University of Glasgow

School *of* Computing Science

Knowledge & Data Engineering Systems

The Price of Labelling: A Two-Phase Federated Self-Learning Approach

Tahani Aladwani, Christos Anagnostopoulos, Shameem Parambath & Fani Deligianni

ECML PKDD Conference 2024, Mon, 9 Sept 2024 – Fri, 13 Sept 2024, Vilnius

School *of* Computing Science
Knowledge & Data Engineering Systems

Introduction

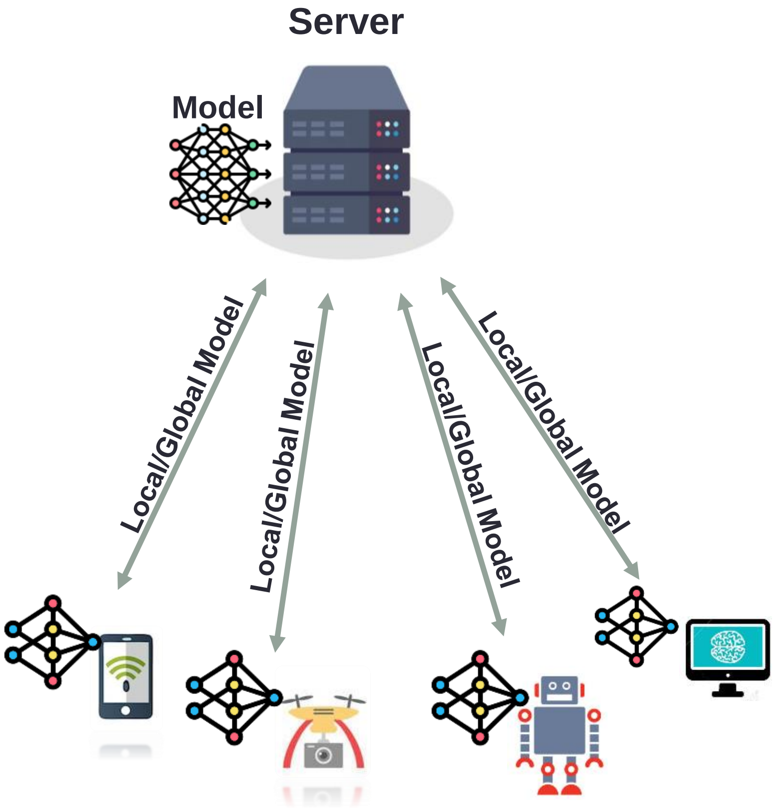
Key (ideal) assumptions in Federated Learning (FL) :

1. **Supervised Learning:** All clients possess **<u>sufficient</u>** training data with ground-truth labels.

2. **Sumi Supervised Learning:** Subset of clients or server have **<u>adequate labelled samples to train supervised models</u>**, ensuring generalization across ‘unlabelled’ clients.

3. **Self-Learning:** Operates under the assumption that data are independent and identically distributed (IID).

4. The model can generate **<u>high-quality pseudo-labels</u>** by considering **<u>only labelled data</u>** during the training.



Introduction

Distributed data in real-world scenarios:

- Data can be non-IID, leading to common issues such as **class imbalance & distribution shift across clients**.
- Existence of un-labeled data across clients, due to various factors like **limited resources, labeling costs, and human errors**

Challenge: create high-quality pseudo-labels without addressing these issues.

- Model performance heavily relies on the quality and distribution of the training data.
- High degree of heterogeneity among client data significantly decreases model performance.

University of Glasgow

School *of* Computing Science
Knowledge & Data Engineering Systems

Overview of the problem

Disparity between ideal key assumptions & realistic scenarios prompt us to contemplate the following question:

What is the *price* of learning a global model using *scarce* and *skewed* distributed *labelled* data, while capitalizing on partially labelled and fully unlabelled data across clients?

Overview & Fundamentals

Consider a set $\mathcal{N} = \{n_1, \dots, n_{\mathcal{N}}\}$ of distributed clients. Each client $n_i \in \mathcal{N}$ possesses a dataset $\mathcal{D}_i$ containing $\mathcal{C} = \{0, \dots, \mathcal{C} - 1\}$ classes (labels) of data, which can be **labelled and/or unlabelled**.

