This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Federated Active Learning Under Extreme Non-IID and Global Class Imbalance

Chen-Chen Zong, Sheng-Jun Huang*
Nanjing University of Aeronautics and Astronautics, China
{chencz,huangsj}@nuaa.edu.cn

Abstract

Federated active learning (FAL) seeks to reduce annotation cost under privacy constraints, yet its effectiveness degrades in realistic settings with severe global class imbalance and highly heterogeneous clients. We conduct a systematic study of query-model selection in FAL and uncover a central insight: the model that achieves more class-balanced sampling, especially for minority classes, consistently leads to better final performance. Moreover, global-model querying is beneficial only when the global distribution is highly imbalanced and client data are relatively homogeneous; otherwise, the local model is preferable. Based on these findings, we propose FairFAL, an adaptive class-fair FAL framework. FairFAL (1) infers global imbalance and local–global divergence via lightweight prediction discrepancy, enabling adaptive selection between global and local query models; (2) performs prototype-guided pseudo-labeling using global features to promote class-aware querying; and (3) applies a two-stage uncertainty–diversity balanced sampling strategy with k-center refinement. Experiments on five benchmarks show that FairFAL consistently outperforms state-of-the-art approaches under challenging long-tailed and non-IID settings. The code is available at https://github.com/chenchenzong/FairFAL.

1. Introduction

Federated learning (FL) has emerged as a central paradigm for collaborative model training without sharing raw data [27, 32, 37], enabling privacy preservation across decentralized ecosystems such as mobile devices, hospitals, and edge–cloud infrastructures. By localizing data storage and computation, FL addresses rising privacy and regulatory constraints while leveraging rich yet siloed data sources. Active learning (AL) aims to reduce annotation cost by selectively querying informative samples [22, 42, 44], and is especially valuable in domains where labeling is expensive, including medical imaging, autonomous driving, and industrial inspection. Integrating these two paradigms yields federated active learning (FAL) [8, 10, 28, 62], which enables decentralized clients to collaboratively identify valuable samples without revealing raw data. This combination is particularly compelling for real-world deployments, where data are inherently distributed, privacy-sensitive, and annotation budgets remain severely constrained.

Despite its promise, FAL remains under-explored in realistic settings characterized by severe global distribution imbalance and highly heterogeneous clients—conditions that frequently arise in practice. Although recent studies [8, 28, 62] have begun to consider extreme non-independent and identically distributed (non-IID) scenarios, they often treat heterogeneity merely as a data-partitioning issue rather than a global class imbalance problem. As a result, most methods implicitly assume that the overall label distribution is relatively balanced or can be recovered through aggregation. However, real-world federated systems commonly exhibit long-tailed global distributions, where rare but critical classes appear sparsely across clients. Under such regimes, existing acquisition strategies struggle to surface minority classes, leading to systematic bias toward head categories and inefficient use of annotation budgets.

To address these limitations, we investigate a fundamental question: in FAL, two query-selection models naturally exist–the globally aggregated global model and the independently trained local model. **Which model enables sampling that better preserves class balance, particularly for rare classes, and how does such balance influence the performance of the final global model?**

Our empirical analysis yields a consistent insight: across varying levels of global imbalance and client heterogeneity, the model that achieves more class-balanced sampling reliably delivers superior final performance. Furthermore, when global class imbalance is severe but client distributions are relatively homogeneous, global-model–based sampling proves more effective; in all other regimes, the local model generally becomes the preferred choice.

Building on these insights, we introduce FairFAL, an adaptive, class-**fair FAL** strategy designed to handle extreme non-IID data and globally long-tailed class distribu-

*Corresponding Author.

tions. FairFAL incorporates three key components. First, we exploit the prediction discrepancies between the global and local models on each client to estimate both the severity of global class imbalance and the degree of local heterogeneity, enabling an adaptive model-selection mechanism that chooses the more suitable query selector per client. Second, motivated by the observation that the global model consistently provides stronger and more coherent representations, we compute per-class prototypes using global features and perform prototype-guided pseudo-labeling to enforce class-aware querying. Third, to reduce redundancy and promote coverage, we form class-specific candidate pools and apply k-center selection in a gradient-embedding space that combines global features with model predictions, ensuring that the final queries are both informative and diverse. Experiments on five benchmark datasets show that FairFAL consistently outperforms state-of-the-art FAL baselines under challenging federated settings.

2. Related Work

Federated learning (FL) enables multiple clients to collaboratively train a shared model without exchanging raw data. FedAvg [37] established the standard framework, but its performance degrades significantly under statistical heterogeneity (non-IID data). To mitigate this, prior work has explored three main directions. Regularization-based approaches [15, 23, 27, 32] constrain local updates or feature representations to reduce client drift. Aggregation-based methods [25, 35, 41, 60] adjust client contributions based on divergence or data quality. Personalized FL [11, 34, 49, 56] further decouples global and client-specific components to improve adaptability across heterogeneous distributions. Beyond the standard setting, FL has also been extended to semi-supervised [5, 33], multi-task [36, 46], continual [45, 58], and domain-adaptive [24, 54] learning, highlighting its versatility in real-world applications.

Active learning (AL) aims to reduce annotation cost by selectively querying the most informative samples from an unlabeled pool. Existing methods fall into three main categories. Uncertainty-based approaches [6, 13, 19, 31, 39, 57] focus more on ambiguous samples near the decision boundary, typically measured through entropy [19] or margin [6]. Diversity-based strategies [9, 30, 38, 43, 47] such as Coreset [43] focus on selecting representative points that cover the feature space. Hybrid methods [3, 40, 48, 59, 61] integrate both, with BADGE [3] being a well-known example of two-stage uncertainty–diversity balanced sampling. Recent developments extend AL to low-budget [4, 16], semi-supervised [14, 21], and cross-domain [50, 59] scenarios.

Federated active learning (FAL) integrates AL into FL to enable label-efficient model training under privacy constraints. Early studies [1, 2, 52] directly applied centralized AL strategies using either a global or a local model, but such single-model designs degrade quickly under strong client heterogeneity. Recent methods [8, 10, 28, 62] therefore adopt hybrid designs that combine global and local signals. For example, LoGo [28] performs local clustering followed by global uncertainty scoring, while KAFAL [8] and IFAL [62] exploit prediction discrepancies between global and local models to guide acquisition.

