DELVING INTO INSTANCES: A UNIFIED PERSPECTIVE FOR FAIR FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Fairness in federated learning (FL) is commonly studied at the client or group level to reduce performance disparities across clients or groups. Such coarse-grained views overlook an important source of unfairness: *similar instances may receive inconsistent predictions, even within the same client*. In this work, we find that *instances provide a unified and finer-grained perspective for fairness in FL*. Instance-level prediction consistency not only characterizes *intra-client unfairness* among similar instances within a client, but also provides a basis for identifying and mitigating *inter-client unfairness* through differences in the consistency of instances across clients. Based on this insight, we propose Federated Instance-Consistency Intervention Learning (FedICIL), a unified instance-level framework for fair FL. Within each client, FedICIL constructs intra-client neighborhoods according to instance similarity and explicitly regularizes predictions to be consistent within each neighborhood. To make the intervention more targeted under limited local update budgets, FedICIL further identifies parameters that are sensitive to instance-level prediction inconsistency and selectively updates them. Beyond the client level, clients communicate a lightweight instance-consistency signal, which is used to adapt aggregation weights according to the remaining inconsistency of their local predictions. In this way, the same instance-level signal is exploited to address both intra-client inconsistency and inter-client performance disparities. Experiments on diverse benchmarks demonstrate that FedICIL achieves a better trade-off among predictive performance, inter-client fairness, intra-client fairness, and group fairness compared with existing methods. Code is available at https://anonymous.4open.science/r/FedICIL-E863/.

1 INTRODUCTION

Federated learning (FL) enables multiple clients to collaboratively train a global model without sharing their raw data (McMahan et al., 2017). However, heterogeneous data distributions across clients can lead to substantial disparities in model behavior, making fairness a fundamental challenge in federated systems (Mukhtiar et al., 2025; Yue et al., 2026). Existing studies have primarily considered fairness at the *client level*, aiming to reduce differences in predictive performance across clients (Li et al., 2020c; 2021b; Guan et al., 2025). Fairness has also been investigated at the *group level*, where models are expected to provide comparable outcomes across demographic groups such as sex and age (Hardt et al., 2016; Ezzeldin et al., 2023; Su et al., 2024a; Hamman & Dutta, 2024; Zhang et al., 2025b;a). While these perspectives are important, they treat clients or demographic groups as the primary units of fairness and therefore focus mainly on aggregate performance disparities.

In this work, we found that *instances provide a finer-grained and more fundamental unit for examining fairness in FL*. From an individual fairness perspective, similar individuals should receive similar predictions (Dwork et al., 2012; Li et al., 2023; Lin et al., 2025). This principle reveals a form of unfairness that can be hidden by client-level or group-level metrics: *similar instances may receive substantially different predictions even when the overall performance of their client is comparable to that of other clients*. For example, as illustrated in Figure 1(a), two clients may achieve similar overall accuracy, while the models within each client exhibit inconsistent predictions for semantically similar instances. Conversely, the degree of instance-level inconsistency may differ substantially across clients, indicating that aggregate client performance alone does not fully characterize how reliably a model treats individual instances.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

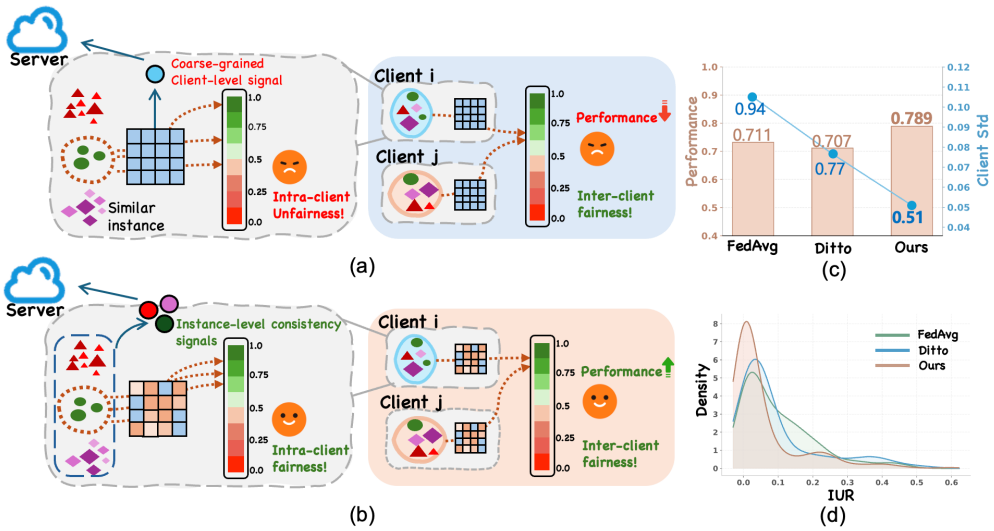


Figure 1: **Illustration** of fairness at different levels in FL. (a) Balanced performance across clients can coexist with inconsistent predictions among similar instances within each client. (b) Our method exploits instance-level consistency signals during both local optimization and global aggregation. (c) At the client level, our method achieves a better trade-off between global performance and cross-client performance disparity. (d) At the instance level, our method yields lower IUR, indicating more consistent predictions for similar instances. More details can be seen in Appendix § A.

This observation leads to a key insight of this work: *instance-level prediction consistency can serve as a unified signal for both intra-client and inter-client fairness.* At the *intra-client* level, consistency measures whether similar instances belonging to the same client receive similar predictions (Zhang et al., 2024). Reducing this inconsistency directly addresses individual unfairness that is invisible to aggregate client-level metrics (Chen et al., 2024a). At the inter-client level, the consistency of instances provides information about how reliably different clients are served by the federated model. Clients exhibiting larger residual inconsistency can therefore be identified as having greater unmet fairness needs and treated differently during global aggregation. Thus, rather than designing separate mechanisms for different fairness levels, we explore whether a common instance-level signal can connect and jointly address both types of unfairness.

This perspective differs fundamentally from conventional fair FL methods. Existing client-level approaches typically reweight client objectives according to training difficulty or performance (Li et al., 2020c; Cheng et al., 2024; Yue et al., 2023), introduce personalized local models (Li et al., 2021b; Hao et al., 2025), or adapt aggregation according to client performance and update information (Chen et al., 2024b; He et al., 2025; Guan et al., 2025). Group-fair FL methods instead rely on demographic-group information or group-level statistics to reduce disparities across sensitive groups (Ezzeldin et al., 2023; Su et al., 2024a). These approaches can improve aggregate fairness, but they do not explicitly constrain the consistency of predictions among similar instances. More importantly, client- or group-level statistics provide only coarse summaries and may fail to reveal localized prediction inconsistencies. Our central idea is therefore to move the fairness intervention from the client/group to the instance, and then use the resulting instance-level signal at both local and global levels.

Based on this insight, we propose *Federated Instance-Consistency Intervention Learning* (FedICIL), a unified framework for addressing both intra- and inter-client fairness in FL. FedICIL constructs intra-client neighborhoods based on instance similarity and introduces an instance-consistency loss to encourage consistent predictions among similar instances. It further leverages loss sensitivity to identify parameters that are most sensitive to local prediction discrepancies and selectively updates a subset under a constrained update budget, focusing optimization on the parameters most relevant to instance-level inconsistency. Meanwhile, each client reports a lightweight consistency signal to the server, enabling fairness-aware reweighted aggregation. Clients with higher remaining prediction inconsistency receive smaller aggregation weights, reducing their influence on the global model. We conduct extensive experiments on diverse FL benchmarks covering a range of tasks, includ-

ing natural language understanding, general image classification, and medical image classification. Experimental results show that FedICIL achieves a better overall trade-off among predictive performance and multiple fairness dimensions, including inter-client, intra-client, and group fairness.

Our contributions are summarized as follows:

- We introduce an *instance-level perspective on fairness in FL*. Rather than treating clients or demographic groups as the primary units of fairness, we identify instance-level prediction consistency as a finer-grained signal that can characterize both intra-client unfairness and inter-client disparities.
- We propose *Federated Instance-Consistency Intervention Learning* (FedICIL), a unified instance-level fairness framework. FedICIL constructs intra-client neighborhoods and enforces prediction consistency, selectively intervenes on inconsistency-sensitive parameters under a constrained update budget, and uses the resulting consistency signal for fairness-aware global aggregation.
- We conduct extensive experiments on diverse FL benchmarks and medical datasets. FedICIL achieves a favorable trade-off among global predictive performance, inter-client fairness, and intra-client consistency, and further demonstrates its effectiveness for demographic group fairness across sex and age groups.

2 Related Work

Heterogeneous Federated Learning. Statistical heterogeneity across clients, commonly referred to as the non-IID problem, remains a major challenge in FL (Li et al., 2020a; Zhu et al., 2021; Ma et al., 2022; Ye et al., 2024a; Lu et al., 2024; Karami & Karami, 2025). FedAvg (McMahan et al., 2017) can suffer from performance degradation under heterogeneous data. To mitigate this issue, FedProx (Li et al., 2020b) constrains local updates with a proximal term, while SCAFFOLD (Karimireddy et al., 2020) corrects client drift using control variates. Other methods address heterogeneity at the representation level. MOON (Li et al., 2021a) uses model-contrastive learning, and FedProto (Tan et al., 2022) aligns class prototypes across clients. More recently, FL has been extended to pretrained-model adaptation. LoRA (Hu et al., 2022) reduces trainable parameters. Federated LoRA methods address data heterogeneity (Sun et al., 2024; Yan et al., 2025) or learn personalized adapters (Hao et al., 2025), while FedAPT (Su et al., 2024b) generates sample-adaptive prompts. These methods mainly target task performance under heterogeneity. Our work instead studies fairness through prediction consistency among similar instances, covering both conventional FL and pretrained-model adaptation.

