# Local or Global: Selective Knowledge Assimilation for Federated Learning with Limited Labels

Yae Jee Cho†
Carnegie Mellon University
Pittsburgh, PA
yaejeec@andrew.cmu.edu

Gauri Joshi
Carnegie Mellon University
Pittsburgh, PA
gaurij@andrew.cmu.edu

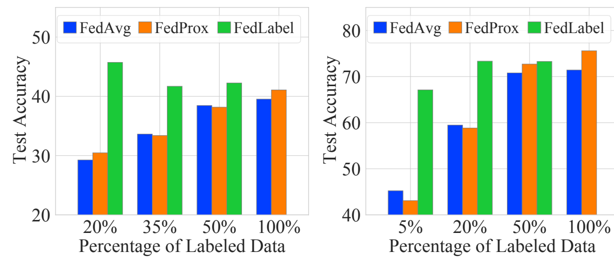
Dimitrios Dimitriadis†
Amazon
Seattle, WA
dbdim@amazon.com

## Abstract

*Many existing FL methods assume clients with fully-labeled data, while in realistic settings, clients have limited labels due to the expensive and laborious process of labeling. Limited labeled local data of the clients often leads to their local model having poor generalization abilities to their larger unlabeled local data, such as having class-distribution mismatch with the unlabeled data. As a result, clients may instead look to benefit from the global model trained across clients to leverage their unlabeled data, but this also becomes difficult due to data heterogeneity across clients. In our work, we propose* FedLabel *where clients selectively choose the local or global model to pseudo-label their unlabeled data depending on which is more of an expert of the data. We further utilize both the local and global models' knowledge via global-local consistency regularization which minimizes the divergence between the two models' outputs when they have identical pseudo-labels for the unlabeled data. Unlike other semi-supervised FL baselines, our method does not require additional experts other than the local or global model, nor require additional parameters to be communicated. We also do not assume any server-labeled data or fully labeled clients. For both cross-device and cross-silo settings, we show that* FedLabel *outperforms other semi-supervised FL baselines by* 8-24%*, and even outperforms standard fully supervised FL baselines (*100% *labeled data) with only* 5-20% *of labeled data.*

## 1. Introduction

Federated learning (FL) [27] enables collaborative learning across clients without explicit disclosure of their local data [18, 38]. In FL, a server updates its global model by aggregating the local gradients obtained from clients' training on their datasets. These clients can be a number of edge-devices such as cell-phones (cross-device) [44] or a handful of hospitals, for example, willing to train a model for disease prediction without sharing patients' private data (cross-silo) [32]. While FL can indeed allow clients to train a single global model without private data sharing, a crucial yet often overlooked limitation found in realistic FL scenarios is that labels can be scarce [13, 9, 2, 21, 34]. In cross-device settings, owners of edge-devices rarely go through the effort of labeling all of their local data such as photos, resulting in only a few labeled samples and a large volume of unlabeled data. Similarly in cross-silo settings, such as hospitals collaborating for predicting diseases, labeling is often a laborious process where healthcare experts are required to process volumes of patients' data [15, 12]. Such scarcity of labels can lead to severe performance degadation as shown in Fig. 1.




(a) CIFAR10 (Cross-Device) (b) OrganAMNIST (Cross-Silo)

Figure 1: Test accuracy of the global model for varying amount of labeled local data in each client. The fewer the labeled data, the lower the test accuracy for standard FL algorithms (FedAvg, FedProx) while our proposed FedLabel performs significantly higher. For CIFAR10, FedLabel achieves an even higher test accuracy than when 100% of labels are used.

A naïve approach to tackle label scarcity in clients is using standard semi-supervised learning (SSL) methods devised for general machine learning (ML) applications at each client, using its own local data. For example, consistency regularization methods [41] or pseudo-labeling methods [35] can be directly used with each client's local data with its local model. However, with limited labels, the feature-label pair distribution of the labeled data can be different from that of the unlabeled data as shown in Fig. 2 and previous work [47, 3]. We call this difference between the two distributions as *class distribution mismatch*. As such, the limited

[†]Work done while at Microsot Research.

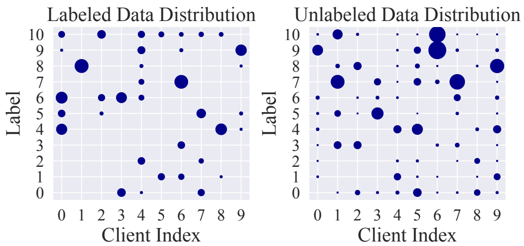



Figure 2: Class distribution mismatch (CIFAR10) is shown just by limiting the number of labeled data to 20% without artificially biasing the labeled data distribution.

labeled data can poorly generalize to a large number of unlabeled local data. This is also shown in Table 1, where the 'Only Local' performance largely degrades for a smaller number of labeled data. Due to these scenarios, *leveraging only the local knowledge of a client may not be enough to fully utilize its unlabeled data*. Thus, in FL, a client can seek to utilize the global knowledge shared across clients through the aggregation of the local updates to the global model. The problem, however, with only using the global knowledge for leveraging the unlabeled data at the clients is that the data distribution amongst clients' local data is heterogeneous [33, 29, 11, 26, 20, 39]. Due to this *data heterogeneity*, for clients whose local data distribution differs from the overall global data distribution, the global model may also not be useful in assisting the clients to leverage their unlabeled data as shown in Table 1, 'Only Global' case in which the test accuracy is lower than selectively leveraging both global and local models.

Based on the observations above, either the local or global model, or both, can be useful for clients to leverage their unlabeled data depending on the labeled data's generalizability to the unlabeled data and clients' data heterogeneity. Therefore, to utilize the setting of FL where clients have access to both the knowledge from their local data and the global model, we propose a selective knowledge assimilation method named FEDLABEL where each client chooses between its local and global model to pseudo-label its unlabeled data based on each model's confidence score. Moreover, with our proposed global-local consistency regularization, we fully utilize both the local and global models when both have useful knowledge of the unlabeled data.

Most relevant line of recent work to FEDLABEL has proposed the server to identify multiple experts for each client. The experts can be other clients with similar data distributions [17], or the local, global, and the mixture of the local and global models by model splicing [2]. However, the additional computational and communication overhead for the server to find and send the appropriate experts for each client can become exponentially costly with the increasing participating clients [40]. Moreover, these works treat all experts equally, taking an average of their knowledge for leveraging each client's unlabeled data. Due to this, we show

| High Data Heterogeneity | 20% Labeled Data | 50% Labeled Data |
|---|---|---|
| Only Local | 32.57 (±2.20) | 41.21 (±1.80) |
| Only Global | 37.51 (±1.80) | 38.43 (±0.94) |
| Global+Local (**ours**) | **44.51** (±1.85) | **50.53** (±1.74) |

Table 1: Test acc. (CIFAR10) when only the local or global model is used for pseudo-labeling which largely underperforms the case when both models are selectively used by our proposed FEDLABEL.

that these methods' performance degrades significantly (see Section 5.2) when the number of labels decreases and the data heterogeneity gets higher. In our work, *we use only the natural two experts that are available in FL, local and global, and show that this is enough to fully utilize the unlabeled data* at the clients when executed properly for *both high and lower label scarcity and data heterogeneity cases*. Other related work proposes methods with restrictive assumptions such as the server having labeled data that is similar to the data distribution of the clients [8] or several clients having fully labeled data [23]. In FL, the server does not have access to client data and labels are scarce, making these assumptions practically improbable.