3. Preliminary

Federated learning (FL) enables a federation of clients to collaboratively train a shared model without exchanging raw data. Let K clients hold local datasets $\{\mathcal{D}_1, \dots, \mathcal{D}_K\}$, where each sample (x_i, y_i) consists of an input $x_i \in \mathcal{X}$ and a label $y_i \in \mathcal{C} = \{1, \dots, C\}$ for a C-class task. Denote the global model parameters by θ^g and the local parameters on client k by θ_k. The training objective aggregates client-level losses in a weighted manner:

$$\min_{\theta^g} \sum_{k=1}^{K} \frac{n_k}{N} \mathcal{L}_k(\theta_k), \quad n_k = |\mathcal{D}_k|, \quad N = \sum_{k=1}^{K} n_k. \tag{1}$$

In each communication round, client k trains θ_k on its private data and sends the resulting model update to the server. The server aggregates all client updates, typically via FedAvg [37], to obtain the new global parameters θ^g. This paradigm enables privacy-preserving learning across heterogeneous data silos while avoiding direct data sharing.

Active learning (AL) aims to minimize annotation cost by iteratively querying the most informative unlabeled samples for labeling. Given a labeled set $\mathcal{D}_L = \{(x_i, y_i)\}$ and an unlabeled pool $\mathcal{D}_U = \{x_j\}$, a model with parameters θ is iteratively trained on $\mathcal{D}_L$ and improved through selective acquisition. At each AL round, the learner selects a query set $\mathcal{Q} \subset \mathcal{D}_U$ using an acquisition function $\mathcal{A}(\cdot)$ that scores the informativeness of each unlabeled instance. The queried samples are labeled and added to $\mathcal{D}_L$, while the unlabeled pool is updated as $\mathcal{D}_U \leftarrow \mathcal{D}_U \setminus \mathcal{Q}$. This iterative procedure continues until the annotation budget is exhausted, enabling label-efficient model training.

Federated active learning (FAL) integrates the privacy-preserving collaboration of FL with the label-efficiency of AL, enabling decentralized clients to identify globally informative samples without sharing raw data. Unlike conventional AL, which operates with a single centralized model, FAL performs a full federated training process in every AL round. Formally, consider K clients, each initialized with a labeled set $\mathcal{D}_L^{(k)}$ and an unlabeled pool $\mathcal{D}_U^{(k)}$. At each AL round t, each client trains its local model on $\mathcal{D}_L^{(k)}|_t$ for several communication rounds, after which the server aggregates the local updates to obtain the new global model. Once federated training completes, each client scores its unlabeled pool $\mathcal{D}_U^{(k)}|_t$ using an acquisition function $\mathcal{A}^{(k)}(\cdot)$,

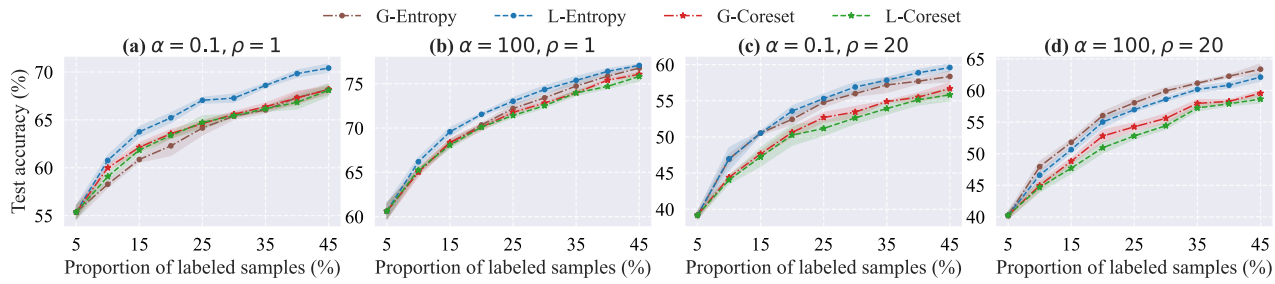


Figure 1. Comparison of global (G) and local (L) query selectors under varying (α, ρ) configurations on CIFAR-10. Test accuracy is plotted against the proportion of labeled samples for Entropy and Coreset sampling.

selects a query set $\mathcal{Q}^{(k)}|_t \subset \mathcal{D}_U^{(k)}|_t$, and acquires their labels. The labeled and unlabeled sets are updated via $\mathcal{D}_L^{(k)}|_{t+1} \leftarrow \mathcal{D}_L^{(k)}|_t \cup \mathcal{Q}^{(k)}|_t$ and $\mathcal{D}_U^{(k)}|_{t+1} \leftarrow \mathcal{D}_U^{(k)}|_t \setminus \mathcal{Q}^{(k)}|_t$. This process alternates between federated training and active querying until the annotation budget is exhausted, gradually improving the global model with minimal supervision.

4. Observation

We begin by investigating how global and local models perform as query selectors in FAL, and how their effectiveness is related to the class balance of the queried data under varying levels of local heterogeneity (α) and global class imbalance (ρ). Here, α denotes the concentration parameter of the widely used Dirichlet partition [27, 32, 37], where smaller values correspond to stronger cross-client distribution shifts. The global imbalance ratio ρ measures the long-tailed severity of the aggregated dataset [7, 26], with larger values indicating more skewed class frequencies.

Specifically, we conduct experiments on CIFAR-10 [29] with $\alpha \in \{0.1, 100\}$ and $\rho \in \{1, 20\}$[1], evaluating both global and local models as query selectors using two representative AL strategies: entropy-based sampling [44] and coreset sampling [43], representing uncertainty- and diversity-based querying, respectively. At each AL round, 5% of the training data is queried for labeling. We then evaluate the results from both quantitative and qualitative perspectives:

(1) Mean accuracy curves. Figure 1 reports the mean test accuracy over $S=5$ random seeds. These curves summarize the learning trajectory of each method, revealing its stability, convergence behavior, and whether it consistently maintains stronger performance throughout the AL process.

(2) Robust paired analysis. To obtain a more aggregated and statistically grounded comparison, we compute the area under the learning curve (AULC) for each seed and conduct a paired analysis between any two methods i and j (Table 1). For seed s, let $\Delta_s^{ij} = \text{AULC}_s^i - \text{AULC}_s^j$. We evaluate the set $\{\Delta_s^{ij}\}_{s=1}^S$ using three complementary statistics:

- **Positive ratio** (π^+) [12] measures cross-seed consistency: $\pi_{ij}^+ = \frac{1}{S}\sum_{s=1}^S \mathbb{1}(\Delta_s^{ij} > 0)$. Larger values (e.g., ≥ 0.8) indicate that i consistently outperforms j.
- **One-sided Wilcoxon p-value** [51] evaluates whether i significantly outperforms j. Small $p(\leq 0.05)$ provides strong evidence favoring i, $0.05 < p \leq 0.10$ indicates a weak trend, and $p > 0.10$ suggests no difference.
- **Hodges–Lehmann (HL) estimator** [18] provides a robust estimate of the typical performance gain:

$$\widehat{\Delta}_{HL}^{ij} = \text{median}\left\{\frac{\Delta_s^{ij} + \Delta_t^{ij}}{2} : 1 \leq s \leq t \leq S\right\}.$$

Compared with the mean, HL is less sensitive to outliers and better reflects the gain in percentage points.

