102976.

Di, Z., Zhu, Z., Li, X., & Liu, Y. (2024). Federated learning with local openset noisy labels. In *European conference on computer vision* (pp. 38–56). Springer.

Duan, S., Liu, C., Cao, Z., Jin, X., & Han, P. (2022). Fed-DR-filter: Using global data representation to reduce the impact of noisy labels on the performance of federated learning. *Future Generation Computer Systems*, *137*, 336–348.

Ejigu, G. F., Adhikary, A., & Hong, C. S. (2025). Relaxed contrastive learning for robust federated models with noisy labels and limited clients. In *2025 27Th international conference on advanced communications technology (ICACT)* (pp. 1–6). IEEE.

Gecer, M., & Garbinato, B. (2024). Federated learning for mobility applications. *ACM Computing Surveys*, *56*(5), 1–28.

Giap, T.-T., Kieu, T.-D., Le, T.-L., & Tran, T.-H. (2025). Feddc: Label noise correction with dynamic clients for federated learning. *IEEE Internet of Things Journal*, *12*(8), 10266–10277.

Han, W., & Yan, X. (2023). A privacy preserving federated learning aggregation algorithm for noise label. In *2023 9Th international conference on computer and communications (ICCC)* (pp. 2170–2176). IEEE.

Hu, K., Gong, S., Zhang, Q., Chaowen, S., Xia, M., & Jiang, S. (2024). An overview of implementing security and privacy in federated learning. *Artificial Intelligence Review*, *57*(8), 204.

Huang, G., & Shu, T. (2024). Decentralized federated learning over noisy labels: A majority voting method. OpenReview preprint.

Ji, X., Zhu, Z., Xi, W., Gadyatskaya, O., Song, Z., Cai, Y., & Liu, Y. (2024). Fedfixer: Mitigating heterogeneous label noise in federated learning. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 12830–12838). (*vol. 38*).

Jia, N., Qu, Z., Ye, B., Wang, Y., Hu, S., & Guo, S. (2025). A comprehensive survey on communication-efficient federated learning in mobile edge environments. *IEEE Communications Surveys & Tutorials*.

Jiang, X., Li, J., Wu, N., Wu, Z., Li, X., Sun, S., Xu, G., Wang, Y., Li, Q., & Liu, M. (2025). Fnbench: Benchmarking robust federated learning against noisy labels. arXiv:2505.06684.

Jiang, X., Sun, S., Li, J., Xue, J., Li, R., Wu, Z., Xu, G., Wang, Y., & Liu, M. (2024). Tackling noisy clients in federated learning with end-to-end label correction. In *Proceedings of the 33rd ACM international conference on information and knowledge management* (pp. 1015–1026).

Jiang, X., Sun, S., Wang, Y., & Liu, M. (2022). Towards federated learning against noisy labels via local self-regularization. In *Proceedings of the 31st ACM international conference on information & knowledge management* (pp. 862–873).

Jiang, X., & Zhang, J. (2025). Fedclean: A general robust label noise correction for federated learning. In *Forty-second international conference on machine learning*.

Johnson, J. M., & Khoshgoftaar, T. M. (2022). A survey on classifying big data with label noise. *ACM Journal of Data and Information Quality*, *14*(4), 1–43.

Kang, R., Li, Q., & Lu, H. (2024). Federated machine learning in finance: A systematic review on technical architecture and financial applications. *Applied and Computational Engineering*, *102*, 61–72.

Kim, D., Oh, K., Lee, Y., & Woo, H. (2025). Overview of fair federated learning for fairness and privacy preservation. *Expert Systems with Applications*, (p. 128568).

Kim, T., Kim, D., & Yun, S.-Y. (2024). Revisiting early-learning regularization when federated learning meets noisy labels. arXiv:2402.05353.

Li, D., Qian, H., Li, Q., Tan, Z., Gan, Z., Wang, J., & Li, X. (2024a). Fedrgl: Robust federated graph learning for label noise. arXiv:2411.18905.

Li, J., Li, G., Cheng, H., Liao, Z., & Yu, Y. (2024b). Feddiv: Collaborative noise filtering for federated learning with noisy labels. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 3118–3126). (*vol. 38*).

Li, J., Pei, J., & Huang, H. (2022). Communication-efficient robust federated learning with noisy labels. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 914–924).