As summarized in Table 2, previous work in SSFL: i) assumes restrictive settings such as the server or several clients having good-enough fully labeled data, ii) imposes additional computational/communication burden at the server to find and send more experts other than the naturally occurring local and global models, and iii) does not consider the poor generalization of limited data to the unlabeled data such as class distribution mismatch. Improving on these drawbacks, we propose our novel method FEDLABEL, which:

- Is robust to both the limited generalizability of the labeled data such as class distribution mismatch and data heterogeneity by using just two experts, global and local, to leverage a large number of unlabeled data (80-95%) with just a few labeled data (5-20%).
- Does not require the server having any labeled data or a few clients to have fully labeled data. It also does not require additional experts to be computed or communicated other than the local and global model used in standard FL algorithms [27, 33].
- Leverages unlabeled data by adaptively choosing either the local or global model based on the confidence of the model's prediction for pseudo-labeling with our proposed global-local consistency regularization that minimizes the divergence between the models' outputs when their pseudo-labels are identical.
- Achieves 8-24% test accuracy improvement compared to the other SSFL baselines, and even achieves a higher test accuracy than fully supervised scenario (100% labeled data) with only using 5-20% of labels for extensive experiments (3 tasks for cross-device and 2 tasks for cross-silo).

| Method | Requires Server Labeled Data | Requires Fully Labeled Clients | Requires Additionally Computed Expert(s) | Requires Additional Comm.- of Params. | Robust to Class-Dist. Mismatch |
|---|---|---|---|---|---|
| SemiFL [8] | Yes | No | No | No | No |
| Rscfed [23] | No | Yes | No | No | No |
| FedTriNet [2] | No | No | Yes | Yes | No |
| FedMatch [17] | No | No | Yes | Yes | No |
| **FedLabel (ours)** | **No** | **No** | **No** | **No** | **Yes** |

Table 2: Comparison of related work with FEDLABEL.

## 2. Related Work

**Semi-Supervised Learning for General ML.** Labels are required to train well-performing models for classification tasks, but realistically labels are often scarce and expensive to obtain [25, 24, 7, 31, 1]. To tackle such label scarcity for general ML problems, there has been a wide range of works including methods such as consistency regularization [41, 16, 36], pseudo-labeling [35, 46, 42], virtual adversarial training [28], and per-sample weighting of unlabeled data [30]. As one of the most popular methods, consistency regularization leverages unlabeled data by minimizing a model's prediction difference between the original image and perturbed versions of the image [41, 16, 36]. Another commonly used method is pseudo-labeling, a simple approach to apply thresholding to the max probability of the prediction and provide a pseudo label for a data sample [35]. Pseudo-labeling has also been popularly used in variations such as dynamic thresholding based on the sample's relatedness to the labeled data [42], or class-based thresholding [46]. However, the aforementioned work does not jointly consider the poor generalizability that the limited labeled data can have to the unlabeled data and the data heterogeneity across clients. We show in Section 5.2 that naively combining these SSL methods to FL significantly fails, while FEDLABEL is robust to both limited labeled data and data heterogeneity.

**Semi-Supervised Learning for FL.** Recently, several SSFL methods have been proposed such as utilizing inter-client consistency across clients [17], leveraging the labeled data at the server or several clients for communication-efficient or personalized SSL for FL [8, 13, 23], or using node classification on graphs for handling data with new label domains [37]. FedMatch [17] and FedTriNet [2] are the closest to our work. In FedMatch, the server finds clients with similar data distributions for each client and sends the predictions from similar data clients for inter-client consistency regularization. In FedTriNet, each client handles three models (local, global, and the combination of the two through model splicing) to leverage the unlabeled data through a simple average of predictions across the three experts, and also requires the clients to send back their loss values to the server. Both approaches not only impose additional communication and computational costs for the clients and server but also their performance highly depends on finding the right parameter such as the number of helper clients or how to splice the models. Moreover, both methods do not consider weighing the experts differently to further combat limited data generalizability and data heterogeneity. We show in Section 5.2 that both FedMatch and FedTrinet indeed fail further for a lower percentage of labeled data and higher data heterogeneity, while FEDLABEL in fact performs even better in these cases by simply leveraging only the local and global models.

## 3. Problem Formulation

We consider a FL setup (cross-silo and cross-device) where $M$ clients are connected to a central server to train a well-performing global model for a $N$-class classification task. Each client's local data, denoted as $\mathcal{D}_k$ , $k \in [M]$, is consisted of the labeled set $(\mathbf{x}, y) \in \mathcal{D}_{L,k}$ and unlabeled set $\xi \in \mathcal{D}_{U,k}$, i.e., $\mathcal{D}_k = \mathcal{D}_{L,k} \bigcup \mathcal{D}_{U,k}$ where $\mathbf{x}, \ \xi \in \mathbb{R}^d$ is the input and $y \in [1, N]$ is the label.

**Conventional FL Algorithms' Assumption.** Previous work in FL [19, 29] assumes an ideal scenario where each client $k$ has its unlabeled data $\mathcal{D}_{U,k}$ labeled which we call as the *hypothetically labeled* unlabeled dataset denoted as $\mathcal{D}_{\overline{U},k}$. With the 'hypothetical' fully labeled dataset of each client $k$ denoted as $\overline{\mathcal{D}}_k = \mathcal{D}_{L,k} \bigcup \mathcal{D}_{\overline{U},k}$, conventional FL algorithms assumes fully labeled data where the server aims to find the model parameter $\mathbf{w} \in \mathbb{R}^q$ that minimizes:

$$F(\mathbf{w}) = \sum_{k=1}^{M} p_k F_k(\mathbf{w}), \ F_k(\mathbf{w}) := \frac{1}{|\overline{\mathcal{D}}_k|} \sum_{\xi \in \overline{\mathcal{D}}_k} f(\mathbf{w}, \xi)$$

where $p_k$ is the aggregating weight and $f(\mathbf{w}, \xi)$ is the loss function for sample $\xi := (\mathbf{x}, y)$ and parameter vector $\mathbf{w}$.

**Realistic FL with Limited Labeled Data.** In practice, much of the available local data may not have ground-truth labels. In fact, the number of unlabeled data can be much larger than the labeled data, i.e., $|\mathcal{D}_{L,k}| \ll |\mathcal{D}_{U,k}|$. Then the server can only use the labeled data and effectively minimize

$$F_{\mathcal{L}}(\mathbf{w}) = \sum_{k=1}^{M} p_k F_{\mathcal{L},k}(\mathbf{w}), \ F_{\mathcal{L},k}(\mathbf{w}) := \frac{1}{|\mathcal{D}_{L,k}|} \sum_{\xi \in \mathcal{D}_{L,k}} f(\mathbf{w}, \xi)$$

where $\mathbf{w}^*_{\mathcal{L}} = \arg\min_{\mathbf{w}} F_{\mathcal{L}}(\mathbf{w})$ becomes more different to the solution of the ideal objective $\mathbf{w}^* = \arg\min_{\mathbf{w}} F(\mathbf{w})$ as the distribution of $\bigcup_{k=1}^{M} \mathcal{D}_{L,k}$ differs more from the hypothetical ideal dataset $\bigcup_{k=1}^{M} \overline{\mathcal{D}}_k$. Our goal is to find an algorithm that can find the model parameter $\mathbf{w}^*$ by using only the labels from $\bigcup_{k=1}^{M} \mathcal{D}_{L,k}$ and the unlabeled data $\bigcup_{k=1}^{M} \mathcal{D}_{U,k}$.