Fair Federated Learning. Fairness in FL has been studied from several perspectives, including client-level performance fairness (Mohri et al., 2019; Li et al., 2020c; 2021b; Guan et al., 2025), group fairness (Hardt et al., 2016; Ezzeldin et al., 2023; Su et al., 2024a), and individual fairness (Dwork et al., 2012; Zemel et al., 2013; Li et al., 2023). Client-level fairness aims to reduce performance disparities across participating clients. AFL (Mohri et al., 2019) considers adversarial mixtures of client distributions, while q-FFL (Li et al., 2020c) places greater emphasis on clients with higher losses. Ditto (Li et al., 2021b) approaches the problem through personalization, whereas FedPW (Guan et al., 2025) combines parameter importance with client dynamics. Group fairness instead focuses on disparities across groups defined by sensitive attributes (Hardt et al., 2016). FairFed (Ezzeldin et al., 2023) adjusts aggregation using group-fairness information, and mFairFL (Su et al., 2024a) jointly considers group and client fairness. Recent work also explores server-side fairness intervention (He et al., 2025) and demographic fairness in federated medical imaging (Li et al., 2026). At the individual level, fairness requires similar individuals to receive similar outcomes under a task-relevant similarity metric (Dwork et al., 2012). Prior work addresses this principle through fair representation learning (Zemel et al., 2013) and adversarial training in federated settings (Li et al., 2023). FedICIL differs by directly incorporating prediction consistency among similar instances into federated optimization, and by using the same consistency objective to guide sensitivity-based parameter updates and fairness-aware global aggregation.

3 PRELIMINARIES

Federated Learning. Consider an FL system with K clients, where client k holds a private dataset $\mathcal{D}_k = \{(x_i^k, y_i^k)\}_{i=1}^{n_k}$. The clients jointly optimize a global model $f(\cdot; w)$ without sharing local data. The global objective is $\mathcal{L}_G(w) = \sum_{k=1}^{K} \lambda_k \mathcal{L}_k(w)$, where $\mathcal{L}_k(w) = \frac{1}{n_k} \sum_{(x_i^k, y_i^k) \in \mathcal{D}_k} \ell(f(x_i^k; w), y_i^k)$ and $\lambda_k = n_k / \sum_j n_j$. At round t, each participating client initializes from w^{t-1} and performs local optimization to obtain w_k^t, after which the server aggregates the local models as

$$w^t = \sum_{k=1}^{K} \lambda_k w_k^t. \tag{1}$$

Here, w may denote either the full model parameters or a parameter-efficient subset.

Fairness in Federated Learning. Fairness in FL can be characterized at multiple levels. We consider *inter-client fairness*, which concerns disparities in predictive performance across clients, and *intra-client fairness*, which requires similar instances within the same client to receive consistent predictions (Dwork et al., 2012). Specifically, letting $M_k(w)$ denote the predictive performance on client k, inter-client disparity can be measured by $\mathrm{Var}(M_k(w)_{k=1}^K)$, where a smaller value indicates more balanced client performance. For two instances x_i^k, x_j^k from client k, intra-client individual fairness can be formulated as $\Delta(f(x_i^k; w), f(x_j^k; w)) \leq L\, d(x_i^k, x_j^k)$, where $d(\cdot, \cdot)$ and $\Delta(\cdot, \cdot)$ measure instance similarity and prediction discrepancy, respectively. We additionally consider *group fairness*, which concerns disparities in outcomes or error rates across groups defined by sensitive attributes (Hardt et al., 2016) such as sex and age.

4 METHODOLOGY

FedICIL is designed to jointly address intra-client and inter-client unfairness from a unified instance-level perspective. It consists of two components: *Instance-Consistency Interventional Local Optimization* and *fairness-aware Global Aggregation*. The former encourages consistent predictions among similar instances and selectively updates inconsistency-sensitive parameters under a constrained update budget. The latter uses the remaining prediction inconsistency to reweight client contributions, extending the same instance-level signal from local optimization to global aggregation. The overall workflow is illustrated in Figure 2.

4.1 INSTANCE-CONSISTENCY INTERVENTIONAL LOCAL OPTIMIZATION

Local optimization addresses intra-client prediction inconsistency through neighborhood-based regularization and sensitivity-guided parameter updates. The consistency objective determines both the prediction discrepancies to reduce and the parameters to prioritize.

Instance Consistency Modeling. For each instance $x_i^k \in \mathcal{D}_k$, let $p_i^k = \mathrm{softmax}\big(f(x_i^k; w_k^{t,e})\big)$ denote the predictive distribution at local step e, while $h_i^k = \mathrm{softmax}\big(f(x_i^k; w^{t-1})\big)$ denotes the predictive representation. Accordingly, the similarity between two instances within client k is defined as $s_{ij}^k = -\|h_i^k - h_j^k\|_2$. For each x_i^k, we construct a neighborhood $\mathcal{N}_i^k$ containing its top-H most similar same-label instances.

Instance-Consistency Loss. The neighborhood provides a local reference for the target-class confidence of each instance. We define the neighborhood consensus as

$$\bar{q}_i^k = \frac{1}{|\mathcal{N}_i^k|} \sum_{x_j^k \in \mathcal{N}_i^k} h_j^k[y_i^k]. \tag{2}$$

Here, $h_j^k[y_i^k]$ denotes the target-class probability of neighbor x_j^k. The inconsistency residual of each instance and the corresponding local consistency loss are

$$r_i^k = \big(p_i^k[y_i^k] - \bar{q}_i^k\big)^2, \qquad \mathcal{L}_{\mathrm{IC}}^k = \frac{1}{n_k} \sum_{i=1}^{n_k} r_i^k. \tag{3}$$

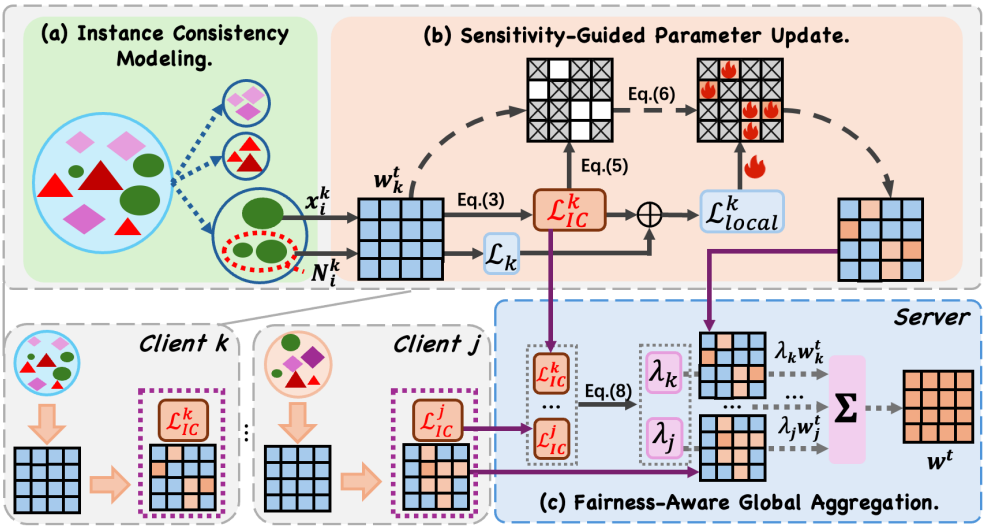


Figure 2: **Architecture illustration of FedICIL.** (a) On each client, ***Instance Consistency Modeling*** groups local instances according to similarity and establishes a neighbor set for each sample. (b) Using the ***Instance-Consistency Loss*** $\mathcal{L}_{\mathrm{IC}}^{k}$ induced by local neighborhoods, each client masks insensitive parameters and performs ***Sensitivity-Guided Parameter Update*** via $\mathcal{L}_{\mathrm{local}}^{k}$. (c) The local models, together with $\mathcal{L}_{\mathrm{IC}}^{k}$ as an inconsistency score, are then transmitted to the server for ***Fairness-Aware Global Aggregation***.

Here, r_i^k measures an individual instance's deviation from its neighborhood consensus, while $\mathcal{L}_{\mathrm{IC}}^{k}$ summarizes prediction inconsistency within client k. The local training objective is then

$$\mathcal{L}_{\mathrm{local}}^{k}(w_k^{t,e}) = \mathcal{L}_k(w_k^{t,e}) + \beta \mathcal{L}_{\mathrm{IC}}^{k}(w_k^{t,e}), \tag{4}$$

where $\beta > 0$ controls the consistency regularization strength. This objective combines task learning with the individual fairness principle of consistent predictions for similar instances.