(α, ρ) \ Method	**Entropy**	**Coreset**
	$(winner, \pi_{ij}^+, p, \widehat{\Delta}_{HL}^{ij})$	
$(0.1, 1)$	$(L, 1.0, 0.03, 201\text{ pp})$	$(G, 1.0, 0.03, 28\text{ pp})$
$(100, 1)$	$(L, 1.0, 0.03, 50\text{ pp})$	$(G, 0.8, 0.06, 21\text{ pp})$
$(0.1, 20)$	$(L, 1.0, 0.03, 66\text{ pp})$	$(G, 1.0, 0.03, 50\text{ pp})$
$(100, 20)$	$(G, 1.0, 0.03, 106\text{ pp})$	$(G, 1.0, 0.03, 92\text{ pp})$

Table 1. Pairwise comparison between global (G) and local (L) query selectors under different (α, ρ) configurations on CIFAR-10. Each entry reports the winner (strategy i), positive ratio (π_{ij}^+), one-sided Wilcoxon p-value, and Hodges–Lehmann effect size $\widehat{\Delta}_{HL}^{ij}$ (in percentage points, pp) for Entropy and Coreset sampling.

Observation 1. *For uncertainty-based sampling, the local model generally outperforms the global model, except in cases where global class imbalance is severe while client distributions remain relatively homogeneous.*

As presented in Figure 1 and Table 1, the local model consistently surpasses the global model under uncertainty-based querying across most (α, ρ) settings. This trend is most pronounced when the global distribution is relatively balanced (small ρ) and client data are highly heterogeneous (small α). In this regime, although each client only considers its own data, the aggregation of diverse client-specific queries naturally produces a globally balanced query set.

[1] Local data distribution visualizations are shown in Appendix Sec. A.

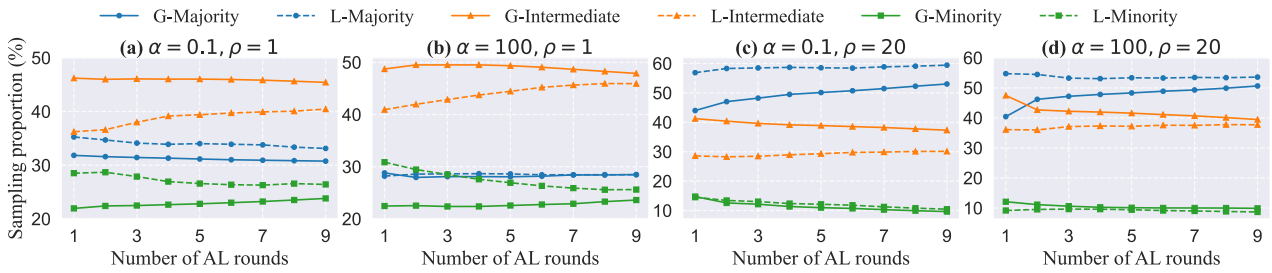


Figure 2. Cumulative sampling proportions of majority, intermediate, and minority classes for global (G) and local (L) query selectors under different (α, ρ) settings on CIFAR-10, where the three groups denote the top 3, middle 4, and bottom 3 classes by sample proportion.

The global model, in contrast, tends to yield smoother and less discriminative uncertainty estimates due to cross-client averaging, making it less sensitive to subtle client-specific decision boundaries. However, when global imbalance becomes severe (large ρ) and client distributions are nearly homogeneous (large α), the diversity effect vanishes: local-model querying increasingly reflects the global long-tailed skew, leading to imbalanced acquisitions. In such cases, the global model gains a noticeable advantage by exploiting shared knowledge across clients to counteract class skew during sampling. These observations are further corroborated in Figure 2, which visualizes how the cumulative sampling ratios of majority, intermediate, and minority classes evolve under different (α, ρ) configurations.

Observation 2. *The relative advantage of the local and global models is largely determined by which one achieves more class-balanced sampling—especially in acquiring minority classes—which closely aligns with final performance.*

Figures 1 and 2 reveal a strong, consistent alignment between the balance of queried samples and downstream accuracy: models that acquire more minority-class instances invariably achieve higher performance. Under severe global class imbalance (large ρ), minority acquisition becomes the dominant factor. In Figure 2 (c), the global and local models obtain nearly the same number of minority samples in the initial AL rounds, yielding almost identical accuracy curves in Figure 1 (c). As the gap in minority sampling grows and eventually stabilizes, so does the performance gap, suggesting a direct correspondence. When the global distribution is balanced (small ρ), however, minority sampling alone no longer explains performance differences. In Figures 2 (a,b), both models acquire similar proportions of minority samples, yet the local model achieves clearly higher accuracy in Figure 1 (a). This is because the local model's queried set exhibits a more balanced class distribution, not only among minority classes but across the entire label space.

Observation 3. *For diversity-based sampling, the global model consistently outperforms the local model.*

Figures 1 and Table 1 show that, across all settings, the global model maintains a clear and consistent advantage when diversity-based strategies (i.e., Coreset sampling) are used. Unlike uncertainty-driven querying, where performance is tied to the predictive scores, diversity-based sampling primarily relies on the structure of the feature space. Its effectiveness depends on how well feature embeddings capture global coverage, inter-sample separation, and manifold geometry. Since the global model is optimized using aggregated updates from all clients, it develops more discriminative and globally aligned feature representations. These richer embeddings enable more faithful estimation of pairwise distances, allowing the global model to identify samples that better span the underlying data manifold.

5. Methodology

Based on the observations, we find that class-balanced sampling, especially for rare classes, is crucial for effective FAL. To this end, we propose FairFAL, an adaptive class-**Fair FAL** framework. First, guided by **Observation 1**, we design an *adaptive model-selection* mechanism that estimates both the severity of global class imbalance and the discrepancy between local and global class distributions on each client. This enables FairFAL to dynamically choose the most suitable query model for each client, either global or local. Second, following **Observation 2**, we promote class-balanced querying through a *prototype-guided pseudo-labeling* strategy that performs class-level sample selection. In particular, motivated by **Observation 3**, we exploit the global model's well-generalized feature representations to compute per-class prototypes from labeled data on each client. The similarity between unlabeled samples and these prototypes is used to assign pseudo-labels, thereby guiding class-aware sample acquisition. Finally, to further enhance query diversity, we introduce a *two-stage balanced sampling* mechanism that jointly considers uncertainty and diversity. Concretely, we first construct a candidate pool by selecting high-uncertainty samples within each class, and then perform k-center sampling in a gradient-embedding space to obtain the final, diverse query set.