Li, M., Xu, P., Hu, J., Tang, Z., & Yang, G. (2025). From challenges and pitfalls to recommendations and opportunities: Implementing federated learning in healthcare. *Medical Image Analysis*, (p. 103497).

Li, M., & Zhu, C. (2024). Noisy label processing for classification: A survey. arXiv:2404.04159.

Li, Q., Duan, C., & Chen, S. (2023). Robust federated learning with parameter classification and weighted aggregation against noisy labels. In *Globecom 2023-2023 ieee global communications conference* (pp. 2445–2450). IEEE.

Liang, S., Huang, J., Hong, J., Zeng, D., Zhou, J., & Xu, Z. (2023). Fednoisy: Federated noisy label learning benchmark. arXiv:2306.11650.

Lu, Y., Chen, L., Zhang, Y., Zhang, Y., Han, B., Cheung, Y.-m., & Wang, H. (2024). Federated learning with extremely noisy clients via negative distillation. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 14184–14192). (*vol. 38*).

Ma, Y., Yang, B., Tang, Y., Zhan, W., & Yang, W. (2025). Harnessing feature distribution consistency for federated learning with noisy labels. In *2025 IEEE International conference on image processing (ICIP)* (pp. 1408–1413). IEEE.

Mao, W., Yu, B., Lv, Y., Xie, Y., & Zhang, C. (2025). Federated semi-supervised learning with contrastive representations against noisy labels. *Applied Soft Computing*, (p. 113421).

Mishra, R., & Gupta, H. P. (2025). Fed-NL: A federated learning approach to suppress noise in participant datasets to reduce communication rounds for convergence. *IEEE Transactions on Mobile Computing*.

Mishra, R., Gupta, H. P., & Dutta, T. (2022). Noise-resilient federated learning: Suppressing noisy labels in the local datasets of participants. In *Ieee infocom 2022-ieee conference on computer communications workshops (infocom wkshps)* (pp. 1–2). IEEE.

Morafah, M., Chang, H., Chen, C., & Lin, B. (2025). Federated learning client pruning for noisy labels. *ACM Transactions on Modeling and Performance Evaluation of Computing Systems*, *10*(2), 1–25.

Ouyang, C., Mao, J., Li, Y., Li, T., Zhu, D., Zhou, C., & Xu, Z. (2025). Federated learning for extreme label noise: Enhanced knowledge distillation and particle swarm optimization. *Electronics*, *14*(2), 366.

Pu, R., Yu, L., Zhan, S., Xu, G., Zhou, F., Ling, C. X., & Wang, B. (2025). FedELR: When federated learning meets learning with noisy labels. *Neural Networks*, *187*, 107275.

Rong, Q., Yuan, L., Li, G., Li, J., Zhang, L., & Ding, X. (2023). A static bi-dimensional sample selection for federated learning with label noise. In *International conference on database systems for advanced applications* (pp. 735–744). Springer.

Sabah, F., Chen, Y., Yang, Z., Azam, M., Ahmad, N., & Sarwar, R. (2024). Model optimization techniques in personalized federated learning: A survey. *Expert Systems with Applications*, *243*, 122874.

Shi, J., Zhang, K., Guo, C., Yang, Y., Xu, Y., & Wu, J. (2024). A survey of label-noise deep learning for medical image analysis. *Medical Image Analysis*, *95*, 103166.

Song, B., Zhao, S., Dang, L., Wang, H., & Xu, L. (2025). A survey on learning from data with label noise via deep neural networks. *Systems Science & Control Engineering*, *13*(1), 2488120.

Song, H., Kim, M., Park, D., Shin, Y., & Lee, J.-G. (2022). Learning from noisy labels with deep neural networks: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(11), 8135–8153.

Sun, W., Yan, R., Jin, R., Zhao, R., & Chen, Z. (2024). Curriculum-based federated learning for machine fault diagnosis with noisy labels. *IEEE Transactions on Industrial Informatics*, *20*(12), 13820–13830.

Tam, K., Li, L., Zhao, Y., & Xu, C. (2023). Fedcoop: Cooperative federated learning for noisy labels. In *Ecai 2023* (pp. 2298–2306). IOS Press.