Sensitivity-Guided Parameter Update. Different parameters have different sensitivities to local prediction inconsistency. Under a constrained update budget, FedICIL uses $\mathcal{L}_{\mathrm{IC}}^{k}$ to identify where the consistency intervention should be concentrated. For the m-th trainable parameter, the sensitivity score is

$$g_m^k = \left| \frac{\partial \mathcal{L}_{\mathrm{IC}}^{k}(w_k^{t,e})}{\partial w_{k,m}^{t,e}} \right|. \tag{5}$$

The top-ρ proportion of parameters forms an update mask $\mathcal{M}_k$. These parameters are updated using the joint local objective, while the remaining parameters are kept unchanged:

$$w_{k,m}^{t,e+1} = \begin{cases} w_{k,m}^{t,e} - \eta \dfrac{\partial \mathcal{L}_{\mathrm{local}}^{k}(w_k^{t,e})}{\partial w_{k,m}^{t,e}}, & m \in \mathcal{M}_k, \\ w_{k,m}^{t,e}, & m \notin \mathcal{M}_k. \end{cases} \tag{6}$$

Thus, instance-level consistency guides parameter selection, while the selected parameters jointly optimize task performance and local consistency. Depending on the training setting, the candidate parameter set can comprise the full model or a parameter-efficient subset.

4.2 Fairness-Aware Global Aggregation

To connect local consistency intervention with inter-client collaboration, FedICIL reuses the consistency loss as an aggregation signal. After local optimization, each participating client $k \in \mathcal{S}^t$ uploads its local model w_k^t together with the residual inconsistency score

$$\Delta_k^t = \mathcal{L}_{\mathrm{IC}}^{k}(w_k^t). \tag{7}$$

A larger Δ_k^t indicates greater remaining prediction inconsistency among similar instances. This score summarizes instance-level behavior within each client and allows it to inform the contribution of that client during aggregation. The original coefficient λ_k is reweighted as

$$\lambda_k^t = \frac{\lambda_k \exp(-\gamma \Delta_k^t)}{\sum_{j \in \mathcal{S}^t} \lambda_j \exp(-\gamma \Delta_j^t)}, \tag{8}$$

where $\gamma > 0$ controls the reweighting strength. The exponential factor places a stronger penalty on larger residual inconsistency, while λ_k retains the local data-size information. The global model is updated as

$$w^t = \sum_{k \in \mathcal{S}^t} \lambda_k^t w_k^t. \tag{9}$$

FedICIL therefore uses the same consistency signal to guide local parameter intervention and global client reweighting. This connects the treatment of similar instances within clients with the aggregation of models across clients, providing a unified framework for addressing intra-client and inter-client unfairness.

5 Experiments

5.1 Experimental Setup

Task Setup. We conduct experiments across different task settings: (1) **natural language understanding (NLU)**, including CoLA, QNLI, RTE, MRPC, MNLI (Wang et al., 2018), and 20NG (Lang, 1995); (2) **cross-domain image classification** on Office-Caltech-10 (Gong et al., 2012; Guan et al., 2025); and (3) **medical image classification** on a private brain MRI dataset comprising 3,421 patients and 10,986 MRI examinations for four-class brain tumor classification. These experiments cover diverse data modalities, tasks, and model architectures, enabling a comprehensive evaluation.

The NLU experiments follow the experimental setting of FRLoRA (Yan et al., 2025), using RoBERTa-base (Liu et al., 2019) with LoRA (Hu et al., 2022) across five heterogeneous clients. The Office-Caltech-10 (Gong et al., 2012) experiments follow the setting of FedPW (Guan et al., 2025), using ResNet-10 (He et al., 2016) across 20 clients. For both settings, we adopt the corresponding data partitioning, client configuration, and training protocol from these works.

For the private brain MRI dataset, we use an ImageNet-pretrained (Deng et al., 2009) ResNet-50 (He et al., 2016) across three clients. We compare FedICIL with representative standard, heterogeneous, client-fair, and group-fair FL methods. All experiments are conducted on a single NVIDIA A40 GPU. Detailed dataset statistics, baseline descriptions, and implementation settings are provided in Appendix B.

Evaluation Metrics. We evaluate FedICIL from four complementary perspectives: predictive performance, inter-client fairness, intra-client prediction consistency, and group fairness. For **predictive performance**, we report the Matthews correlation coefficient (MCC) on CoLA and accuracy on the remaining tasks. For **inter-client fairness**, we mainly report the standard deviation of client accuracies (Std), where a lower value indicates more balanced performance across clients. We additionally report the maximum–minimum accuracy gap (Gap) and worst-client accuracy (Worst) when analyzing training dynamics. For **intra-client prediction consistency**, following the individual fairness principle that similar instances should receive similar predictions (Dwork et al., 2012; Zemel et al., 2013), we use the *Individual Unfairness Residual* (IUR). Specifically, IUR measures the average Jensen–Shannon divergence (Lin, 1991) between the prediction of each instance and the mean prediction of its H nearest neighbors in the representation space:

$$\mathrm{IUR} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{JS}\left(p_i, \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} p_j \right). \tag{10}$$

For easier comparison, we report its log-transformed form $\mathrm{IUR}^{\dagger} = -\log_{10}(\mathrm{IUR} + 10^{-12})$. A higher $\mathrm{IUR}^{\dagger}$ indicates better intra-client prediction consistency. For the medical dataset, we additionally

Table 1: Performance on four benchmarks in terms of accuracy, client standard deviation, and IUR†. ↑ denotes an improvement over FedAvg, while ↓ denotes a degradation.The best results are marked in **bold**, and the second-best results are underlined.

Method	CoLA			RTE			MRPC			MNLI		
	MCC ↑	Std ↓	IUR† ↑	Acc ↑	Std ↓	IUR† ↑	Acc ↑	Std ↓	IUR† ↑	Acc ↑	Std ↓	IUR† ↑
FedAvg	54.69	15.08	3.14	71.84	8.67	1.17	86.52	2.95	4.22	80.66	7.19	2.58
FedAdagrad	53.10 ↓1.59	9.60 ↑5.48	2.65 ↓0.49	70.40 ↓1.44	5.69 ↑2.98	1.08 ↓0.09	84.80 ↓1.72	**1.34** ↑1.61	2.60 ↓1.62	76.46 ↓4.20	6.97 ↑0.22	2.47 ↓0.11
FedProx	57.30 ↑2.61	12.34 ↑2.74	2.49 ↓0.65	72.92 ↑1.08	11.59 ↓2.92	1.25 ↑0.08	85.29 ↓1.23	4.38 ↓1.43	4.27 ↑0.05	81.10 ↑0.44	4.71 ↑2.48	2.58 →0.00
SCAFFOLD	55.34 ↑0.65	9.24 ↑5.84	4.41 ↑1.27	73.65 ↑1.81	4.52 ↑4.15	1.49 ↑0.32	85.78 ↓0.74	2.72 ↑0.23	4.52 ↑0.30	73.01 ↓7.65	4.36 ↑2.83	**3.10** ↑0.52
q-FFL	56.96 ↑2.27	7.34 ↑7.74	2.86 ↓0.28	74.73 ↑2.89	5.27 ↑3.40	1.20 ↑0.03	87.25 ↑0.73	5.26 ↓2.31	4.36 ↑0.14	82.10 ↑1.44	4.10 ↑3.09	2.68 ↑0.10
AFL	61.16 ↑6.47	7.17 ↑7.91	3.20 ↑0.06	73.29 ↑1.45	5.47 ↑3.20	1.31 ↑0.14	85.78 ↓0.74	3.33 ↓0.38	4.97 ↑0.75	81.85 ↑1.19	**2.74** ↑4.45	2.44 ↓0.14
Ditto	56.00 ↑1.31	14.68 ↑0.40	2.72 ↓0.42	75.45 ↑3.61	7.89 ↑0.78	1.33 ↑0.16	85.05 ↓1.47	4.86 ↓1.91	4.03 ↓0.19	81.74 ↑1.08	4.37 ↑2.82	2.59 ↑0.01
FedPW	49.72 ↓4.97	19.74 ↓4.66	2.82 ↓0.32	70.76 ↓1.08	7.00 ↑1.67	1.21 ↑0.04	83.58 ↓2.94	8.53 ↓5.58	2.58 ↓1.64	75.62 ↓5.04	5.42 ↑1.77	2.40 ↓0.18
FedHEAL	53.32 ↓1.37	9.45 ↑5.63	2.79 ↓0.35	70.40 ↓1.44	10.44 ↓1.77	1.17 →0.00	83.09 ↓3.43	5.66 ↓2.71	6.49 ↑2.27	76.91 ↓3.75	6.52 ↑0.67	2.75 ↑0.17
FedICIL (Ours)	**61.89** ↑7.20	**6.93** ↑8.15	**11.86** ↑8.72	**80.87** ↑9.03	**3.09** ↑5.58	**2.64** ↑1.47	**87.99** ↑1.47	2.20 ↑0.75	**12.00** ↑7.78	**84.63** ↑3.97	5.65 ↑1.54	2.97 ↑0.39

evaluate **group fairness** using the accuracy gap between sex groups (Sex Gap) and between patients younger than 65 and those aged 65 or older (Age Gap), where lower values indicate smaller demographic disparities. (Seyyed-Kalantari et al., 2021a;b; Xing et al., 2025). To evaluate the trade-off across predictive performance and multiple fairness dimensions, we additionally report **Mean Rank** over the summary metrics, where a lower value indicates a more favorable overall trade-off.