With clients having only a few labeled data and a larger number of unlabeled data, the labeled data can have limited generalization properties to the large unlabeled data due to factors such as class-distribution mismatch (the discrepancy between the distribution $\mathcal{D}_{L,k}$ and $\mathcal{D}_{\overline{U},k}$) or the mere limited number of its samples. Moreover, data heterogeneity (the discrepancy across the distributions of the clients' local data $\overline{\mathcal{D}}_k,\ k \in [M]$) exacerbates the difficulty of SSFL. We show in our work that our proposed FedLabel enables clients to leverage their unlabeled data to the full extent just by selectively using the local and global model despite the difficulties caused by data heterogeneity and limited generalizability of the labeled data as explained in detail in the next section.

## 4. FedLabel: Choose Local or Global

In this section, we first introduce the novel semi-supervised loss function of FedLabel over the clients' unlabeled data and then present the end-to-end algorithm of FedLabel for implementation when used in realistic FL frameworks. An overview of how FedLabel leverages unlabeled data for each client $k \in [M]$ is in Fig. 3.

### 4.1. Semi-supervised Loss of FedLabel.

In the most commonly used vanilla FL [27], there are two natural sources the clients can learn from: *the global model* which is trained across different clients, and *the local model* which is trained further with local SGD with their own local data. Whether the local or global model, or even both, can be effective for labeling the unlabeled data depends on which is more knowledgeable on the data based on what each has learned. For instance, if a client has very few labeled data, leading to class distribution mismatch or limited generalization capability to the unlabeled data, it can perform badly at correctly matching the labels for the unlabeled data. Instead, the global model can be more effective in giving the correct labels for the unlabeled data since it has seen more data from different clients. On the other hand, if the local labeled data sufficiently generalizes well to the local unlabeled data, the local model is more likely to give correct pseudo labels. This is also observed in Table 1 where smaller number of labels leads to the 'Only Global' performing better, but for larger number of labels 'Only Local' performs better. There can also be cases where both the local and global models are not useful for leveraging the unlabeled data, and it is best not to use either model.

Based on this observation, we propose FedLabel that adaptively chooses either the local or global model based on the confidence score of each model's logits for pseudo-labeling. If both models' confidence does not exceed a certain threshold, we do not use that unlabeled data. Such binary choice of the model, however, can lead to losing relevant information from the other discarded model despite it having the same hard-label prediction as the chosen model. To assimilate both information from the local and global model for such scenarios, FedLabel adds global-local consistency regularization that minimizes the divergence between the local and global models' outputs. To the best of our knowledge, adaptive selection of the global and local model for pseudo-labeling and assimilating more knowledge, when needed, with global-local consistency regularization is a novel method that has not been previously proposed.

**1) Obtaining Global and Local Models (same as Standard FedAvg [27]).** With superscript $(t, r)$ denoting the communication round $t$ and local iteration $r$, for each $t$ the server selects a set of clients $\mathcal{C}^{(t,0)}$ uniformly at random and sends the global model $\mathbf{w}^{(t,0)}$ to clients in $\mathcal{C}^{(t,0)}$. The clients in $\mathcal{C}^{(t,0)}$ initialize their local model for supervised training as $\mathbf{w}_{\mathcal{L},k}^{(t,0)} = \mathbf{w}^{(t,0)}$ to perform $\tau$ local iterations with learning rate $\eta$ to obtain their respective *supervised* local models as:

(Perform Local SGD on *Labeled* Data)

$$\mathbf{w}_{\mathcal{L},k}^{(t,\tau)} = \mathbf{w}_{\mathcal{L},k}^{(t,0)} - \eta \sum_{l=0}^{\tau-1} \nabla F_{\mathcal{L},k}(\mathbf{w}_{\mathcal{L},k}^{(t,l)}, \xi_{\mathcal{L},k}^{(t,l)}) \tag{1}$$

(Compute the *Supervised* Local Update)

$$\Delta \mathbf{w}_{\mathcal{L},k}^{(t,0)} = \mathbf{w}_{\mathcal{L},k}^{(t,\tau)} - \mathbf{w}_{\mathcal{L},k}^{(t,0)} \tag{2}$$

where $\nabla F_{\mathcal{L},k}(\mathbf{w}_{\mathcal{L},k}^{(t,l)}, \xi_{\mathcal{L},k}^{(t,l)}) = \frac{1}{b} \sum_{\xi \in \xi_{\mathcal{L},k}^{(t,l)}} \nabla f(\mathbf{w}_{\mathcal{L},k}^{(t,l)}, \xi)$ is the stochastic gradient computed with mini-batch $\xi_{\mathcal{L},k}^{(t,l)}$ of size $b$ randomly sampled from $\mathcal{D}_{L,k}$. Note that FedLabel does not alter the local update procedure used in standard FL algorithms [27, 33], and can be easily extended to using different methods to obtain the local model $\mathbf{w}_{\mathcal{L},k}^{(t,\tau)}$.

**2) Confidence-based Selection with Thresholding.** After performing local SGD on the labeled data as in (1), clients have the global model $\mathbf{w}^{(t,0)}$ and the local model trained with labeled data $\mathbf{w}_{\mathcal{L},k}^{(t,\tau)}$. Each client $k \in \mathcal{C}^{(t,0)}$ gets the logits from each the global and local model from the unlabeled data $\xi \in \mathcal{D}_{U,k}$ denoted as $\mathbf{s}(\mathbf{w}^{(t,0)},\ \xi)$ and $\mathbf{s}(\mathbf{w}_{\mathcal{L},k}^{(t,\tau)},\ \xi)$ respectively where $\mathbf{s}(\cdot,\cdot) : \mathbb{R}^q \times \mathbb{R}^d \to \mathbb{R}^{N\times 1}$. We then use function $h(\cdot) : \mathbb{R}^{N\times 1} \to \mathbb{R}$ that calculates the confidence-score (variance) of the logits and select the logit with the higher confidence-score to pseudo-label the unlabeled data. While we use variance to calculate confidence as in previous literature [4], FedLabel is not restricted to this metric. We include an ablation study on what to use as the confidence score of the logits in Appendix A. Formally, we have that

$$\begin{gathered}\text{Binary Selection of Logit: } \mathbf{s}^*(\xi) = \underset{\mathbf{s} \in \mathcal{S}^{(t)}(\xi)}{\arg\max}\, h(\mathbf{s}), \\ \mathcal{S}^{(t)}(\xi) := \{\mathbf{s}(\mathbf{w}^{(t,0)},\ \xi), \mathbf{s}(\mathbf{w}_{\mathcal{L},k}^{(t,\tau)},\ \xi)\}\end{gathered} \tag{3}$$

Pseudo Label from Thresholding:

$$\hat{y}_\xi = \underset{i \in [N]}{\arg\max}\, \mathbf{s}^*(\xi) \mathbb{1}(\max \mathbf{s}^*(\xi) > \beta) \tag{4}$$

We discard the instances of $\hat{y}_\xi = 0$ which indicates that the selected logit did not pass the thresholding function in (4).