Clients are categorized into three types based on their data:

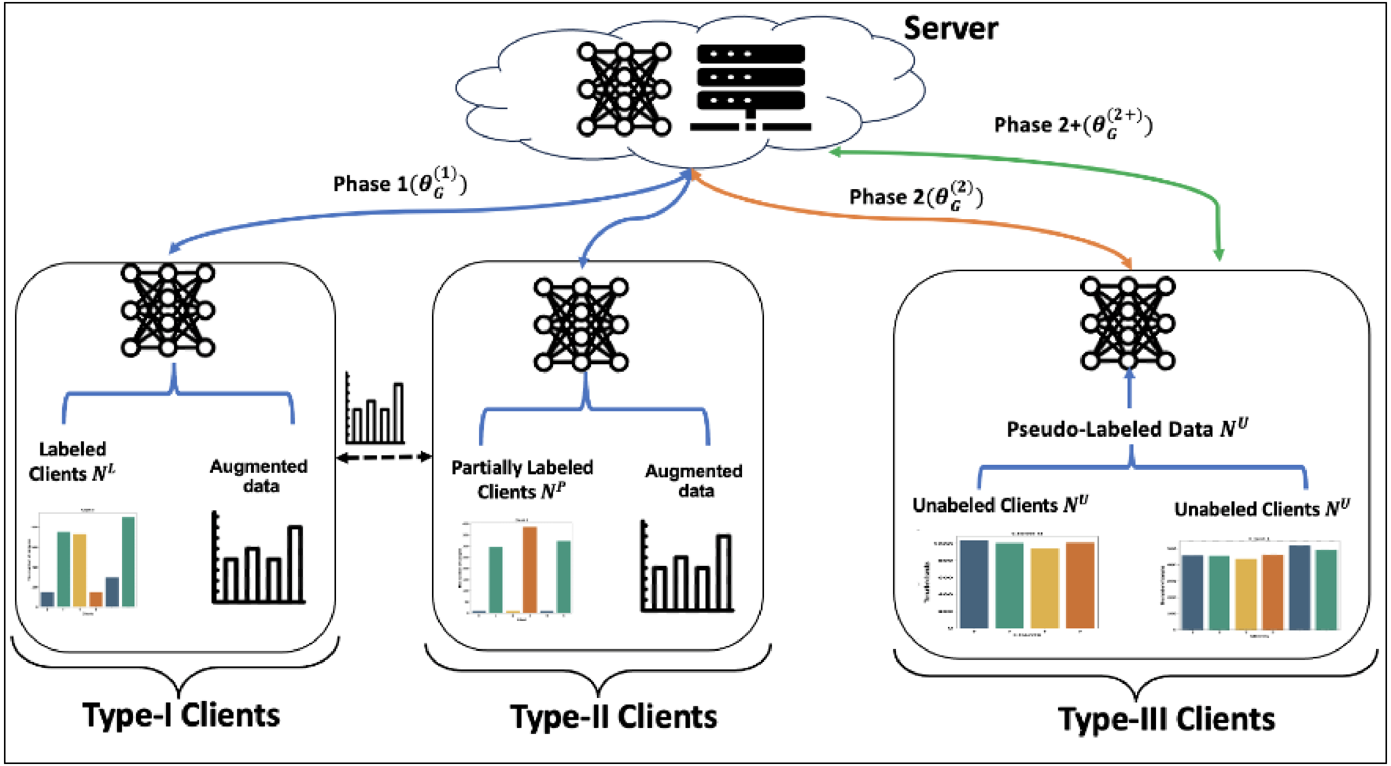
– **Type I** clients (**labelled clients**) $n_i \in \mathcal{N}^L \subset \mathcal{N}$, denoted as $\mathcal{D}_i^L = \{(x_k, y_k)\}_{k=1}^{\mathcal{D}_i^L}$, y_k is the label.

– **Type II** clients (**partially labelled clients**) $n_i \in \mathcal{N}^P \subset \mathcal{N}$ have **labelled and unlabelled** samples, i.e., $\mathcal{D}_i^P = \{(x_k, y_k \vee \perp)\}_{k=1}^{\mathcal{D}_i^P}$, $\perp$.