5.2 Quantitative Comparison

Results on NLU Tasks. FedICIL offers a more favorable overall trade-off between predictive performance and fairness than the competing methods in Table 1. It achieves the best predictive performance on all four benchmarks and is the only method to improve all three metrics over FedAvg on every benchmark. On RTE, FedICIL improves global accuracy from 71.84% to 80.87% and reduces Std from 8.67 to 3.09. In comparison, AFL achieves the lowest Std on MNLI but obtains a lower IUR† than FedAvg, showing that more balanced client performance can still leave local prediction inconsistency unresolved. The training curves in Figure 3 further show higher worst-client accuracy and a smaller gap between the best- and worst-performing clients. Together with the lower IUR distributions in Figure 4, these results show that the benefits of instance-level intervention extend to both prediction consistency within clients and the performance of underperforming clients. Additional results and analyses are provided in Appendix C.1.

Results on Cross-Domain Image Classification. As shown in Table 2, FedICIL achieves the best Std and IUR† while ranking second in average accuracy, yielding the best Mean Rank of 1.33. Compared with FedPW, it improves average accuracy from 59.90% to 62.51% and reduces Std from 7.70 to 4.13. The largest gain occurs on DSLR, the weakest domain under FedPW, where accuracy increases from 49.33% to 66.00%. Although accuracy decreases on Amazon and Caltech, the gains on DSLR and Webcam more than offset these declines. The resulting performance is stronger on average, with better inter-client fairness as assessed by domain-level balance. Ditto achieves higher average accuracy, but its Std of 13.43 is substantially larger and its IUR† is lower. Taken together, these results indicate that FedICIL achieves a favorable trade-off between predictive performance,inter-client fairness, and instance consistency under cross-domain heterogeneity.

Results on Medical Image Classification. FedICIL improves all five metrics over FedAvg and achieves the best Mean Rank of 3.00 in Table 3. It ranks second in predictive accuracy and first in both Sex Gap and IUR†. While Ditto achieves higher accuracy, its Sex Gap is 2.591, compared with 0.383 for FedICIL. FedICIL combines strong predictive performance with improvements over FedAvg across all evaluated fairness dimensions, while its Client Std and Age Gap are not the lowest. This yields a favorable overall trade-off among predictive performance, inter-client fairness, demo-

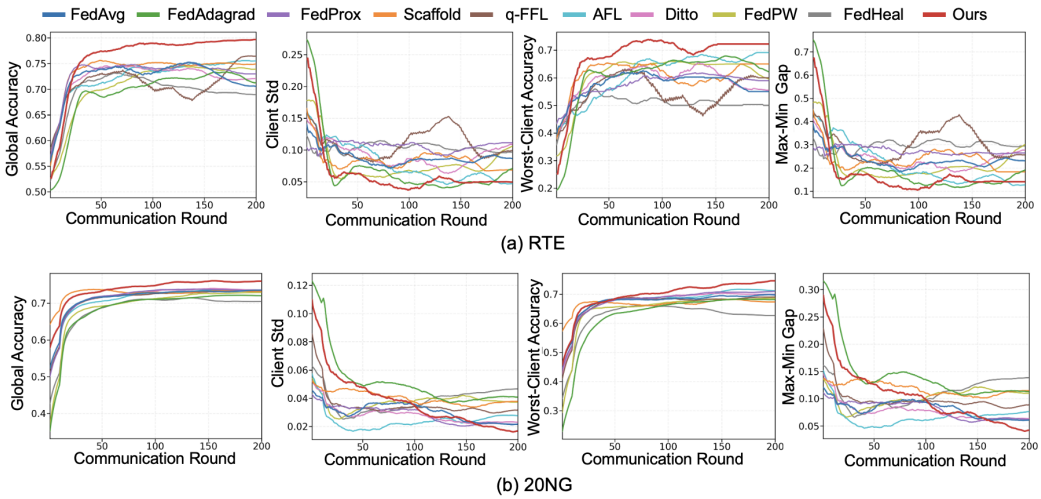


Figure 3: Illustration of round-wise evaluation metrics over communication rounds on (a) RTE and (b) 20NG, covering global accuracy, client accuracy standard deviation, worst-client accuracy, and max–min accuracy gap.

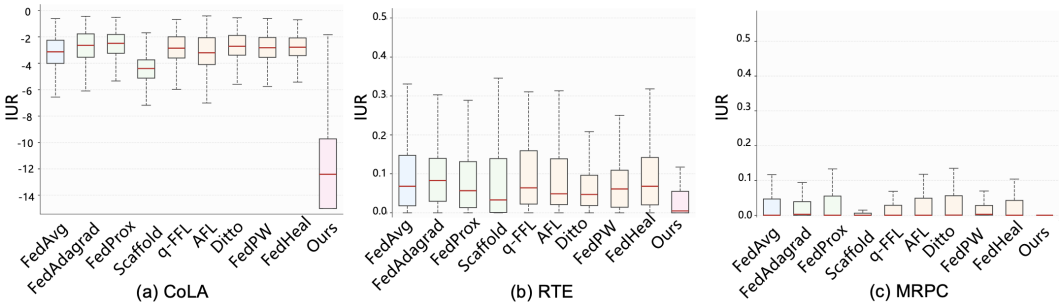


Figure 4: Illustration of IUR distributions on (a) CoLA, (b) RTE, and (c) MRPC.

Table 2: Results on cross-domain image classification using Office-Caltech-10. $(\cdot)$ denotes the rank of each method, and Mean Rank is calculated over AVG, Std, and $\text{IUR}^\dagger$. Std is computed across the 4 domains. The best results are marked in **bold**, and the second-best results are underlined.

Method	Amazon	DSLR	Caltech	Webcam	**AVG** ↑	**Std** ↓	**IUR**† ↑	**Mean Rank** ↓
AFL	20.95	42.67	29.46	32.41	31.37 (11)	8.96 (4)	0.7459 (4)	6.33 (7)
q-FFL	69.58	40.00	60.63	51.72	55.48 (6)	12.64 (7)	0.7463 (3)	5.33 (4)
FedHEAL	54.21	66.67	53.57	58.62	58.27 (5)	6.03 (2)	0.7118 (10)	5.67 (5)
Ditto	73.54	78.67	48.53	72.21	**68.24** (1)	13.43 (9)	0.7310 (9)	6.33 (7)
FedFV	67.89	41.33	60.71	50.00	54.99 (8)	11.70 (6)	0.7390 (7)	7.00 (9)
FedCKA	70.21	40.67	61.07	48.62	55.14 (7)	13.09 (8)	0.7561 (2)	5.67 (5)
FedGCR	71.16	57.33	60.09	46.21	58.70 (4)	10.25 (5)	0.7431 (5)	4.67 (3)
FedSaC	34.32	51.07	56.07	74.83	54.07 (9)	16.68 (10)	0.7312 (8)	9.00 (10)
FedAA	68.63	18.00	34.73	15.86	34.31 (10)	24.39 (11)	0.5979 (11)	10.67 (11)
FedPW	67.68	49.33	62.23	60.34	59.90 (3)	7.70 (3)	0.7426 (6)	4.00 (2)
FedICIL	62.00	66.00	56.88	65.17	62.51 (2)	**4.13** (1)	**0.7670** (1)	**1.33 (1)**

graphic fairness, and instance consistency. Importantly, Sex Gap decreases from 1.139 to 0.383 and Age Gap from 4.480 to 2.719 relative to FedAvg, even though neither disparity is directly optimized. These results suggest that instance-level intervention can also reduce demographic disparities, extending its benefits beyond client-level performance balance.

5.3 Ablation and Efficiency Analysis

Effectiveness of Instance-Consistency Interventional Local Optimization. To assess the effectiveness of neighborhood-based consistency learning, we examine two variants that replace the top-H neighborhood $\mathcal{N}_i$ with all same-label instances within the client or remove $\mathcal{L}_{\text{IC}}$ by setting $\beta = 0$

Table 3: Results on medical image classification using the private brain MRI dataset. $(\cdot)$ denotes the rank of each method, and Mean Rank is calculated over all five metrics. The best results are marked in **bold**, and the second-best results are underlined.

Method	ACC ↑	Client Std ↓	Sex Gap ↓	Age Gap ↓	IUR† ↑	Mean Rank ↓
FedAvg	76.969 (7)	5.079 (9)	1.139 (5)	4.480 (6)	0.9777 (3)	6.00 (7)
AFL	75.483 (9)	3.105 (7)	1.883 (6)	8.425 (9)	0.9423 (8)	7.80 (9)
q-FFL	75.929 (8)	3.367 (8)	0.716 (3)	1.482 (3)	0.9370 (9)	6.20 (8)
Ditto	**81.278** (1)	1.838 (3)	2.591 (7)	7.346 (8)	0.9894 (2)	4.20 (3)
FedHEAL	77.266 (6)	**0.846** (1)	2.786 (9)	2.625 (4)	0.9616 (4)	4.80 (5)
FedPW	77.415 (5)	1.681 (2)	0.658 (2)	0.605 (2)	0.9447 (7)	3.60 (2)
FairFed	77.563 (4)	2.186 (5)	1.047 (4)	6.989 (7)	0.9605 (5)	5.00 (6)
mFairFL	78.306 (3)	2.010 (4)	2.673 (8)	**0.373** (1)	0.9496 (6)	4.40 (4)
FedICIL	78.752 (2)	2.198 (6)	**0.383** (1)	2.719 (5)	**1.0061** (1)	**3.00 (1)**

Table 4: **Ablation studies** on RTE in terms of accuracy, client standard deviation, maximum performance gap, worst-client accuracy, and IUR†. The best results are marked in **bold**.