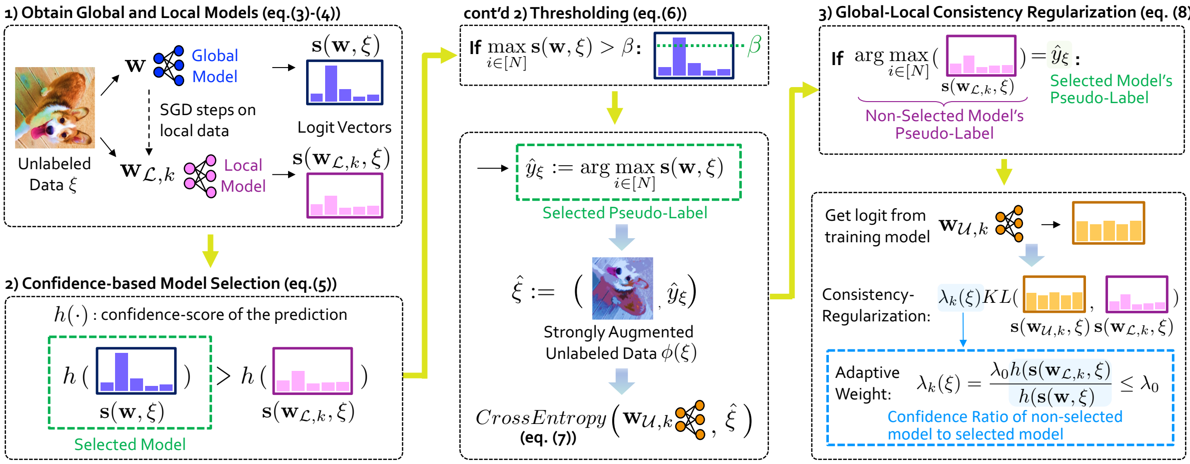



Figure 3: Overview of FEDLABEL leveraging unlabeled data for each client $k \in [M]$. FEDLABEL consists of 3 main steps to leverage unlabeled data: 1) Obtaining global and local models (eq. (1)-(2)), 2) Confidence-based selection with thresholding for obtaining the pseudo-label for the cross-entropy loss (eq. (3)-(4)), and 3) the global-local consistency regularizing term to assimilate knowledge from both global and local models when applicable (eq. (6)-(7)). Details of the end-to-end algorithm of FEDLABEL is in Algorithm 1.

Given that $\hat{y}_\xi \neq 0$, we have the pseudo-label $\hat{y}_\xi$ obtained via (4) for the unlabeld data $\xi \in \mathcal{D}_{U,k}$. Then, with the separately set local model to be trained over the unlabeled data defined as $\mathbf{w}_{\mathcal{U},k}$, we have the cross-entropy loss for FEDLABEL as

$$CrossEntropy(\mathbf{w}_{\mathcal{U},k},\ \hat{\xi}),\ \hat{\xi} := (\psi(\xi), \hat{y}_\xi) \tag{5}$$

where $\psi(\cdot)$ is a strong-augumentation (RandAugment [6]) of the data sample $\xi$ with its hyperparameters set to $(1, 10)$.

**3) Global-Local Consistency Regularization.** The cross-entropy loss of FEDLABEL in (5) uses the pseudo-label selected via a binary selection between the local and global model which essentially discards the logits of the model that has not been selected. However, there can be cases where the discarded model's soft logits which we denote as $\mathbf{s}^{-*}(\xi)$ also point to the same label $\hat{y}_\xi$, i.e., $\arg\max_{i\in[N]} \mathbf{s}^{-*}(\xi) = \hat{y}_\xi$ as the selected model. In such cases with only the cross-entropy term in (5), FEDLABEL can lose the useful information contained in the discarded model. To cover such scenarios, FEDLABEL utilizes the knowledge of the discarded model when it predicts the same label as the selected model by our proposed global-local consistency regularizing term as:

$$\lambda_k(\xi) KL(\mathbf{s}(\mathbf{w}_{\mathcal{U},k}, \xi), \mathbf{s}^{-*}(\xi)) \mathbb{1}(\arg\max_{i\in[N]} \mathbf{s}^{-*}(\xi) = \hat{y}_\xi), \tag{6}$$

$$\text{where } \lambda_k(\xi) := \lambda_0 h(\mathbf{s}^{-*}(\xi))/h(\mathbf{s}^*(\xi)) \tag{7}$$

Note that the global-local consistency regularizing term (KL-divergence) is weighed by $\lambda_k(\xi)$ which is the confidence-score ratio by the discarded model to the selected model, i.e., $\lambda_0 h(\mathbf{s}^{-*}(\xi))/h(\mathbf{s}^*(\xi)) \leq \lambda_0$, so that the less confident the discarded model is compared to the selected model, the lower the regularization weight. The maximum value of $\lambda_k(\xi),\ \xi \in \mathcal{D}_{\mathcal{U},k},\ \forall k \in [M]$ is $\lambda_0$ which is achieved when the confidence score of the discarded model and the selected model is identical. We show in Section 5.2 that this consistency regularization term indeed helps FEDLABEL improve its performance from at least 5% to at most 9%.

**FEDLABEL's Final Semi-supervised Loss.** Combining the cross-entropy loss from the confidence-based selection in (5) and the global-local consistency regularization loss in (6), we have the final semi-supervised loss of FEDLABEL denoted as $F_{\mathcal{U},k}(\mathbf{w})$ for each client $k \in [M]$ as below:

$$\begin{aligned} F_{\mathcal{U},k}(\mathbf{w}_{\mathcal{U},k}) &= \frac{1}{|\mathcal{D}_{\mathcal{U},k}|} \sum_{\xi \in \mathcal{D}_{\mathcal{U},k}} (CrossEntropy(\mathbf{w}_{\mathcal{U},k},\ \hat{\xi}) \\ &+ \lambda_k(\xi) KL(\mathbf{s}(\mathbf{w}_{\mathcal{U},k}, \xi), \mathbf{s}^{-*}(\xi)) \mathbb{1}(\arg\max_{i\in[N]} \mathbf{s}^{-*}(\xi) = \hat{y}_\xi)) \end{aligned} \tag{8}$$

With the final semi-supervised loss of FEDLABEL proposed in (8), now we are ready to elaborate on how FEDLABEL is implemented in the subsequent subsection.