– **Type III** clients (**unlabelled clients**) $n_i \in \mathcal{N}^L \subset \mathcal{N}$ have **all samples unlabelled**, , i.e., $\mathcal{D}_i^L = \{(x_k, \perp)\}_{k=1}^{\mathcal{D}_i^U}$.

Focus: labelled samples are much fewer than unlabelled ones, i.e., $|\mathcal{D}^L| \ll |\mathcal{D}^U|$

2-Phase Federated Self-Learning Framework (**2PFL**)



2-Phase Federated Self-Learning Framework (2PFL)

1. Local Data Augmentation

2PFL adopts **MixUp** to augment data over client .

✓ **In labelled/partially labelled client** $n_i \in \mathcal{N}^L \cup \mathcal{N}^P$, for any two inputs x_k and x_ℓ with labels y_k and y_ℓ, MixUp synthesizes the sample (x', y'):

$$x' = \lambda x_k + (1-\lambda) x_\ell \;\; \boldsymbol{and} \;\; y' = \lambda y_k + (1-\lambda) y_\ell$$

with $\lambda \in (0, 1)$, a blending parameter controlling interpolation between samples.

✓ **In unlabelled client** $n_i \in \mathcal{N}^U$, two randomly selected pseudo-labelled inputs x_k and x_ℓ with high-confidence pseudo-labels $\hat{y}_k$ and $\hat{y}_\ell$, respectively, generate the sample (x', y'):

$$x' = \lambda x_k + (1-\lambda) x_\ell \;\; \boldsymbol{and} \;\; y' = \lambda \hat{y}_k + (1-\lambda) \hat{\boldsymbol{y}}_\ell$$

2-Phase Federated Self-Learning Framework

2. 2PFL Training Phases

2PFL exploits labelled, partially labelled and unlabelled data across all types of clients $(\mathcal{N}^L \cup \mathcal{N}^P \cup \mathcal{N}^U)_{n_i \in \mathcal{N}}$ to minimize the loss function $f^L(\theta_G)$, $f^P(\theta_G)$, and $f^U(\theta_G)$ over **labelled, partially labelled and unlabelled clients**, respectively:

$$\boxed{\min_{\theta_G} f(\theta_G) = \frac{1}{\mathcal{N}^L}\sum_{\ell=1}^{\mathcal{N}^L} \mathcal{L}^L(x_\ell^L, y_\ell^L, \theta_G) + \frac{1}{\mathcal{N}^P}\sum_{\ell=1}^{\mathcal{N}^P} \mathcal{L}^P(x_\ell^P, y_\ell^P, \theta_G) + \frac{1}{\mathcal{N}^U}\sum_{\ell=1}^{\mathcal{N}^U} \mathcal{L}^U(x_\ell^U, y_\ell^U, \theta_G)}$$

$\mathcal{L}$ is task-specific loss function on clients with labelled, partial labelled and unlabelled data.

2-Phase Federated Self-Learning Framework

Phase 1: Engagement of Labelled & Partially Labelled Clients:

Phase 1 trains a global pseudo-labeling model $\boldsymbol{\theta}_{\boldsymbol{G}}^{(1)}$ from decentralized labelled and partially labelled client $n_i \epsilon \mathcal{N}^L \cup \mathcal{N}^P$, using the ground-truth labels optimizing the loss:

$$\boxed{\boldsymbol{\theta}_{\boldsymbol{G}}^{(1)} = \boldsymbol{min} \left[\frac{1}{\mathcal{N}^L} \sum_{\ell=1}^{\mathcal{N}^L} \mathcal{L}_{CE}\left(x_\ell; (\boldsymbol{\theta}_{\boldsymbol{G}}^{(1)}), y_\ell\right)\right]}$$

$\mathcal{L}_{CE}$ is cross-entropy loss and g(·; ·) represents the classifier.
At round $t \leq T_1$, $\boldsymbol{\theta}_{\boldsymbol{G}}^{(1)}$ are disseminated to each labelled client n_i locally updating over E local epochs:

$$\boldsymbol{\theta}_{\boldsymbol{i}}^{\boldsymbol{t,e+1}} = \theta_i^{t,e} - \eta_t \nabla f_t\left(\theta_i^{t,e}\right), e = 1, \ldots, E.$$