Setting	Acc (%)↑	Std (%)↓	Max Gap (%)↓	Worst Client (%)↑	IUR† ↑
FedAvg	71.84	8.67	23.08	70.56	1.17
w/o $\mathcal{L}_{\mathrm{IC}}$	75.45	6.59	18.00	70.00	2.24
w/o $\mathcal{N}_i$	76.53	5.06	13.07	69.12	2.19
w/o $\mathcal{M}_k$	78.34	4.18	10.00	74.00	2.32
w/o Cons. Agg.	79.78	6.81	20.93	64.00	2.43
FedICIL (Ours)	**80.87**	**3.09**	**8.95**	**76.47**	**2.64**

in Equation (4). Both variants reduce IUR† in Table 4, showing that similarity-based neighborhoods and explicit regularization jointly improve intra-client fairness. Removing the loss also increases Max Gap from 8.95 to 18.00, with higher best-client but lower worst-client accuracy. This suggests that the consistency loss also helps mitigate the bias toward high-performing clients. The same consistency objective further guides parameter selection. Compared with random masking under the same update budget, sensitivity-guided masking improves all five RTE metrics, raising accuracy from 78.34% to 80.87% and reducing Std from 4.18 to 3.09.

Effectiveness of Fairness-Aware Global Aggregation. We replace consistency-based reweighting with standard FedAvg aggregation. As shown in Table 4, global accuracy decreases by 1.09 percentage points. Worst-client accuracy drops by 12.47 percentage points, compared with only 0.49 for best-client accuracy. This asymmetric decline suggests that consistency-based reweighting is particularly beneficial to lower-performing clients. Reusing the local consistency signal to adjust aggregation weights therefore complements local optimization, improving both predictive performance and inter-client fairness.

Communication Efficiency. To evaluate the communication efficiency of FedICIL, we use the same parameter update budget as LoRA. As shown in Table 5, FedICIL incurs an average uplink cost of 35.22 MB per client per round. While this exceeds the cost of LoRA, it is 92.94% lower than that of full-parameter fine-tuning.

Table 5: Client uplink communication on RTE.

Method	Budget (M)	Uplink (MB)
Full	124.65	498.59
LoRA	1.77	7.09
FedICIL	1.77	35.22 ± 6.68

6 Conclusion

In this work, we study fairness in FL from an instance-level perspective and propose FedICIL to jointly improve intra-client and inter-client fairness through prediction consistency. FedICIL addresses prediction inconsistencies that aggregate fairness measures can overlook, using instance consistency to guide both local optimization and global aggregation. Extensive experiments on NLU, cross-domain image classification, and medical image classification demonstrate that FedICIL achieves a favorable trade-off across predictive performance and multiple fairness dimensions.

AI USE STATEMENT

We only used generative AI tools for minor tasks such as grammar checking and expression polishing. The generative AI tools did not contribute to research ideation, experimental design, analysis, or drafting of technical content. All scientific claims, experimental results, and writing remain the sole responsibility of the authors.

REFERENCES

Huiqiang Chen, Tianqing Zhu, Tao Zhang, Wanlei Zhou, and Philip S. Yu. Privacy and fairness in federated learning: On the perspective of tradeoff. *ACM Computing Surveys*, 56(2):1–37, 2024a. doi: 10.1145/3606017.

Yuhang Chen, Wenke Huang, and Mang Ye. Fair federated learning under domain skew with local consistency and domain diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12077–12086, 2024b.

Shu-Ling Cheng, Chin-Yuan Yeh, Ting-An Chen, Eliana Pastor, and Ming-Syan Chen. FedGCR: Achieving performance and fairness for federated learning with distinct client types via group customization and reweighting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11498–11506, 2024. doi: 10.1609/aaai.v38i10.29031.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226. ACM, 2012. doi: 10.1145/2090236.2090255. URL https://doi.org/10.1145/2090236.2090255.

Yahya H. Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A. Salman Avestimehr. FairFed: Enabling group fairness in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7494–7502, 2023. doi: 10.1609/aaai.v37i6.25911. URL https://doi.org/10.1609/aaai.v37i6.25911.

Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2066–2073. IEEE, 2012. doi: 10.1109/CVPR.2012.6247911. URL https://doi.org/10.1109/CVPR.2012.6247911.

Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. Technical Report CNS-TR-2007-001, California Institute of Technology, 2007. URL https://authors.library.caltech.edu/records/5sv1j-ytw97.

Kaiqi Guan, Wenke Huang, Xianda Guo, Yueyang Yuan, Bin Yang, and Mang Ye. Rethinking fair federated learning from parameter and client view. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0630. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/1b1205172b8a57ad504002a081b37859-Abstract-Conference.html.

Faisal Hamman and Sanghamitra Dutta. Demystifying local & global fairness trade-offs in federated learning using partial information decomposition. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1033880329f87ab8302b4fb39a52aebc-Abstract-Conference.html.

Jie Hao, Yuman Wu, Ali Payani, Myungjin Lee, and Mingrui Liu. Personalized federated fine-tuning for heterogeneous data: An automatic rank learning approach via two-level LoRA. *arXiv preprint arXiv:2503.03920*, 2025.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/6a9659feb1216f14f7384ba499518b38-Abstract.html.

Jialuo He, Wei Chen, and Xiaojin Zhang. FedAA: A reinforcement learning perspective on adaptive aggregation for fair and robust federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17085–17093, 2025. doi: 10.1609/aaai.v39i16.33878. URL https://doi.org/10.1609/aaai.v39i16.33878.

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*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Mehdi Karami and Amin Karami. Harmony in federated learning: A comprehensive review of techniques to tackle heterogeneity and non-IID data. *Cluster Computing*, 28(9):570, 2025. doi: 10.1007/s10586-025-05250-y.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5132–5143. PMLR, 2020.

Ken Lang. NewsWeeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*, pp. 331–339. Elsevier, 1995.

Agostina J. Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H. Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020. doi: 10.1073/pnas.1919012117.

Jie Li, Tianqing Zhu, Wei Ren, and Kim-Kwang Raymond Choo. Improve individual fairness in federated learning via adversarial training. *Computers & Security*, 132:103336, 2023. doi: 10.1016/j.cose.2023.103336. URL https://doi.org/10.1016/j.cose.2023.103336.

Minghan Li, Congcong Wen, Yu Tian, Min Shi, Yan Luo, Hao Huang, Yi Fang, and Mengyu Wang. Fairfedmed: Benchmarking group fairness in federated medical imaging with fairlora. *IEEE Transactions on Medical Imaging*, 45(4):1337–1351, 2026. doi: 10.1109/TMI.2025.3622522. URL https://doi.org/10.1109/TMI.2025.3622522.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10713–10722, 2021a.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020a. doi: 10.1109/MSP.2020.2975749.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pp. 429–450, 2020b.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2020c. URL https://openreview.net/forum?id=ByexElSYDr.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6357–6368. PMLR, 2021b. URL https://proceedings.mlr.press/v139/li21h.html.

Xiaoxiao Li, Yufeng Gu, Nicha Dvornek, Lawrence H. Staib, Pamela Ventola, and James S. Duncan. Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: Abide results. *Medical Image Analysis*, 65:101765, 2020d. doi: 10.1016/j.media.2020.101765.

Feng Lin, Chaoyue Zhao, Xiaoning Qian, Kendra Vehik, and Shuai Huang. Fair collaborative learning (FairCL): A method to improve fairness amid personalization. *INFORMS Journal on Data Science*, 4(1):67–84, 2025. doi: 10.1287/ijds.2024.0029.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Zili Lu, Heng Pan, Yueyue Dai, Xueming Si, and Yan Zhang. Federated learning with non-iid data: A survey. *IEEE Internet of Things Journal*, 11(11):19188–19209, 2024. doi: 10.1109/JIOT.2024.3376548.

Xiaodong Ma, Jia Zhu, Zhihao Lin, Shanxuan Chen, and Yangjie Qin. A state-of-the-art survey on solving non-iid data in federated learning. *Future Generation Computer Systems*, 135:244–258, 2022. doi: 10.1016/j.future.2022.05.003.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4615–4625. PMLR, 2019.

Noorain Mukhtiar, Adnan Mahmood, and Quan Z. Sheng. Fairness in federated learning: Trends, challenges, and opportunities. *Advanced Intelligent Systems*, 7(6):2400836, 2025. doi: 10.1002/aisy.202400836.

Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and H. Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=LkFG3lB13U5.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Computer Vision – ECCV 2010*, pp. 213–226. Springer, 2010. doi: 10.1007/978-3-642-15561-1_16.

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y. Chen, and Marzyeh Ghassemi. CheXclusion: Fairness gaps in deep chest x-ray classifiers. In *Pacific Symposium on Biocomputing*, volume 26, pp. 232–243, 2021a. doi: 10.1142/9789811232701_0022.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27:2176–2182, 2021b. doi: 10.1038/s41591-021-01595-0.