5.1. Adaptive Model-Selection

To adaptively decide whether the global or local model should act as the query selector, without introducing any additional privacy leakage, we estimate two key quantities on each client: (1) the severity of global class imbalance, and (2) the divergence between local and global data distributions. These estimates, obtained solely from locally available labeled data and aggregated statistics, are then mapped to a continuous model-selection score that operationalizes the empirical observations in Section 4.

Consider client k with labeled set $\mathcal{D}_L^{(k)}$ and unlabeled pool $\mathcal{D}_U^{(k)}$. In the first AL round, labeled samples are typically acquired through random querying from $\mathcal{D}_U^{(k)}$, making $\mathcal{D}_L^{(k)}$ a reasonable IID approximation of the client's underlying data distribution. We leverage this property to derive a privacy-preserving estimate of the global long-tailedness, which will later guide the adaptive model-selection process.

For client k, let $n_{k,c}$ denote the number of labeled samples of class c, and define $n_k^{\max} = \max_{c:n_{k,c}>0} n_{k,c}$. We construct a class-balanced subset

$$\mathcal{B}^{(k)} = \bigcup_{c\in\mathcal{C}_k^+} \mathcal{B}_c^{(k)}, \quad \mathcal{B}_c^{(k)} \subseteq \mathcal{D}_L^{(k)}, \quad |\mathcal{B}_c^{(k)}| = n_k^{\max}, \tag{2}$$

where $\mathcal{C}_k^+ = \{c \in \mathcal{C} : n_{k,c} > 0\}$ and each $\mathcal{B}_c^{(k)}$ is formed by sampling (with replacement if necessary) from the labeled instances of class c. This upsampling step normalizes the per-class sample counts and ensures that every observed class contributes equally to the subsequent estimation.

Global class imbalance ratio estimation. Let f^g denote the current global model with parameters θ^g, and let $\mathbf{p}^g(x) = f^g(x)$ be its predictive probability vector. Using the class-balanced set $\mathcal{B}^{(k)}$, each client k forms a local estimate of the global predictive prior:

$$\widehat{\boldsymbol{\pi}}_g^{(k)} = \frac{1}{|\mathcal{B}^{(k)}|} \sum_{x\in\mathcal{B}^{(k)}} \mathbf{p}^g(x). \tag{3}$$

We estimate the global class imbalance ratio on client k via

$$\gamma_k = \frac{\min_{c\in\mathcal{C}_k^+} \widehat{\pi}_{g,c}}{\max_{c\in\mathcal{C}_k^+} \widehat{\pi}_{g,c}} \in (0, 1], \tag{4}$$

where values closer to 1 indicate a more balanced distribution, and smaller values reflect stronger long-tailed skew.

To ensure reliable yet privacy-preserving estimation, each client uploads only the scalar γ_k to the server. The server aggregates these values via simple averaging, $\bar{\gamma} = \frac{1}{K}\sum_{k=1}^K \gamma_k$, and broadcasts $\bar{\gamma}$ to all clients as the global class-imbalance coefficient. Since labeled data in later AL rounds are no longer obtained through random sampling, $\bar{\gamma}$ is computed only in the first round and kept fixed thereafter.

Local–global distributional divergence estimation. To characterize how heterogeneous each client is with respect to the global distribution, we further estimate the discrepancy between the local and global predictive priors. Let $f^{(k)}$ denote the locally trained model on client k with predictive vector $\mathbf{p}^{(k)}(x) = f^{(k)}(x)$. Using the same balanced set $\mathcal{B}^{(k)}$, client k computes a local predictive prior $\widehat{\boldsymbol{\pi}}_\ell^{(k)}$.

We quantify the divergence between the global and local predictive priors by the normalized symmetric difference:

$$d_k = \frac{1}{C}\sum_{c=1}^{C} \frac{\left|\widehat{\pi}_{g,c} - \widehat{\pi}_{\ell,c}^{(k)}\right|}{\widehat{\pi}_{g,c} + \widehat{\pi}_{\ell,c}^{(k)}}, \tag{5}$$

which satisfies $d_k \in [0, 1]$ by construction. Larger values indicate stronger deviation of client k from the global predictive distribution, hence higher statistical heterogeneity.

Unlike the global class-balance coefficient $\bar{\gamma}$ estimated only once, d_k is recomputed at every AL round to capture the evolving discrepancy between local and global labeled distributions as querying progresses.

Continuous model-selection score. Section 4 indicates that the global model is more suitable when the global distribution is highly imbalanced and clients are relatively homogeneous, whereas local models dominate in balanced or highly heterogeneous regimes.

To capture this dependency in a smooth and data-driven manner, we define a continuous model-selection score $s_k \in [0, 1)$ for each client k, based on the global balance coefficient $\bar{\gamma}$ and the local–global divergence d_k:

$$s_k = 1 - \frac{1}{2}\big(d_k + \bar{\gamma}\big). \tag{6}$$

This score naturally recovers the four regimes discussed in Section 4. Larger s_k values occur when the global distribution is highly imbalanced (small $\bar{\gamma}$) and the local distribution is well aligned with the global one (small d_k), corresponding to the regime where the global model is preferred. In contrast, a more balanced global distribution or stronger local heterogeneity leads to smaller scores, favoring the local model. We then adopt a simple decision rule:

$$\mathcal{M}^{(k)} = \begin{cases} \text{global model}, & s_k > \delta, \\ \text{local model}, & s_k \le \delta, \end{cases} \tag{7}$$

where δ is a fixed threshold (set to 0.75). This mechanism is lightweight and fully privacy-preserving, relying only on aggregated statistics derived from locally labeled data.

5.2. Prototype-Guided Pseudo-Labeling

As discussed in Section 4, achieving class-balanced sampling is essential for improving FAL performance. While adaptive model selection determines which model should

guide querying, it alone cannot guarantee balanced acquisition when class distributions are highly skewed. We therefore introduce a prototype-guided pseudo-labeling mechanism that explicitly promotes class-level balance.

In long-tailed or class-imbalanced settings, the classifier trained on imbalanced data tends to learn biased decision boundaries, making unlabeled samples more likely to be assigned to head classes. To counteract this bias, we replace direct classifier predictions with a prototype-based assignment: each class is represented by a feature prototype, and pseudo labels are determined by similarity between sample embeddings and these prototypes.