## 4.2. FEDLABEL Implementation

Recall that at each communication round $t$, the selected set of clients $\mathcal{C}^{(t,0)}$ initialize their local model to the global model, i.e., $\mathbf{w}_{\mathcal{L},k}^{(t,0)} = \mathbf{w}^{(t,0)}$ and perform $\tau$ local iterations to obtain the supervised local update $\Delta\mathbf{w}_{\mathcal{L},k}^{(t,0)}$ (see (2)). Using the same initialization $\mathbf{w}_{\mathcal{U},k}^{(t,0)} := \mathbf{w}^{(t,0)}$, the clients now obtain the semi-supervised local update by minimizing the semi-supervised loss $F_{\mathcal{U},k}(\mathbf{w}_{\mathcal{U},k})$ (see (8)) via mini-batch SGD for $\tau'$ iterations on the unlabeled data $\mathcal{D}_{\mathcal{U},k}, k \in \mathcal{C}^{(t,0)}$. Concretely, we have the semi-supervised local update as:

**Algorithm 1** FEDLABEL Framework

**Initialize** $\mathbf{w}^{(0,0)}$, **Output:** Global model $\mathbf{w}^{(T,0)}$,
**For** $t = 0, ..., T-1$ **communication rounds do**:
  **Global server:** Select $m$ clients for $\mathcal{C}^{(t,0)}$ uniformly at random and send $\mathbf{w}^{(t,0)}$ to clients in $\mathcal{C}^{(t,0)}$
  **Clients** $k \in \mathcal{C}^{(t,0)}$ **in parallel do:**
    Set $\mathbf{w}_{\mathcal{L},k}^{(t,0)} = \mathbf{w}^{(t,0)}$, $\mathbf{w}_{\mathcal{U},k}^{(t,0)} = \mathbf{w}^{(t,0)}$
    Get $\mathbf{w}_{\mathcal{L},k}^{(t,\tau)} = \Delta\mathbf{w}_{\mathcal{L},k}^{(t,0)} + \mathbf{w}_{\mathcal{L},k}^{(t,0)}$ (See (2))
    Get $\Delta\mathbf{w}_{\mathcal{U},k}^{(t,0)}$ (See (10))
    Send $\Delta\mathbf{w}_{k}^{(t,0)} = \Delta\mathbf{w}_{\mathcal{L},k}^{(t,0)} + \Delta\mathbf{w}_{\mathcal{U},k}^{(t,0)}$ and aggregation weight $r_k^{(t,0)}$ to the server
  **Global server:** Update $\mathbf{w}^{(t+1,0)} = \mathbf{w}^{(t,0)} + \sum_{k\in\mathcal{C}^{(t,0)}} \frac{r_k^{(t,0)}}{\sum_{k'\in\mathcal{C}^{(t,0)}} r_{k'}^{(t,0)}} \Delta\mathbf{w}_k^{(t,0)}$

(Perform Local SGD on *Unlabeled* Data)

$$\mathbf{w}_{\mathcal{U},k}^{(t,\tau)} = \mathbf{w}_{\mathcal{U},k}^{(t,0)} - \eta \sum_{l=0}^{\tau'-1} \nabla F_{\mathcal{U},k}(\mathbf{w}_{\mathcal{U},k}^{(t,l)}, \xi_{\mathcal{U},k}^{(t,l)}) \quad (9)$$

(Compute the *Semi-supervised* Local Update)

$$\Delta\mathbf{w}_{\mathcal{U},k}^{(t,0)} = \mathbf{w}_{\mathcal{U},k}^{(t,\tau)} - \mathbf{w}_{\mathcal{U},k}^{(t,0)} \quad (10)$$

where $\nabla F_{\mathcal{U},k}(\mathbf{w}_{\mathcal{U},k}^{(t,l)}, \xi_{\mathcal{U},k}^{(t,l)}) = \frac{1}{b}\sum_{\xi\in\xi_{\mathcal{U},k}^{(t,l)}} \nabla f(\mathbf{w}_{\mathcal{U},k}^{(t,l)}, \xi)$ is the stochastic gradient computed using a mini-batch $\xi_{\mathcal{U},k}^{(t,l)}$ of size $b$ that is randomly sampled from client $k$'s local unlabeled dataset $\mathcal{D}_{U,k}$. Now each client $k \in \mathcal{C}^{(t,0)}$ sends back its update $\Delta\mathbf{w}_k^{(t,0)} = \Delta\mathbf{w}_{\mathcal{L},k}^{(t,0)} + \Delta\mathbf{w}_{\mathcal{U},k}^{(t,0)}$ back to the server along with the total number of data samples $r_k^{(t,0)}$ used for obtaining the local updates $\Delta\mathbf{w}_{\mathcal{L},k}^{(t,0)}$ and $\Delta\mathbf{w}_{\mathcal{U},k}^{(t,0)}$. Note that $r_k^{(t,0)}$ is dependent on $t$ because the number of unlabeled data samples that are used for training varies for each communication round depending on how many samples pass the confidence-based threshold in (4). Then finally the server updates its global model as $\mathbf{w}^{(t+1,0)} = \mathbf{w}^{(t,0)} + \sum_{k\in\mathcal{C}^{(t,0)}} \frac{r_k^{(t,0)}}{\sum_{k'\in\mathcal{C}^{(t,0)}} r_{k'}^{(t,0)}} \Delta\mathbf{w}_k^{(t,0)}$. The details of FEDLABEL's implementation is in Algorithm 1.

## 5. Experiments

### 5.1. Experimental Setup

We perform experiments on a wide variety of experiments on **both cross-device and cross-silo settings, partial and full client participation, low and high data heterogeneity, and low and large label scarcity**. We perform 3 tasks for cross-device: Resnet18 with EMNIST (62 labels) [5], Resnet34 with CIFAR10 (10 labels), and Resnet50 with CIFAR100 (100 labels) [22] and 2 tasks for cross-silo: OrganAMNIST (11 labels) and BloodMNIST (8 labels) [43] both with Resnet18. For cross-device, we select 10% of clients for training for each comm. round. For the OrganAMNIST and BloodMNIST we select 30% and 100% of clients.

**Data Partitioning Across Clients.** For cross-device, the data is partitioned across 100 clients using the Dirichlet distribution Dir$(\alpha = 0.1)$ [14], unless mentioned otherwise. The parameter $\alpha$ determines the degree of data heterogeneity (smaller $\alpha$ indicates larger data heterogeneity). For cross-silo, data is partitioned across 10 and 5 clients in total respectively with $\alpha = 0.1$, unless mentioned otherwise. Note that we emulate a realistic setting where the total number of clients is much smaller for cross-silo than for cross-device.

**Data Partitioning Within Clients.** The local training dataset of each client is partitioned into labeled and unlabeled data uniformly at random. Note that even without artificially biasing the dataset partitioning, we were still able to observe class-distribution mismatch for high label scarcity (see Fig. 2). For cross-device, we have $20\% : 80\%$ and $50\% : 50\%$ partitioning of the labeled : unlabeled data for each client's local training data. For cross-silo, we have $5\% : 95\%$ and $20\% : 80\%$ partitioning. Note that we have run more fine-grained data ratio experiments (10,20,50,65 for CIFAR10, and 5,20,35,50 for OrganAMNIST), also shown in Fig. 1, but have selected a few intervals that have shown the most significant difference for presentation.

**Baselines.** We compare FEDLABEL with 3 classes of baselines: i) Supervised FL baselines with Fully Labeled Data denoted as 100% (**FedAvg (100%)**, **FedProx (100%)**), ii) Supervised FL baselines with Partially Labeled Data (**FedAvg, FedProx**), and iii) SSFL baselines with Partially Labeled Data ((**FedAvg+UDA, FedAvg+FixMatch, FedProx+UDA, FedProx+FixMatch, FedTriNet, FedMatch**). The i) supervised FL baselines with 100% of labeled data is the hypothetical *upper bound* that we can achieve by using FL when clients have all their data labeled, and ii) the supervised FL baselines with partially labeled data is the *lower bound* that FL actually achieves in a realistic setting with only a few labeled data. The iii) SSFL baselines are the state-of-the-art methods that tackle label scarcity in FL. We do not compare with SemiFL [8] and Rscfed [2] which impose assumptions such as the server having labeled data or several clients having fully labeled data (see Table 2).