Micah J. Sheller, Brandon Edwards, G. Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R. Colen, and Spyridon Bakas. Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10:12598, 2020. doi: 10.1038/s41598-020-69250-1.

Ha Min Son, Moon Hyun Kim, and Tai-Myoung Chung. Comparisons where it matters: Using layer-wise regularization to improve federated learning on heterogeneous data. *Applied Sciences*, 12(19):9943, 2022. doi: 10.3390/app12199943. URL https://www.mdpi.com/2076-3417/12/19/9943.

Cong Su, Guoxian Yu, Jun Wang, Hui Li, Qingzhong Li, and Han Yu. Multi-dimensional fair federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15083–15090, 2024a. doi: 10.1609/aaai.v38i13.29430. URL https://doi.org/10.1609/aaai.v38i13.29430.

Shangchao Su, Mingzhao Yang, Bin Li, and Xiangyang Xue. Federated adaptive prompt tuning for multi-domain collaborative learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15117–15125, 2024b. doi: 10.1609/aaai.v38i13.29434.

Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving LoRA in privacy-preserving federated learning. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4e243e95c913b367775d71d7182b99d9-Abstract-Conference.html.

Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. FedProto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8432–8440, 2022.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, 2018. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446/.

Feifei Wang, Junni L. Zhang, Yichao Li, Ke Deng, and Jun S. Liu. Bayesian text classification and summarization via a class-specified topic model. *Journal of Machine Learning Research*, 22(89):1–48, 2021a. URL https://jmlr.org/papers/v22/18-332.html.

Tongnian Wang, Kai Zhang, Jiannan Cai, Yanmin Gong, Kim-Kwang Raymond Choo, and Yuanxiong Guo. Analyzing the impact of personalization on fairness in federated learning for healthcare. *Journal of Healthcare Informatics Research*, 8(2):181–205, 2024. doi: 10.1007/s41666-024-00164-7. URL https://doi.org/10.1007/s41666-024-00164-7.

Zheng Wang, Xiaoliang Fan, Jianzhong Qi, Chenglu Wen, Cheng Wang, and Rongshan Yu. Federated learning with fair averaging. *arXiv preprint arXiv:2104.14937*, 2021b.

Huijun Xing, Rui Sun, Jinke Ren, Jun Wei, Chun-Mei Feng, Xuan Ding, Zilu Guo, Yu Wang, Yudong Hu, Wei Wei, Xiaohua Ban, Chuanlong Xie, Yu Tan, Xian Liu, Shuguang Cui, Xiaohui Duan, and Zhen Li. Achieving flexible fairness metrics in federated medical imaging. *Nature Communications*, 16:3342, 2025. doi: 10.1038/s41467-025-58549-0.

Kunda Yan, Sen Cui, Abudukelimu Wuerkaixi, Jingfeng Zhang, Bo Han, Gang Niu, Masashi Sugiyama, and Changshui Zhang. Balancing similarity and complementarity for federated learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 55739–55758. PMLR, 2024a. URL https://proceedings.mlr.press/v235/yan24a.html.

Yunlu Yan, Lei Zhu, Yuexiang Li, Xinxing Xu, Rick Siow Mong Goh, Yong Liu, Salman Khan, and Chun-Mei Feng. A new perspective to boost performance fairness for medical federated learning. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pp. 13–23. Springer Nature Switzerland, 2024b. doi: 10.1007/978-3-031-72117-5_2.

Yunlu Yan, Chun-Mei Feng, Wangmeng Zuo, Rick Siow Mong Goh, Yong Liu, and Lei Zhu. Federated residual low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/906c860f1b7515a8ffec02dcdac74048-Paper-Conference.pdf.

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, 2024a. doi: 10.1145/3625558.

Rui Ye, Rui Ge, Xinyu Zhu, Jingyi Chai, Yaxin Du, Yang Liu, Yanfeng Wang, and Siheng Chen. Fedllm-bench: Realistic benchmarks for federated learning of large language models. *arXiv preprint arXiv:2406.04845*, 2024b. URL https://arxiv.org/abs/2406.04845.

Rui Ye, Wenhao Wang, Jingyi Chai, Dihan Li, Zexi Li, Yinda Xu, Yaxin Du, Yanfeng Wang, and Siheng Chen. OpenFedLLM: Training large language models on decentralized private data via federated learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6137–6147, 2024c. doi: 10.1145/3637528.3671582.

Wentao Yue, Tianyou Lai, Qingyu Mao, Qilei Li, and David Camacho. A review of federated learning under data heterogeneity. *Expert Systems*, 43(6):e70271, 2026. doi: 10.1111/exsy.70271.

Xubo Yue, Maher Nouiehed, and Raed Al Kontar. GIFAIR-FL: A framework for group and individual fairness in federated learning. *INFORMS Journal on Data Science*, 2(1):10–23, 2023. doi: 10.1287/ijds.2022.0022.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 325–333. PMLR, 2013. URL https://proceedings.mlr.press/v28/zemel13.html.

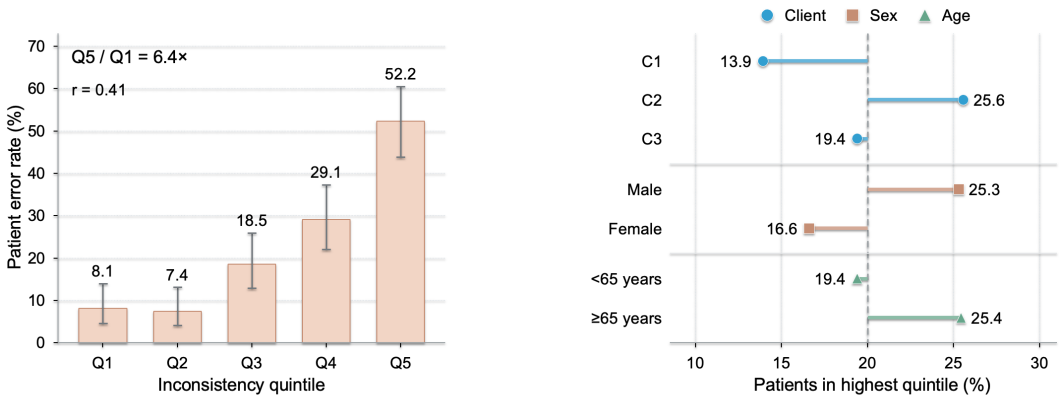
Jiale Zhang, Ye Li, Di Wu, Yanchao Zhao, and Shivakumara Palaiahnakote. SFFL: Self-aware fairness federated learning framework for heterogeneous data distributions. *Expert Systems with Applications*, 269:126418, 2025a. doi: 10.1016/j.eswa.2025.126418.

Li Zhang, Chaochao Chen, Zhongxuan Han, Qiyong Zhong, and Xiaolin Zheng. LoGoFair: Post-processing for local and global fairness in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22470–22478, 2025b. doi: 10.1609/aaai.v39i21.34404.

Yifei Zhang, Dun Zeng, Jinglong Luo, Xinyu Fu, Guanzhong Chen, Zenglin Xu, and Irwin King. A survey of trustworthy federated learning: Issues, solutions, and challenges. *ACM Transactions on Intelligent Systems and Technology*, 15(6):1–47, 2024. doi: 10.1145/3678181.

Hangyu Zhu, Jinjin Xu, Shiqing Liu, and Yaochu Jin. Federated learning on non-iid data: A survey. *Neurocomputing*, 465:371–390, 2021. doi: 10.1016/j.neucom.2021.07.098.

Yongshuo Zong, Yongxin Yang, and Timothy Hospedales. MEDFAIR: Benchmarking fairness for medical imaging. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6ve2CkeQe5S.



(a) Inconsistency and prediction errors (b) Clients and demographic groups

Figure 5: Relationship between instance-level inconsistency, prediction errors, and fairness on the medical dataset. (a) Patient error rates across inconsistency quintiles. (b) Proportions of patients in the highest-inconsistency quintile across clients, sex groups, and age groups.

A Empirical Analysis

Figure 1 motivates our instance-level perspective by showing that balanced client-level performance can coexist with inconsistent predictions among similar instances, while instance consistency is closely related to predictive performance and fairness. Medical FL often involves heterogeneous data distributions across hospitals or clinical sites (Wang et al., 2024; Yan et al., 2024b; Sheller et al., 2020; Li et al., 2020d). It can also exhibit demographic disparities associated with attributes such as age and sex (Xing et al., 2025; Li et al., 2026; Larrazabal et al., 2020; Zong et al., 2023). To provide quantitative evidence for this observation, we analyze the relationship between instance-level inconsistency, prediction errors, and cross-group heterogeneity on the medical dataset based on the predictions of FedAvg.

Patients are ranked globally according to their confidence-inconsistency scores and divided into five equal-sized quintiles, where Q1 and Q5 denote the lowest- and highest-inconsistency groups, respectively. Since Q5 contains the 20% most inconsistent patients overall, 20% serves as the reference proportion in Figure 5(b).