Motivated by **Observation 3**, which demonstrates that the global model provides more discriminative and well-generalized representations than local models, we adopt the global model as the feature extractor for prototype construction. This choice ensures that the resulting prototypes better reflect the global data manifold and yield more reliable class-wise pseudo labels.

Let $\phi^g(\cdot)$ denote the penultimate-layer feature extractor of the global model. On client k, given the labeled set $\mathcal{D}_L^{(k)}$, we first extract ℓ_2-normalized features:

$$\mathbf{z}_i^{(k)} = \frac{\phi^g(x_i)}{\|\phi^g(x_i)\|_2}, \quad (x_i, y_i) \in \mathcal{D}_L^{(k)}. \tag{8}$$

For each class $c \in \mathcal{C}_k^+$, we compute a class prototype by averaging the normalized features of its labeled samples:

$$\boldsymbol{\mu}_c^{(k)} = \frac{1}{|\mathcal{D}_{L,c}^{(k)}|} \sum_{(x_i,y_i)\in\mathcal{D}_L^{(k)},\, y_i=c} \mathbf{z}_i^{(k)}, \tag{9}$$

where $\mathcal{D}_{L,c}^{(k)}$ denotes the subset of labeled instances of class c at client k. These prototypes provide a compact, noise-reduced summary of each class in the representation space.

For any unlabeled sample $x \in \mathcal{D}_U^{(k)}$, we extract its normalized feature $\mathbf{z}^{(k)}(x)$ and compute cosine similarity to all available prototypes:

$$s_c^{(k)}(x) = \left\langle \mathbf{z}^{(k)}(x),\, \boldsymbol{\mu}_c^{(k)} \right\rangle, \quad c \in \mathcal{C}_k^+. \tag{10}$$

Then, a prototype-guided pseudo-label is assigned via

$$\widehat{y}^{(k)}(x) = \arg\max_{c\in\mathcal{C}_k^+} s_c^{(k)}(x), \tag{11}$$

which leverages the geometry of the global feature space rather than raw classifier logits, offering improved robustness under long-tailed and heterogeneous federation.

The resulting pseudo labels induce a partition of the unlabeled pool into class-specific subsets:

$$\widetilde{\mathcal{D}}_{U,c}^{(k)} = \left\{ x \in \mathcal{D}_U^{(k)} \,\middle|\, \widehat{y}^{(k)}(x) = c \right\}, \quad c \in \mathcal{C}_k^+. \tag{12}$$

These subsets form the basis of the class-aware sampling mechanism in the subsequent stage.

5.3. Uncertainty–Diversity Balanced Sampling

While class-wise sampling promotes balanced label acquisition, directly selecting the most uncertain samples within each class often results in severe redundancy: highly uncertain examples tend to cluster in a narrow region of the feature space, wasting annotation budget on near-duplicates. To address this, we introduce a two-stage sampling strategy that jointly enforces uncertainty and diversity.

Stage 1: Class-wise candidate selection. For client k, let $b_c^{(k)}$ denote the uniformly allocated query budget for class c, with $\sum_c b_c^{(k)} = B_k$. Using the adaptively selected query model $\mathcal{M}^{(k)}$, we compute an uncertainty score (e.g., entropy) $u^{(k)}(x)$ for each $x \in \widetilde{\mathcal{D}}_{U,c}^{(k)}$. We then construct an over-complete candidate pool:

$$\mathcal{H}_c^{(k)} = \mathrm{TopK}\Big(\widetilde{\mathcal{D}}_{U,c}^{(k)},\; u^{(k)}(x),\; \kappa\, b_c^{(k)}\Big), \quad \kappa > 1, \tag{13}$$

which retains the $\kappa b_c^{(k)}$ most uncertain samples for class c. This stage preserves informativeness while maintaining sufficient redundancy for diversity refinement.

Stage 2: Diversity sampling. To reduce redundancy within each $\mathcal{H}_c^{(k)}$, we refine candidates in a global gradient-embedding space consistently using the global model.

We compute the gradient embedding for any x as

$$\mathbf{g}^{(k)}(x) = \psi\big(x; \phi^g, f^g\big) \in \mathbb{R}^d, \tag{14}$$

where $\psi(\cdot)$ denotes the gradient of the classification loss with respect to the global classifier parameters, using global features $\phi^g(x)$. This gradient embedding captures both semantic structure and class-discriminative geometry, providing a stable and globally aligned space for enforcing diversity—particularly beneficial for rare classes.

For each class c, we define the embedding sets for labeled (anchors) and candidate samples as:

$$\mathcal{A}_c^{(k)} = \{\mathbf{g}^{(k)}(x) : (x, y) \in \mathcal{D}_L^{(k)},\, y = c\}, \tag{15}$$

$$\mathcal{G}_c^{(k)} = \{\mathbf{g}^{(k)}(x) : x \in \mathcal{H}_c^{(k)}\}. \tag{16}$$

Treating $\mathcal{A}_c^{(k)}$ as fixed anchors, we select $b_c^{(k)}$ additional points from $\mathcal{H}_c^{(k)}$ such that the combined set achieves good coverage in the embedding space. This is formulated as a class-wise k-center objective:

$$\mathcal{Q}_c^{(k)} = \arg\min_{\substack{\mathcal{Q}'\subset\mathcal{H}_c^{(k)} \\ |\mathcal{Q}'|=b_c^{(k)}}} \left\{ \max_{x\in\mathcal{H}_c^{(k)}} \min_{a\in\mathcal{A}_c^{(k)}\cup\mathcal{Q}'} d\big(\mathbf{g}^{(k)}(x), \mathbf{g}^{(k)}(a)\big) \right\}. \tag{17}$$

However, Eq. (17) is NP-hard [43], so we adopt the greedy k-center algorithm [43] to get an efficient 2-approximation.