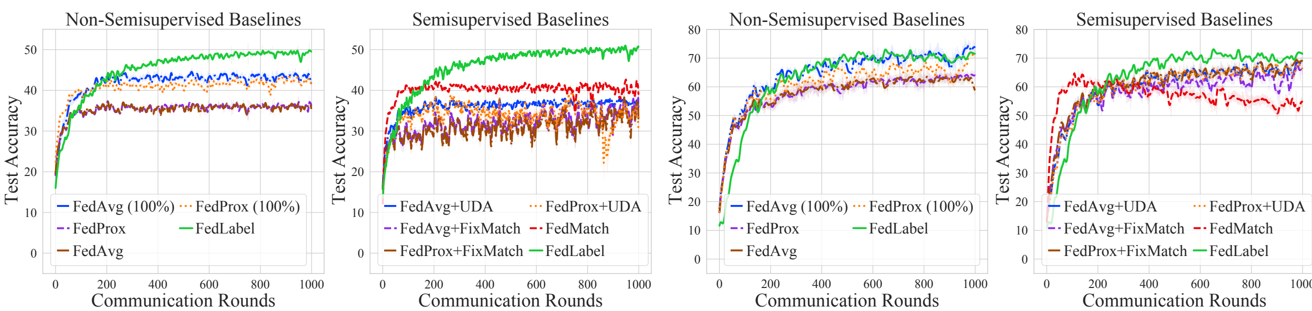
### 5.2. Experimental Results

We thoroughly evaluate FEDLABEL in the following aspects: achieves high accuracy for i) both low and high label scarcity (5-20% and 20-50%), and ii) both low and high data heterogeneity. We also perform ablation studies on FEDLABEL as follows: iii) effect of the global-local consistency regularizing term modulated by $\lambda_0$ in (7), iv) effect of the number of local steps $\tau$ to obtain the local model in (1), v) effect of the thresholding parameter $\beta$ in (4), and vi) effect of different confidence measures $h(\cdot)$ in (3). We defer the results for v) and vi) to Appendix A due to space constraints.

**Effectiveness of FEDLABEL for High Label Scarcity.** We

| High Client Data Heterogeneity | | Cross-Device Setting | | | Cross-Silo Setting | |
|---|---|---|---|---|---|---|
| | | EMNIST | CIFAR10 | CIFAR100 | OrganAMNIST | BloodMNIST |
| Supervised, Fully Labeled | FedAvg (100%) | 78.36 (±0.73) | 44.67 (±1.09) | 24.67 (±0.55) | 71.09 (±1.00) | 70.53 (±1.13) |
| | FedProx (100%) | **78.60** (±0.89) | **44.72** (±1.63) | **24.87** (±0.19) | **73.34** (±0.86) | **74.53** (±0.39) |
| Supervised, Partially Labeled | FedAvg | 63.05 (±1.52) | 28.83 (±2.94) | 11.33 (±0.22) | 41.90 (±2.18) | 53.18 (±0.11) |
| | FedProx | 63.31 (±0.16) | 31.09 (±1.15) | 11.37 (±0.15) | 48.28 (±0.78) | 56.53 (±0.68) |
| Semi-Supervised, Partially Labeled | FedAvg+UDA | 68.98 (±1.06) | 31.32 (±2.84) | 10.80 (±0.57) | 55.14 (±0.92) | 53.18 (±0.68) |
| | FedAvg+FixMatch | 67.98 (±1.59) | 23.99 (±2.73) | 10.16 (±0.20) | 52.76 (±2.95) | 47.92 (±1.55) |
| | FedProx+UDA | 71.93 (±1.37) | 31.99 (±2.26) | 10.57 (±0.13) | 54.15 (±2.53) | 59.48 (±1.17) |
| | FedProx+FixMatch | 70.56 (±0.96) | 22.38 (±2.28) | 10.02 (±0.54) | 55.49 (±2.65) | 50.88 (±1.12) |
| | FedTriNet | 60.31 (±0.25) | 29.56 (±1.53) | 10.53 (±1.83) | 50.82 (±1.21) | 59.38 (±1.05) |
| | FedMatch | 63.48 (±0.57) | 31.94 (±1.78) | 10.86 (±1.33) | 48.99 (±1.38) | 58.89 (±2.24) |
| | FedLabel (**ours**) | **79.33** (±1.97) | **46.05** (±1.09) | **18.42** (±1.46) | **69.43** (±1.58) | **71.46** (±1.89) |

Table 3: Test accuracy for high label scarcity on each client's local data (20% of labeled data for cross-device and 5% for cross-silo). FedLabel achieves a significantly higher test accuracy by approximately 8-16%, 15-24%, 8%, 14-20%, and 12-23% for EMNIST, CIFAR10, CIFAR100, OrganAMNIST, and BloodMNIST respectively. For EMNIST, CIFAR10, and BloodMNIST, FedLabel achieves *an even higher test accuracy* than the supervised fully labeled case (100% of the data labeled) by approximately 1%, 2%, and 1%.




(a) 50% of Labeled Data for CIFAR10

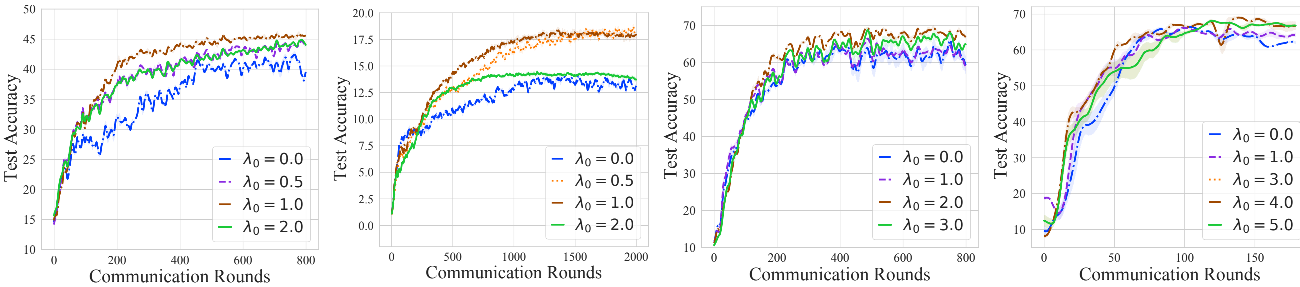
(b) 20% of Labeled Data for OrganAMNIST

Figure 4: Test accuracy for lower label scarcity, 50% and 20% of labeled data, for each client's local training data for CIFAR10 and OrganAMNIST respectively. FedLabel outperforms the baselines by 8-15% and 2-10% for CIFAR10 and OrganAMNIST respectively. Comparing to results in Table 3, FedLabel performance gap with other baselines is higher when there is higher label scarcity.

evaluate FedLabel with the top-1 test accuracies for 20% and 5% of labeled data on each client for cross-device and cross-silo respectively in Table 3. Compared to the other SSFL baselines, FedLabel achieves a higher test accuracy by 8-16%, 15-24%, 8%, 14-20%, and 12-23% for EMNIST, CIFAR10, CIFAR100, OrganAMNIST, and BloodMNIST respectively. Surprisingly, for EMNIST, CIFAR10, and BloodMNIST, **FedLabel achieves an even higher test accuracy than the fully supervised case** (100% of labeled data) by approximately 1%, 2%, and 1%. In other words, FedLabel performs even better than the ideal scenario when 100% of the data is labeled, with only 5-20% of labeled data. We reason that this is due to the small amount of noise introduced by FedLabel when leveraging the unlabeled data which can improve the generalization performance as shown in previous work [45, 10]. FedLabel's also outperforms by 7-27% than the supervised baselines with partially labeled data, while some of the SSFL baselines with partially labeled data perform even worse. Next, we evaluate FedLabel when there are more labels in the subsequent paragraph.