As shown in Figure 5(a), prediction errors generally increase with instance-level inconsistency. The patient error rate rises from 8.1% in Q1 to 52.2% in Q5, with the latter being approximately 6.4× higher. This result reveals a clear association between instance-level inconsistency and predictive errors.

Figure 5(b) further shows that highly inconsistent patients are not uniformly distributed across clients and demographic groups. The proportions of patients falling into Q5 vary noticeably across clients, sex groups, and age groups, indicating that the prevalence of high inconsistency differs across these groups. Together, these patterns connect instance-level inconsistency with both prediction errors and cross-group heterogeneity, providing empirical support for using instance consistency as a common signal for predictive performance and fairness.

B Experimental Details

Our NLU experimental setup follows FRLoRA (Yan et al., 2025), which adopts the empirical federated settings of (Ye et al., 2024b;c). The Office-Caltech-10 experiments follow FedPW (Guan et al., 2025), while the medical image classification experiments are conducted on a private brain MRI dataset. Detailed information about the three experimental settings is provided below.

B.1 Datasets

NLU Tasks. The NLU experiments are conducted on six benchmarks covering different language understanding tasks:

Table 6: Detailed information of the NLU datasets.

Datasets	**RTE**	**CoLA**	**QNLI**	**MRPC**	**MNLI**	**20NG**
Task	Inference	Linguistic acceptability	Inference	Paraphrase	Inference	Text classification
# Labels	2	2	2	2	3	20
# Train	2.49k	8.55k	105k	3.67k	393k	11.31k
Evaluation split	Validation	Validation	Validation	Validation	MNLI-m	Test
# Clients (Total)	5	5	5	5	5	5
# Clients (Train)	5	5	5	5	5	5
Metric	Accuracy	Matthews correlation	Accuracy	Accuracy	Accuracy	Accuracy

Table 7: Detailed information of the Office-Caltech-10 dataset.

Domain	**Amazon**	**DSLR**	**Caltech**	**Webcam**	**Overall**
Task	Object classification				
# Labels	10	10	10	10	10
# Images	958	157	1,123	295	2,533
# Clients (Total)	5	5	5	5	20
# Clients (Train)	5	5	5	5	20
Metric	Accuracy				

- **RTE** (Wang et al., 2018): A binary textual entailment task constructed from news and Wikipedia text, where the goal is to determine whether a given premise entails a hypothesis.
- **CoLA** (Wang et al., 2018): A linguistic acceptability task consisting of English sentences annotated according to whether they are grammatically acceptable.
- **QNLI** (Wang et al., 2018): A sentence-pair classification task derived from question answering, where the model determines whether a context sentence contains the answer to a given question.
- **MRPC** (Wang et al., 2018): A paraphrase identification task consisting of sentence pairs from news sources, labeled according to whether the two sentences are semantically equivalent.
- **MNLI** (Wang et al., 2018): A multi-genre natural language inference task that classifies the relationship between a premise and a hypothesis as entailment, contradiction, or neutral.
- **20NG** (Lang, 1995; Wang et al., 2021a): A document classification benchmark containing messages from 20 different newsgroups, where the task is to assign each document to its corresponding topic category.

As shown in Table 6, the 20NG dataset contains 20 categories and MNLI contains 3 categories, while RTE, CoLA, QNLI, and MRPC are binary classification tasks. Following Yan et al. (2025), the training set of each benchmark is partitioned across five clients using Dirichlet sampling ($D_k \sim \mathrm{Dir}(\beta_{\mathrm{Dir}})$), where a smaller β_{Dir} indicates stronger data heterogeneity. By default, β_{Dir} is set to 0.5, and all five clients participate in each communication round. We report Matthews correlation coefficient (MCC) for CoLA and accuracy for the remaining NLU tasks. # Train denotes the number of training samples.

Cross-Domain Image Classification. The cross-domain image classification experiments are conducted on Office-Caltech-10:

- **Office-Caltech-10** (Gong et al., 2012): A cross-domain object classification benchmark constructed from the 10 object categories shared by Office-31 (Saenko et al., 2010) and Caltech-256 (Griffin et al., 2007). It contains 2,533 images collected from four visual domains: Amazon, DSLR, Webcam, and Caltech. Although these domains share the same label space, they exhibit substantial differences in image appearance and acquisition conditions, making the dataset suitable for evaluating learning under cross-domain heterogeneity.

Office-Caltech-10 contains four domains with 10 shared object categories (Table 7). We follow FedPW (Guan et al., 2025) in both client allocation and cross-domain evaluation. Each domain is

Table 8: Detailed information of the private brain MRI dataset.

Client	C1	C2	C3	Overall
Task	Brain tumor classification			
# Labels	4	4	4	4
# Patients	948	1,137	1,336	3,421
# MRI Exams	3,078	3,631	4,277	10,986
# Train Patients	670	805	945	2,420
# Val Patients	91	109	128	328
# Test Patients	187	223	263	673
Sensitive Attributes	Sex and Age			
Metric	Accuracy			

assigned five clients, giving 20 clients in total. We report average accuracy across domains to measure predictive performance. We use the sample standard deviation of the four domain accuracies to assess cross-domain fairness.

Medical Image Classification. We further consider a private brain MRI dataset for four-class brain tumor classification, including pituitary tumors, schwannomas, gliomas, and meningiomas. As shown in Table 8, the dataset contains 10,986 MRI examinations from 3,421 patients distributed across three clients. The data are split at the patient level into training, validation, and test sets with a ratio of 70%/10%/20%, ensuring that examinations from the same patient remain in the same split. The number of patients in each split is fixed across random seeds, while the patient identities vary. Since different patients have different numbers of MRI examinations, the number of examinations in each split may vary slightly across runs.

Each MRI examination is represented using the T1AX sequence, and patient-level predictions are obtained by averaging the predicted probabilities across examinations from the same patient. The dataset also contains sex and age information, which enables additional evaluation of demographic group fairness.

B.2 Baselines

We compare FedICIL with representative methods covering standard FL, heterogeneous FL, client-fair FL, and group-fair FL.

Standard and Heterogeneous FL. **FedAvg** (McMahan et al., 2017) is the standard federated learning baseline that aggregates locally trained models according to client data sizes. **FedProx** (Li et al., 2020b) introduces a proximal term between the local and global models to alleviate client drift under heterogeneous data. **SCAFFOLD** (Karimireddy et al., 2020) uses control variates to correct local updates caused by heterogeneous client objectives. **FedAdagrad** (Reddi et al., 2021) incorporates adaptive server-side optimization into federated training.

Client-Fair FL. **AFL** (Mohri et al., 2019) optimizes an agnostic federated objective with an emphasis on poorly performing client distributions. **q-FFL** (Li et al., 2020c) assigns greater importance to clients with higher losses to improve client-level performance fairness. **Ditto** (Li et al., 2021b) combines global federated training with personalized local models. **FedHEAL** (Chen et al., 2024b) and **FedPW** (Guan et al., 2025) further target the trade-off between global performance and client-level fairness under heterogeneous data.

For Office-Caltech-10, we additionally compare with **FedFV** (Wang et al., 2021b), **FedCKA** (Son et al., 2022), **FedGCR** (Cheng et al., 2024), **FedSaC** (Yan et al., 2024a), and **FedAA** (He et al., 2025), following the set of representative heterogeneous and fairness-oriented FL baselines considered in the cross-domain experiments.

Group-Fair FL. For the medical image experiments, we additionally include two group-fair FL methods. **FairFed** (Ezzeldin et al., 2023) adapts aggregation according to group-fairness information, while **mFairFL** (Su et al., 2024a) jointly considers group- and client-level fairness. These methods allow us to further compare FedICIL under demographic fairness criteria defined by sex and age.

B.3 Implementation Details

For FedICIL, we set the update ratio ρ, the instance-consistency coefficient β, and the neighborhood size H according to the task-specific configurations. We fix $\gamma = 1$ in all experiments.

NLU Tasks. RoBERTa-base (Liu et al., 2019) is used with LoRA (Hu et al., 2022), where the rank is set to $r = 16$ and the scaling factor to $\alpha = 32$. All methods are trained for 200 communication rounds with a batch size of 32 and a learning rate of 2×10^{-4} using cosine annealing. The maximum sequence length is 256. FedICIL uses the same parameter update budget as LoRA for a fair comparison. For FedICIL, the neighborhood size is set to $H = 20$ and the instance-consistency coefficient is set to $\beta = 1$ for all NLU tasks.

Cross-Domain Image Classification. Following FedPW (Guan et al., 2025), ResNet-10 is trained for 200 communication rounds with 10 local epochs per round. Optimization uses SGD with a learning rate of 0.001, momentum of 0.9, and a batch size of 16. For FedICIL, we set the update ratio to $\rho = 0.50$, the instance-consistency coefficient to $\beta = 0.75$, and the neighborhood size to $H = 5$. We repeat each experiment with three different random seeds. We also follow FedPW's evaluation protocol, averaging the results over the final five communication rounds of each run and then across the three runs.

Medical Image Classification. An ImageNet-pretrained ResNet-50 is used as the backbone. All methods are trained for 50 communication rounds with one local epoch per round using SGD with a learning rate of 0.001 and a batch size of 32. For FedICIL, we set the update ratio to $\rho = 0.75$, the instance-consistency coefficient to $\beta = 1.0$, and the neighborhood size to $H = 5$. MRI examinations from the same patient are additionally excluded from the neighborhood candidates. We repeat each experiment with three different random seeds. For each run, we select the checkpoint with the best validation performance and evaluate it on the test set. We report the mean test results across the three runs.