Final query set. The final query set for client k is the union of class-wise selections:

$$\mathcal{Q}^{(k)} = \bigcup_{c\in\mathcal{C}_k^+} \mathcal{Q}_c^{(k)}. \tag{18}$$

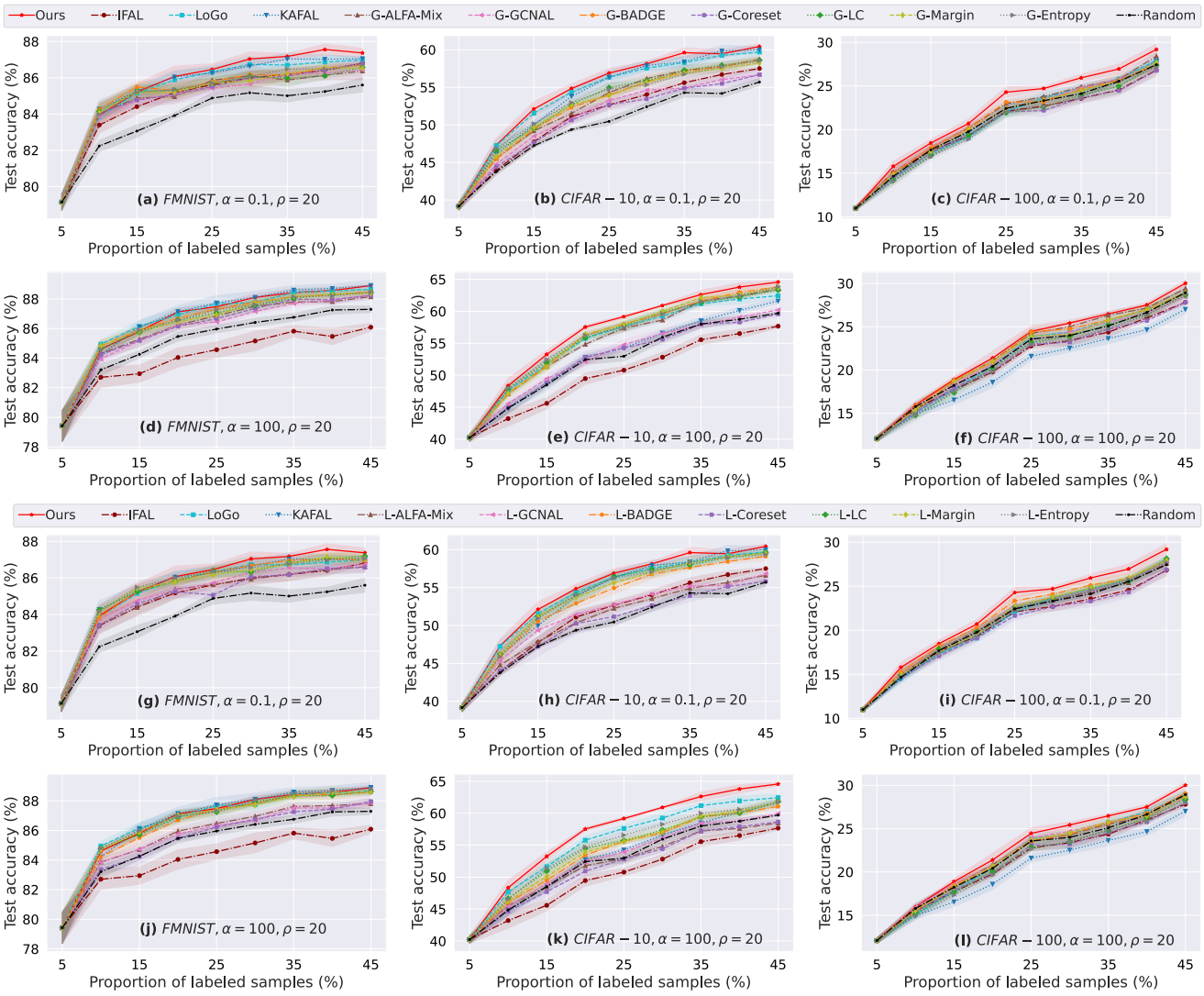


Figure 3. Test accuracy vs. proportion of labeled samples on FMNIST, CIFAR-10, and CIFAR-100 under global imbalance $\rho = 20$ and two heterogeneity levels ($\alpha = 0.1$ and $\alpha = 100$). Each subplot corresponds to a dataset–heterogeneity setting, and curves compare FairFAL with FAL and AL baselines using global (G-) or local (L-) models. Numerical results are provided in Tables 5 and 6 in the Appendix.

Overall, the two-stage procedure first identifies informative samples based on uncertainty and then promotes representative coverage through diversity under class-specific budget constraints. This yields query sets that are simultaneously informative, diverse, and class-balanced, enabling more sample-efficient and fairer FAL.

6. Experiments

Datasets. We evaluate FairFAL on five benchmarks spanning natural and medical image classification. The natural-image datasets include FMNIST [53], CIFAR-10 [29], and CIFAR-100 [29], while the medical benchmarks comprise OctMNIST and DermaMNIST from MedMNIST [55]. To highlight the effect of global class imbalance, we set the overall imbalance ratio to ρ=20 in all experiments. Data are then partitioned across 10 clients using a Dirichlet distribution with α=0.1 and 100, representing highly heterogeneous and nearly homogeneous client splits, respectively.

Baselines. We compare FairFAL against 11 query strategies grouped into five categories: (1) Random sampling; (2) Uncertainty-based methods, including Entropy [19], Margin [6], and Least Confidence (LC) [31]; (3) Diversity-based method Coreset [43]; (4) Hybrid approaches such as BADGE [3], GCNAL [9], and ALFA-Mix [39], which jointly leverage uncertainty and diversity cues; (5) FAL methods, including KAFAL [8], LoGo [28], and IFAL [62].

We provide training configurations, dataset details, and additional experimental results in Appendix Sec. B.

Component / (α, ρ)	w/ Model selection		w/ Class-wise sampling		w/ Two stage balanced sampling			
	$\mathcal{M}^{(k)}$	$\tilde{\mathcal{M}}^{(k)}$	Local	Global	$\kappa = 2$	$\kappa = 3$	$\kappa = 4$ **(Final)**	$\kappa = 5$
$(0.1, 20)$	59.33	58.49	59.14	59.95	60.61	60.38	60.44	60.28
$(100, 20)$	63.65	61.89	63.39	64.02	64.60	64.58	64.57	64.17

Table 2. Ablation results of FairFAL under different (α, ρ) configurations. We evaluate three key components: adaptive model selection, prototype-guided class-wise sampling, and the two-stage uncertainty–diversity balanced sampling strategy. Modules are added sequentially from left to right, and the final configuration corresponds to κ=4. All results are reported in terms of final-round test accuracy (%).

	Random	KAFAL	LoGo	IFAL	**Ours**
OctMNIST	68.30	70.40	70.00	68.40	**72.80**
DermaMNIST	72.32	73.27	73.62	72.97	**73.77**

Table 3. Final-round test accuracy comparison on OctMNIST and DermaMNIST under extreme client heterogeneity α=0.1.