**Comparison with Cases for Lower Label Scarcity.** In Fig. 4, we show results for larger portions of labeled data, 50% for CIFAR10 and 20% for OrganAMNIST, than the results shown in Table 3 which was for 20% and 5% of labeled data respectively. For the larger number of labeled data, FedLabel still outperforms the baselines by 8-15% and 2-10% for CIFAR10 and OrganAMNIST respectively. However, the performance gap is lower than the 15-24% and 14-20% improvement shown in Table 3 for higher label scarcity. Hence, this shows that while FedLabel still outperforms other baselines for lower label scarcity, it outperforms the other baselines with a higher gap when clients have a smaller number of labels, i.e., high label scarcity.

**Robustness of FedLabel to Data Heterogeneity.** In Table 4, we show the test accuracy for smaller data heterogeneity ($\alpha = 1$) where FedLabel still outperforms the other baselines by 4-10% and 7-20% for CIFAR10 and OrganAMNIST respectively. However, compared to the results in Table 3 which is for higher data heterogeneity ($\alpha = 0.1$), the performance gap between FedLabel and the other baselines is smaller by around 11-14%. This implies that FedLabel works better when there is high data heterogeneity, while the other baselines perform worse when there is higher data heterogeneity across clients.




(a) CIFAR10 (b) CIFAR100 (c) OrganAMNIST (d) BloodMNIST

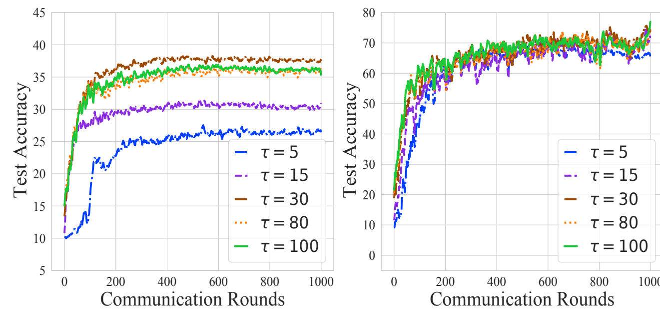
Figure 5: Ablation study on the effect of the global-local consistency regularizing term by modulating $\lambda_0$ in (7). For all datasets, $\lambda_0 = 0$ gives the lowest test accuracy, showing that without the regularizing term, we lose useful information from the discarded model from the binary selection between the local and global model. For larger $\lambda_0 > 0$, the test accuracy improves by approximately 5-9%.

| Low Inter-Client Data Heterogeneity | CIFAR10 | OrganAMNIST |
|---|---|---|
| FedAvg (100%) | 57.17 (±0.34) | 81.49 (±0.96) |
| FedProx (100%) | **57.45** (±0.51) | **82.07** (±1.10) |
| FedAvg | 40.11 (±0.67) | 59.74 (±1.52) |
| FedProx | 41.45 (±0.58) | 61.86 (±1.23) |
| FedAvg+UDA | 42.07 (±0.94) | 63.05 (±1.91) |
| FedAvg+FixMatch | 38.51 (±0.96) | 70.14 (±1.42) |
| FedProx+UDA | 42.08 (±0.83) | 65.12 (±2.03) |
| FedProx+FixMatch | 42.94 (±0.86) | 70.58 (±1.13) |
| FedTriNet | 42.67 (±0.79) | 74.16 (±1.42) |
| FedMatch | 44.23 (±0.88) | 72.57 (±1.58) |
| FEDLABEL (**ours**) | **48.89** (±0.91) | **79.76** (±0.83) |

Table 4: Test accuracy for labeled data 20% and 5% respectively for CIFAR10 and OrganAMNIST with lower data heterogeneity ($\alpha = 1$). FEDLABEL outperforms the baselines by 4-10% and 7-20% for CIFAR10 and OrganAMNIST respectively.

**Effect of Global-Local Consistency Regularization.** In FEDLABEL, we use global-local consistency regularization which is weighted by the parameter $\lambda_0$ (see (7)). We evaluate the effectiveness of this term by varying $\lambda_0$ in Fig. 5. For all datasets, $\lambda_0 = 0$ gives the lowest test accuracy, showing that without the regularizing term, we are losing useful information from the discarded model due to the binary selection between the local and global model. As we increase $\lambda_0$, we see significant improvement in the test accuracy of 5-9%. However, we also observe that when $\lambda_0$ exceeds a certain threshold the improvement decreases. The intensity of the global-local consistency regularizing term can be modulated by tuning $\lambda_0$ when appropriate.

**Number of Training Steps to Obtain the Local Model.** To obtain the local model (see Eq. (1)), clients perform $\tau$ local SGD steps on the received global model with their labeled data. Hence, with larger $\tau$, the more the local model well reflects the client's local data. In Fig. 6, we perform an ablation study on the effect of $\tau$. The smallest $\tau = 5$ yields the worst performance across the range of $\tau$, showing that for the local model to well reflect the client's local data and bring distinct information from the global model, $\tau$ needs to be set to a moderately large value. For larger $\tau$, the performance improves and gradually saturates indicating that the local




(a) CIFAR10 (b) OrganAMNIST

Figure 6: Ablation study on the number of local steps $\tau$ to obtain the local model $\mathbf{w}_{\mathcal{L},k}$ for client $k \in [M]$. The smallest $\tau = 5$ yields the worst performance showing that for the local model to well reflect the client's local data, $\tau$ needs to be moderately large.

model converges. We provide additional results on the effect of $\tau$ for a larger number of labeled data in Appendix A, giving further insight into how the quality of the local model affects FEDLABEL's performance.

## 6. Concluding Remarks

In conclusion, we propose FEDLABEL, a SSFL framework that works well for both low and high data heterogeneity cases, as well as for both limited and larger portions of labeled data. We use a confidence-based binary selection of the local or global model for pseudo-labeling with global-local consistency regularization. Unlike previous work, FEDLABEL does not require additional computation to find new experts, additional communication of parameters, server labeled data, or any fully labeled clients. In both cross-device and cross-silo settings, we show that FEDLABEL largely outperforms other SSFL baselines, especially when there is high data heterogeneity and label scarcity, by at most 24%. FEDLABEL even outperforms fully-supervised FL baselines which use fully-labeled data with only using 5-20% of labeled data. Currently, FEDLABEL does not consider the possible noise that can be present in the labeled data of the clients caused, for instance, clients mislabeling their data due to lack of expertise. Thus, for future work, we aim to extend the FEDLABEL to be robust to label noise.