C Additional Experiments

C.1 Additional Results on NLU Tasks

Table 9 reports the final results on QNLI and 20NG, including accuracy, client accuracy standard deviation, and IUR. Fig. 6 presents the round-wise training curves on MNLI, MRPC, and QNLI, including global accuracy, client accuracy standard deviation, worst-client accuracy, and max–min accuracy gap. Fig. 7 illustrates the IUR distributions on MNLI, QNLI, and 20NG.

As shown in Table 9, FedICIL improves five of the six metrics over FedAvg and achieves the **best Mean Rank of 2.50**. On 20NG, all evaluated alternatives to FedAvg obtain lower IUR†. 20NG contains semantically related categories (Wang et al., 2021a). For FedICIL, the consistency objective aligns target-class confidence with neighborhood consensus, but does not directly constrain non-target probabilities. Similar instances can still distribute probability mass differently among related non-target classes, and these differences are captured by IUR. This objective–metric mismatch offers a possible explanation for the gains in accuracy and client balance without a corresponding improvement in IUR†.

C.2 Hyperparameter Sensitivity

We vary the update ratio ρ from 10% to 100% on RTE, selecting parameters by their sensitivity scores. Figure 8 shows that the 40% setting achieves higher accuracy and lower client Std than the 100% setting. Restricting updates to a sensitivity-selected subset can therefore improve both measures while using a smaller budget. The 90% setting further reduces Std with comparable accuracy, offering a different balance between update budget and client disparity. Meanwhile, the 20% and 80% settings exhibit low IUR but poorer accuracy and client balance. Thus, a low consistency residual alone does not identify a suitable update ratio; predictive performance and client disparity must also be considered.

Table 9: Additional results on QNLI and 20NG in terms of accuracy, client standard deviation (Std), and IUR†. Changes relative to FedAvg are shown in the second line. Mean Rank is computed over the six metrics, where lower is better. The best results are marked in **bold**, and the second-best results are underlined.

Method	QNLI			20NG			Mean Rank ↓
	Acc ↑	Std ↓	IUR† ↑	Acc ↑	Std ↓	IUR† ↑	
FedAvg	88.19	1.61	2.86	69.62	2.36	**6.78**	4.17
FedAdagrad	83.87 ↓4.32	1.28 ↑0.33	3.24 ↑0.38	68.63 ↓0.99	4.05 ↓1.69	5.91 ↓0.87	6.33
FedProx	88.03 ↓0.16	1.27 ↑0.34	3.00 ↑0.13	69.86 ↑0.24	2.49 ↓0.13	6.54 ↓0.24	3.92
SCAFFOLD	86.11 ↓2.08	2.80 ↓1.19	3.11 ↑0.24	69.04 ↓0.58	5.04 ↓2.68	1.68 ↓5.10	7.83
q-FFL	**89.29** ↑1.10	1.10 ↑0.51	3.00 ↑0.14	69.40 ↓0.22	2.28 ↑0.08	1.74 ↓5.04	3.92
AFL	87.68 ↓0.51	3.53 ↓1.92	3.39 ↑0.53	69.05 ↓0.57	2.50 ↓0.14	6.76 ↓0.02	5.33
Ditto	87.11 ↓1.08	2.30 ↓0.69	2.65 ↓0.22	69.42 ↓0.20	2.29 ↑0.07	1.73 ↓5.05	6.33
FedHEAL	87.20 ↓0.99	2.85 ↓1.24	3.41 ↑0.54	68.32 ↓1.30	4.38 ↓2.02	1.79 ↓4.99	7.00
FedPW	81.75 ↓6.44	**1.05** ↑0.56	2.16 ↓0.70	68.93 ↓0.69	3.75 ↓1.39	1.67 ↓5.11	7.67
FedICIL	89.08 ↑0.89	1.55 ↑0.06	**4.19** ↑1.33	**70.33** ↑0.71	**2.26** ↑0.10	3.92 ↓2.86	**2.50**

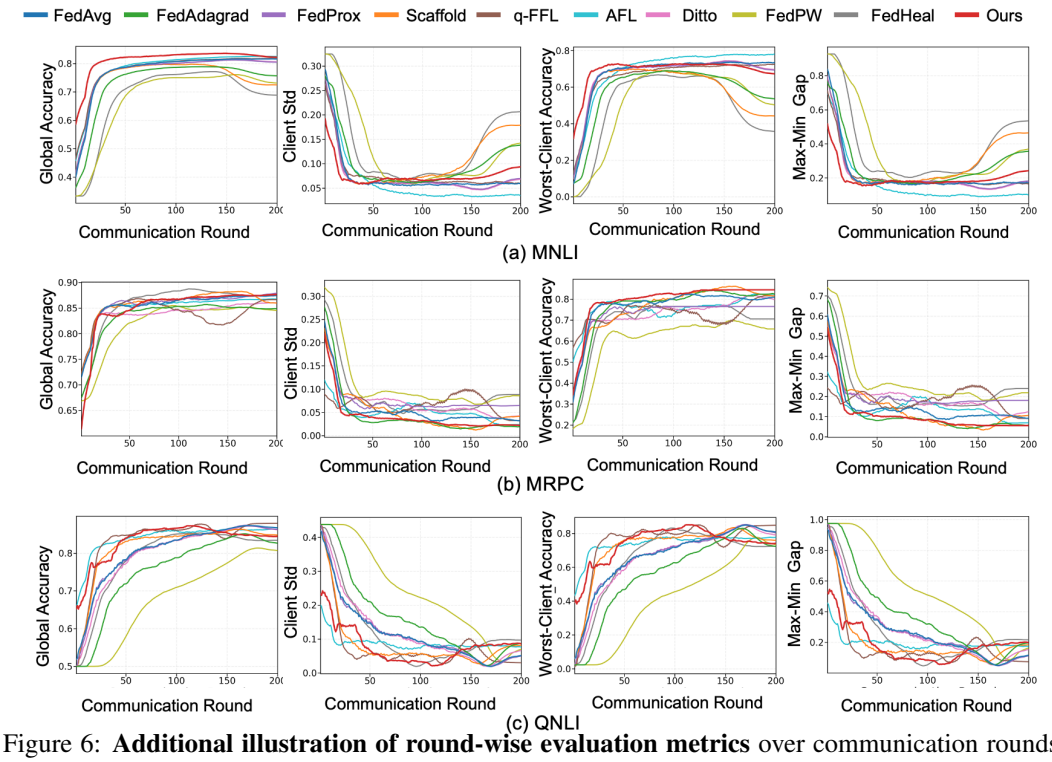


Figure 6: **Additional illustration of round-wise evaluation metrics** over communication rounds on (a) MNLI, (b) MRPC and (c) QNLI, covering global accuracy, client accuracy standard deviation, worst-client accuracy, and max–min accuracy gap.

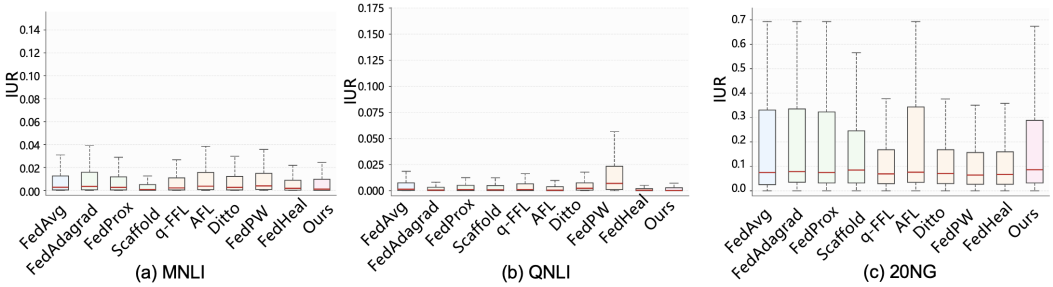


Figure 7: **Illustration of IUR distributions** on (a) MNLI, (b) QNLI, and (c) 20NG.

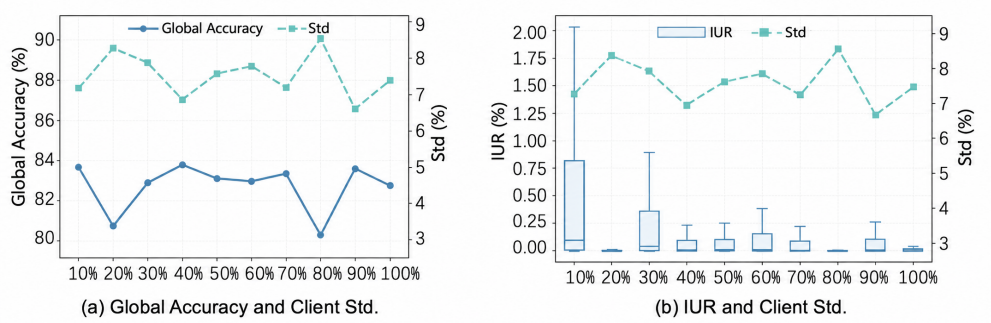


Figure 8: **Performance and fairness** under different mask ratios.