6.1. Performance Comparison

Figure 3 compares test accuracy across AL rounds on FMNIST, CIFAR-10, and CIFAR-100 under severe global class imbalance and both heterogeneous and homogeneous client distributions. FairFAL consistently outperforms all baselines, with its advantage becoming more pronounced as task difficulty increases. The widening gap from FMNIST to CIFAR-100 highlights its robustness and generalization in challenging federated settings. In contrast, the gains of prior AL and FAL methods over Random sampling diminish as the task becomes harder, especially under nearly homogeneous client distributions (α=100). Under high heterogeneity (α=0.1), client-specific distributions naturally inject diversity into the global query set, allowing many baselines to perform reasonably well. However, when this implicit diversity weakens, methods such as IFAL, which lack explicit class-balancing mechanisms, tend to query redundant or biased samples, causing their performance to approach or even fall below that of random sampling. This further highlights the necessity and effectiveness of FairFAL's strategy design, where adaptive model selection works jointly with class-balanced, diversity-aware sampling.

Table 3 reports results on two naturally imbalanced medical benchmarks, OctMNIST and DermaMNIST, under severe client heterogeneity (α=0.1). Unlike the earlier curated datasets, both benchmarks exhibit intrinsic long-tailed label distributions, posing a substantially more challenging setting for class-fair querying. Despite this difficulty, FairFAL consistently achieves the highest accuracy. These results further underscore the robustness of FairFAL in highly imbalanced, real-world clinical scenarios.

6.2. Ablation Studies

Table 2 summarizes the ablation results under two representative configurations, where the components of FairFAL are incrementally enabled from left to right.

Effectiveness of adaptive model selection. We denote by $\mathcal{M}^{(k)}$ the model (global or local) chosen by our adaptive rule on client k, and by $\widetilde{\mathcal{M}}^{(k)}$ its complementary option, i.e., the model not selected. Across both heterogeneity levels, using $\mathcal{M}^{(k)}$ consistently leads to higher accuracy, confirming that adaptively selecting the appropriate query model is crucial for stabilizing querying behavior in FAL.

Role of global vs. local prototypes. To assess the effect of prototype quality on class-wise sampling, we compare prototypes constructed using features from the local and global models. In both cases, prototypes based on global features perform better, suggesting that globally aggregated representations yield cleaner class separation and more reliable similarity estimates, thereby enabling more accurate pseudo-labels and more balanced class-aware acquisition.

Effect of candidate-pool size. We then ablate the oversampling ratio κ used to construct each class-specific candidate pool. The two-stage strategy consistently outperforms its one-stage counterparts across different choices of κ. Although κ=2 achieves the highest accuracy by a small margin, the differences among κ=2, 3, and 4 are minor, suggesting that the method is relatively insensitive to the exact choice of κ. We therefore adopt κ=4 as FairFAL's final configuration, as it offers a larger and more flexible candidate pool for subsequent diversity refinement.

Additional Appendix results further verify FairFAL's robustness across different FL frameworks, client numbers, model architectures, uncertainty measures, and δ values.

7. Conclusion

In this paper, we revisited FAL under realistic challenges from global long-tailed distributions and client heterogeneity. Our systematic analysis shows that class-balanced sampling is a key factor in FAL performance, and that the choice between global and local query selectors depends jointly on global imbalance and local–global divergence. Building on these findings, we propose FairFAL, a class-fair FAL framework that integrates adaptive model selection, prototype-guided pseudo-labeling, and uncertainty–diversity balanced sampling. FairFAL consistently improves sample efficiency and robustness across diverse datasets and settings, showing strong effectiveness in extreme non-IID and severely imbalanced federated scenarios.

Acknowledgements

This work was supported in part by the NSFC (U2441285).

References

[1] Lulwa Ahmed, Kashif Ahmad, Naina Said, Basheer Qolomany, Junaid Qadir, and Ala Al-Fuqaha. Active learning based federated learning for waste and natural disaster image classification. *IEEE Access*, 8:208518–208531, 2020. 2

[2] Jin-Hyun Ahn, Yeeun Ma, Seoyun Park, and Cheolwoo You. Federated active learning (f-al): an efficient annotation strategy for federated learning. *IEEE Access*, 12:39261–39269, 2024. 2

[3] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019. 2, 7, 3, 4

[4] Wonho Bae, Junhyug Noh, and Danica J Sutherland. Generalized coverage for more robust low-budget active learning. In *European Conference on Computer Vision*, pages 318–334. Springer, 2024. 2

[5] Sikai Bai, Shuaicheng Li, Weiming Zhuang, Jie Zhang, Kunlin Yang, Jun Hou, Shuai Yi, Shuai Zhang, and Junyu Gao. Combating data imbalances in federated semi-supervised learning with dual regulators. In *Proceedings of the AAAI conference on artificial intelligence*, pages 10989–10997, 2024. 2

[6] Maria-Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In *International Conference on Computational Learning Theory*, pages 35–50. Springer, 2007. 2, 7, 3, 4

[7] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018. 3

[8] Yu-Tong Cao, Ye Shi, Baosheng Yu, Jingya Wang, and Dacheng Tao. Knowledge-aware federated active learning with non-iid data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22279–22289, 2023. 1, 2, 7, 3, 4

[9] Razvan Caramalau, Binod Bhattarai, and Tae-Kyun Kim. Sequential graph convolutional network for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9583–9592, 2021. 2, 7, 3, 4

[10] Jiayi Chen, Benteng Ma, Hengfei Cui, and Yong Xia. Think twice before selection: Federated evidential active learning for medical image analysis with domain shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11439–11449, 2024. 1, 2

[11] Zhen Chen, Meilu Zhu, Chen Yang, and Yixuan Yuan. Personalized retrogress-resilient framework for real-world medical federated learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 347–356. Springer, 2021. 2

[12] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006. 3

[13] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 2

[14] Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Ö Arık, Larry S Davis, and Tomas Pfister. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *European Conference on Computer Vision*, pages 510–526. Springer, 2020. 2

[15] Yongxin Guo, Xiaoying Tang, and Tao Lin. Fedbr: Improving federated learning on heterogeneous data via local learning bias reduction. In *International conference on machine learning*, pages 12034–12054. PMLR, 2023. 2

[16] Guy Hacohen and Daphna Weinshall. How to select which active learning strategy is best suited for your specific problem and budget. *Advances in Neural Information Processing Systems*, 36:13395–13407, 2023. 2

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[18] Joseph L Hodges Jr and Erich L Lehmann. Estimates of location based on rank tests. In *Selected works of EL Lehmann*, pages 287–300. Springer, 2011. 3

[19] Alex Holub, Pietro Perona, and Michael C Burl. Entropy-based active learning for object recognition. In *2008 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 1–8. IEEE, 2008. 2, 7, 3, 4

[20] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 5

[21] Siyu Huang, Tianyang Wang, Haoyi Xiong, Jun Huan, and Dejing Dou. Semi-supervised active learning with temporal output discrepancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3447–3456, 2021. 2