Tian, Y., Yang, M., Zhou, Y., Wang, J., Ye, Q., Liu, T., Niu, G., & Lv, J. (2024). Learning locally, revising globally: Global reviser for federated learning with noisy labels. arXiv:2412.00452.

Tsouvalas, V., Saeed, A., Ozcelebi, T., & Meratnia, N. (2024). Labeling chaos to learning harmony: Federated learning with noisy labels. *ACM Transactions on Intelligent Systems and Technology*, *15*(2), 1–26.

Tuor, T., Wang, S., Ko, B. J., Liu, C., & Leung, K. K. (2021). Overcoming noisy and irrelevant data in federated learning. In *2020 25Th international conference on pattern recognition (ICPR)* (pp. 5020–5027). IEEE.

Wang, L., Bian, J., & Xu, J. (2024). Federated learning with instance-dependent noisy label. In *Icassp 2024-2024 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 8916–8920). IEEE.

Wang, Q., & Zhou, Y. (2022). FedSPL: Federated self-paced learning for privacy-preserving disease diagnosis. *Briefings in Bioinformatics*, *23*(1), bbab498.

Wang, Z., Zhou, T., Long, G., Han, B., & Jiang, J. (2022). Fednoil: A simple two-level sampling method for federated learning with noisy labels. arXiv:2205.10110.

Wu, N., Yu, L., Jiang, X., Cheng, K.-T., & Yan, Z. (2023). Fednoro: Towards noise-robust federated learning by addressing class imbalance and label noise heterogeneity. In *Proceedings of the thirty-second international joint conference on artificial intelligence* (pp. 4424–4432).

Wu, S., Zhang, G., Dai, F., Liu, B., & Dou, W. (2024). An edge-assisted federated contrastive learning method with local intrinsic dimensionality in noisy label environment. *Software: Practice and Experience*, *54*(9), 1793–1810.

Xu, J., Chen, Z., Quek, T. Q. S., & Chong, K. F. E. (2022). Fedcorr: Multi-stage federated learning for label noise correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10184–10193).

Xu, Y., Liao, Y., Wang, L., Xu, H., Jiang, Z., & Zhang, W. (2024). Overcoming noisy labels and non-iid data in edge federated learning. *IEEE Transactions on Mobile Computing*, *23*(12), 11406–11421.

Yang, K., Fan, W., Hu, M., & Li, J. (2025). Mitigating heterogeneous instance-dependent label noise in federated learning. In *International conference on intelligent computing* (pp. 456–468). Springer.

Yang, M., Qian, H., Wang, X., Zhou, Y., & Zhu, H. (2021). Client selection for federated learning with label noise. *IEEE Transactions on Vehicular Technology*, *71*(2), 2193–2197.

Yang, S., Park, H., Byun, J., & Kim, C. (2022). Robust federated learning with noisy labels. *IEEE Intelligent Systems*, *37*(2), 35–43.

Ye, M., Fang, X., Du, B., Yuen, P. C., & Tao, D. (2023). Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Computing Surveys*, *56*(3), 1–44.

Yu, S., Ahn, J.-H., & Kang, J. (2025). FedEFC: Federated learning using enhanced forward correction against noisy labels. arXiv:2504.05615.

Zeng, B., Yang, X., Chen, Y., Shen, Z., Yu, H., & Zhang, Y. (2024a). FedES: Federated early-stopping for hindering memorizing heterogeneous label noise. In *Proceedings of the thirty-third international joint conference on artificial intelligence* (pp. 5416–5424).

Zeng, B., Yang, X., Chen, Y., Yu, H., Hu, C., & Zhang, Y. (2024b). Federated data quality assessment approach: Robust learning with mixed label noise. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(12), 17620–17634.

Zhan, S., Yu, L., Chen, H., & Ji, T. (2025). FedLTF: Linear probing teaches fine-tuning to mitigate noisy labels in federated learning. In *The 16th asian conference on machine learning (conference track)*.

Zhang, H., Zhang, Y., Li, J., Liu, J., & Ji, L. (2025). A survey on learning with noisy labels in natural language processing: How to train models with label noise. *Engineering Applications of Artificial Intelligence*, *146*, 110157.

Zhou, X., & Wang, X. (2024). Federated label-noise learning with local diversity product regularization. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 17141–17149). (*vol. 38*).