## References


[1] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *CHI '17*, page 2334–2346, New York, NY, USA, 2017. Association for Computing Machinery. 3

[2] Liwei Che, Zewei Long, Jiaqi Wang, Yaqing Wang, Houping Xiao, and Fenglong Ma. Fedtrinet: A pseudo labeling method with three players for federated semi-supervised learning. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 715–724, 2021. 1, 2, 3, 6

[3] Y. Chen, X. Zhu, W. Li, and S. Gong. Semi-supervised learning under class distribution mismatch. In *In The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 1

[4] Yae Jee Cho, Andre Manoel, Gauri Joshi, Robert Sim, and Dimitrios Dimitriadis. Heterogeneous ensemble knowledge transfer for training large models in federated learning. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-22)*, 2022. 4

[5] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. EMNIST: an extension of MNIST to handwritten letters. *arXiv preprint arXiv:1702.05373*, 2017. 6

[6] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. *arXiv preprint arXiv:1909.13719*, 2019. 5

[7] Aron Culotta and Andrew McCallum. Reducing labeling effort for structured prediction tasks. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2005. 3

[8] Enmao Diao, Jie Ding, and Vahid Tarokh. SemiFL: Semi-Supervised Federated Learning for Unlabeled Clients with Alternate Training. In *Advances in Neural Information Processing Systems*, 2022. 2, 3, 6

[9] Chenyou Fan, Junjie Hu, and Jianwei Huang. Private Semi-Supervised Federated Learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*, 2022. 1

[10] Xian-Jin Gui, Wei Wang, and Zhang-Hao Tian. Towards understanding deep learning from noisy labels with small-loss criterion. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*, 2021. 7

[11] Farzin Haddadpour and Mehrdad Mahdavi. On the convergence of local descent methods in federated learning. *arXiv preprint arXiv:1910.14425*, 2019. 2

[12] Zhongyi Han, Benzheng Wei, Yanfei Hong, Tianyang Li, Jinyu Cong, Xue Zhu, Haifeng Wei, and Wei Zhang. Accurate screening of covid-19 using attention-based deep 3d multiple instance learning. *IEEE Transactions on Medical Imaging*, 39(8):2584–2594, 2020. 1

[13] Chaoyang He, Zhengyu Yang, Erum Mushtaq, Sunwoo Lee, Mahdi Soltanolkotabi, and Salman Avestimehr. SSFL: Tackling Label Deficiency in Federated Learning via Personalized Self-Supervision. *preprint*, 2021. 1, 3

[14] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. In *International Workshop on Federated Learning for User Privacy and Data Confidentiality in Conjunction with NeurIPS 2019 (FL-NeurIPS'19)*, Dec. 2019. 6

[15] M. Humbert-Droz, P. Mukherjee, and O. Gevaert. Strategies to address the lack of labeled data for supervised machine learning training with electronic health records: Case study for the extraction of symptoms from clinical notes. *JMIR Med Inform*, 10(3), 2022. 1

[16] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based Semi-supervised Learning for Object Detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3

[17] Wonyong Jeong, Jaehong Yoon, Eunho Yang, and Sung Ju Hwang. Federated semi-supervised learning with inter-client consistency & disjoint learning. In *International Conference on Learning Representations (ICLR)*, 2021. 2, 3

[18] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurelien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adria Gascon, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecny, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrede Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Ozgur, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramer, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019. 1

[19] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for on-device federated learning. *arXiv preprint arXiv:1910.06378*, 2019. 3

[20] A Khaled, K Mishchenko, and P Richtárik. Tighter theory for local SGD on identical and heterogeneous data. In *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, 2020. 2

[21] Woojung Kim, Keondo Park, Kihyuk Sohn, Raphael Shu, and Hyung-Sin Kim. Federated semi-supervised learning with prototypical networks. *arXiv preprint arXiv:2205.13921*, 2022. 1

[22] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Learning multiple layers of features from tiny images. *CIFAR-10 (Canadian Institute for Advanced Research)*, 2009. 6

[23] Xiaoxiao Liang, Yiqun Lin, Huazhu Fu, Lei Zhu, and Xiaomeng Li1. Rscfed: Random sampling consensus federated semi-supervised learning. In *In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[24] Xinyang Lin, , Hanting Chen, Yixing Xu, Chao Xu, Xiaolin Gui, Yiping Deng, and Yunhe Wang. Federated Learning with Positive and Unlabeled Data. *preprint*, 2021. 3

[25] Nan Lu, Zhao Wang, Xiaoxiao Li, Gang Niu, Qi Dou, and Masashi Sugiyama. Federated learning from only unlabeled data with class-conditional-sharing clients. *International Conference on Learning Representations (ICLR)*, 2022. 3

[26] Grigory Malinovsky, Dmitry Kovalev, Elnur Gasanov, Laurent Condat, and Peter Richtárik. From local SGD to local fixed point methods for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. 2

[27] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agøura y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. *International Conference on Artificial Intelligenece and Statistics (AISTATS)*, Apr. 2017. 1, 2, 4

[28] Takeru Miyato, Shin ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 3

[29] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations (ICLR)*, 2021. 2, 3

[30] Zhongzheng Ren, Raymond A. Yeh, and Alexander G. Schwing. Not All Unlabeled Data are Equal: Learning to Weight Data in Semi-supervised Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

[31] Yuji Roh, Geon Heo, and Steven Euijong Whang. A survey on data collection for machine learning: A big data - ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1328–1347, 2021. 3

[32] Adam Sadilek, Luyang Liu, Dung Nguyen, Methun Kamruzzaman, Benjamin Rader, Alex Ingerman, Stefan Mellem, Peter Kairouz, Elaine O. Nsoesie, Jamie MacFarlane, Anil Vullikanti, Madhav Marathe, Paul Eastham, John S. Brownstein, Michael Howell, and John Hernandez1. Privacy-first health research with federated learning. In *npj Digital Medicine*, 2021. 1

[33] Anit Kumar Sahu, Tian Li, Maziar Sanjabi, Manzil Zaheer, Ameet Talwalkar, and Virginia Smith. Federated optimization for heterogeneous networks. In *Proceedings of the 3rd MLSys Conference*, Jan. 2020. 2, 4

[34] Jonathan Scott, Michelle Yeo, and Christoph H. Lampert. Fedprop: Cross-client label propagation for federated semi-supervised learning. *arXiv preprint arXiv:2210.06434*, 2022. 1

[35] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1, 3, 11

[36] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation Consistency Training for Semi-supervised Learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019. 3

[37] Binghui Wang, Ang Li, Hai Li, and Yiran Chen. GraphFL: A Federated Learning Framework for Semi-Supervised Node Classification on Graphs. *preprint*, 2020. 3

[38] Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H Brendan McMahan, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021. 1

[39] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization. *preprint*, May 2020. 2

[40] Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Yongfeng Huang, and Xing Xie. Communication-efficient federated learning via knowledge distillation. *Nature Communications*, 13(1):2032, 2022. 2

[41] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems*, 2020. 1, 3

[42] Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. Dash: Semi-Supervised Learning with Dynamic Thresholding. In *International Conference on Machine Learning (ICML)*, 2021. 3

[43] 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. *arXiv preprint arXiv:2110.14795*, 2021. 6

[44] Wenti Yang, Naiyu Wang, Zhitao Guan, Longfei Wu, Xiaojiang Du, and Mohsen Guizani. A practical cross-device federated learning framework over 5g networks. In *Proceedings of the IEEE Wireless Communications*, 2022. 1

[45] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, , Ivor W. Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019. 7

[46] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flex-Match: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[47] Z. Zhao, L. Zhou, Y. Duan, L. Wang, L. Qi, and Y. Shi. Dc-ssl: Addressing mismatched class distribution in semi-supervised learning. In *In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1