[22] Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. Active learning by querying informative and representative examples. *Advances in neural information processing systems*, 23, 2010. 1

[23] Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. Rethinking federated learning with domain shift: A prototype view. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16312–16322. IEEE, 2023. 2

[24] Yongzhe Jia, Xuyun Zhang, Hongsheng Hu, Kim-Kwang Raymond Choo, Lianyong Qi, Xiaolong Xu, Amin Beheshti, and Wanchun Dou. Dapperfl: domain adaptive federated learning with model fusion pruning for edge devices. *Advances in Neural Information Processing Systems*, 37:13099–13123, 2024. 2

[25] Meirui Jiang, Holger R Roth, Wenqi Li, Dong Yang, Can Zhao, Vishwesh Nath, Daguang Xu, Qi Dou, and Ziyue Xu. Fair federated medical image segmentation via client contribution estimation. In *Proceedings of the IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition, pages 16302–16311, 2023. 2

[26] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of big data*, 6 (1):1–54, 2019. 3

[27] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020. 1, 2, 3

[28] SangMook Kim, Sangmin Bae, Hwanjun Song, and Se-Young Yun. Re-thinking federated active learning based on inter-class diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3944–3953, 2023. 1, 2, 7, 3, 4

[29] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 3, 7, 1, 2

[30] Natsumaro Kutsuna, Takumi Higaki, Sachihiro Matsunaga, Tomoshi Otsuki, Masayuki Yamaguchi, Hirofumi Fujii, and Seiichiro Hasezawa. Active learning framework with iterative clustering for bioimage classification. *Nature communications*, 3(1):1032, 2012. 2

[31] Mingkun Li and Ishwar K Sethi. Confidence-based active learning. *IEEE transactions on pattern analysis and machine intelligence*, 28(8):1251–1261, 2006. 2, 7, 3, 4

[32] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020. 1, 2, 3

[33] Xiaoxiao Liang, Yiqun Lin, Huazhu Fu, Lei Zhu, and Xiaomeng Li. Rscfed: Random sampling consensus federated semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10154–10163, 2022. 2

[34] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1013–1023, 2021. 2

[35] Benteng Ma, Yu Feng, Geng Chen, Changyang Li, and Yong Xia. Federated adaptive reweighting for medical image classification. *Pattern Recognition*, 144:109880, 2023. 2

[36] Othmane Marfoq, Giovanni Neglia, Aurélien Bellet, Laetitia Kameni, and Richard Vidal. Federated multi-task learning under a mixture of distributions. *Advances in neural information processing systems*, 34:15434–15447, 2021. 2

[37] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. 1, 2, 3

[38] Hieu T Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 79, 2004. 2

[39] Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Gholamreza Reza Haffari, Anton Van Den Hengel, and Javen Qinfeng Shi. Active learning by feature mixing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12237–12246, 2022. 2, 7, 3, 4

[40] Viraj Prabhu, Arjun Chandrasekaran, Kate Saenko, and Judy Hoffman. Active domain adaptation via clustering uncertainty-weighted embeddings. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8505–8514, 2021. 2

[41] Yasar Abbas Ur Rehman, Yan Gao, Pedro Porto Buarque De Gusmao, Mina Alibeigi, Jiajun Shen, and Nicholas D Lane. L-dawa: Layer-wise divergence aware weight aggregation in federated self-supervised visual representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16464–16473, 2023. 2

[42] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021. 1

[43] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017. 2, 3, 6, 7, 4

[44] Burr Settles. Active learning literature survey. 2009. 1, 3

[45] Donald Shenaj, Marco Toldo, Alberto Rigon, and Pietro Zanuttigh. Asynchronous federated continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5055–5063, 2023. 2

[46] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. *Advances in neural information processing systems*, 30, 2017. 2

[47] Ruth Urner, Sharon Wulff, and Shai Ben-David. Plal: Cluster-based active learning. In *Conference on learning theory*, pages 376–397. PMLR, 2013. 2

[48] Fan Wang, Zhongyi Han, Zhiyan Zhang, Rundong He, and Yilong Yin. Mhpl: Minimum happy points learning for active source free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20008–20018, 2023. 2

[49] Jiacheng Wang, Yueming Jin, and Liansheng Wang. Personalizing federated medical image segmentation via local calibration. In *European Conference on Computer Vision*, pages 456–472. Springer, 2022. 2

[50] Thorben Werner, Johannes Burchert, Maximilian Stubbemann, and Lars Schmidt-Thieme. A cross-domain benchmark for active learning. *Advances in Neural Information Processing Systems*, 37:62875–62911, 2024. 2

[51] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pages 196–202. Springer, 1992. 3

[52] Xing Wu, Jie Pei, Cheng Chen, Yimin Zhu, Jianjia Wang, Quan Qian, Jian Zhang, Qun Sun, and Yike Guo. Federated active learning for multicenter collaborative disease diagnosis. *IEEE transactions on medical imaging*, 42(7):2068–2080, 2022. 2

[53] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 7, 1, 2

[54] Xiao Xiao, Xinyue Yang, Yucheng Wang, Yilong Hui, Jianchuan Zhou, and Guoqiang Mao. Unsupervised domain adaptive vehicle re-identification: A federated learning scheme. *IEEE Transactions on Intelligent Transportation Systems*, 2025. 2

[55] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023. 7, 1, 2

[56] Mang Ye, Xiuwen Fang, Bo Du, Pong C Yuen, and Dacheng Tao. Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Computing Surveys*, 56(3):1–44, 2023. 2

[57] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 93–102, 2019. 2

[58] Jaehong Yoon, Wonyong Jeong, Giwoong Lee, Eunho Yang, and Sung Ju Hwang. Federated continual learning with weighted inter-client transfer. In *International conference on machine learning*, pages 12073–12086. PMLR, 2021. 2

[59] Jiakang Yuan, Bo Zhang, Xiangchao Yan, Tao Chen, Botian Shi, Yikang Li, and Yu Qiao. Bi3d: Bi-domain active learning for cross-domain 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15599–15608, 2023. 2

[60] Ruipeng Zhang, Qinwei Xu, Jiangchao Yao, Ya Zhang, Qi Tian, and Yanfeng Wang. Federated domain generalization with generalization adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3954–3963, 2023. 2

[61] Chen-Chen Zong and Sheng-Jun Huang. Rethinking epistemic and aleatoric uncertainty for active open-set annotation: An energy-based approach. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10153–10162, 2025. 2

[62] Chen-Chen Zong, Tong Jin, and Sheng-Jun Huang. Inconsistency-based federated active learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 7300–7308. International Joint Conferences on Artificial Intelligence Organization, 2025. Main Track. 1, 2, 7, 3, 4