After completion of epochs, each client $n_i \epsilon \mathcal{N}^L$ sends its local model $\theta_i^{t,E}$ to the server for aggregation:

$$\theta_{G,t}^{(1)} = \frac{1}{|\mathcal{N}^L|} \sum_{n_{i \in \mathcal{N}^L}} \theta_i^{t,E}$$

2-Phase Federated Self-Learning Framework

Phase 1: Engagement of Labelled & Partially Labelled Clients:

At each round t, $\boldsymbol{\theta}_{\boldsymbol{G,t}}^{(\mathbf{1})}$, is distributed to **each** partially labelled client $n_i \epsilon \mathcal{N}^p$ to be used for pseudo-labeling of partially labelled samples in the subsequent training rounds.

Each unlabelled client $n_i \epsilon \mathcal{N}^U$ uses $\boldsymbol{\theta}_{\boldsymbol{G,t}}$ to predict the label $\hat{y}_u$ for the unlabelled input x_u based on previous knowledge captured from previous rounds $\tau < t$.

Select the class $c \ \epsilon \ \mathcal{C}$ with maximum predicted confidence from $\boldsymbol{\theta}_{\boldsymbol{G,t}}$,
i.e., the pseudo-label for x_u is $\hat{y}_u$ = c, such that:

$$c = \arg max_{c' \in \mathcal{C}} \quad p\, \boldsymbol{\theta}_{\boldsymbol{G,t}}(c' | x_u) \ \geq \varphi$$

2-Phase Federated Self-Learning Framework

Phases 2 & 2+: Engagement of Unlabelled Clients & Fine-tuning:

The unlabelled clients (along with the rest) are engaged in Phase 2 to enhance the robustness of the global $\boldsymbol{\theta}_{\boldsymbol{G}}^{(2)}$.

We **progressively** incorporate pseudo-labelled samples with high confidence obtained from previous rounds into the subsequent.

Benefit: This allows the global model to generate increasingly high-quality pseudo-labels for unlabelled samples in unlabelled clients.

Experimental Evaluation

Experimental Set-up:

- Images: MNIST, EMNIST, MEDMNIST, Fashion-MNIST; classes $|\mathcal{C}|$ = (10, 47, 6, 10), respectively.
- Number of samples per class differs from one client to another (non-iid).
- Clients: $|\mathcal{N}| \in \{10, 20, 50\}$, split the clients into **Types I, II and III** based on the ratio **2:3:5**.

Baselines

- **Baseline 1:** FL benchmark (**FedAvg**): all clients have **fully labelled data without class imbalance**.
- **Baseline 2: PL-FL**, which involves only **Type II** clients. All clients have **partially labelled data with class imbalance**.
- **Baseline 3: L&PL-FL**, which involves **Type I & II** clients **with class imbalance**.

Experimental Results

Impact of pseudo-labeling confidence on training phases

Dataset	Method	Phase1	Phase2	Phase2+
MNIST	**2PFL**	**96.93%**	**95.02%**	**97.31%**
	FedAvg	88.07%	88.67%	86.29%
	PL-FL	79.65%	85.10%	85.10%
	L&PL-FL	88.59%	90.01%	90.01%
F-MNIST	**2PFL**	**86.24%**	**88.05%**	**89.01%**
	FedAvg	81.15%	83.18%	82.16%
	PL-FL	76.70%	75.81%	75.77%
	L&PL-FL	71.43%	75.60%	72.43%
EMNIST	**2PFL**	**94.4%**	**94.8%**	**96.00%**
	FedAvg	72.47%	86.10%	84.35%
	PL-FL	53.30%	77.72%	83.45%
	L&PL-FL	84.38%	79.37%	78.20%
MEDMNIST	**2PFL**	**95.38%**	**98.53%**	**98.92%**
	FedAvg	54.69%	74.39%	71.41%
	PL-FL	49.76%	67.79%	59.54%
	L&PL-FL	86.45%	78.90%	74.88%

Experimental Results

Comparison assessment with baselines

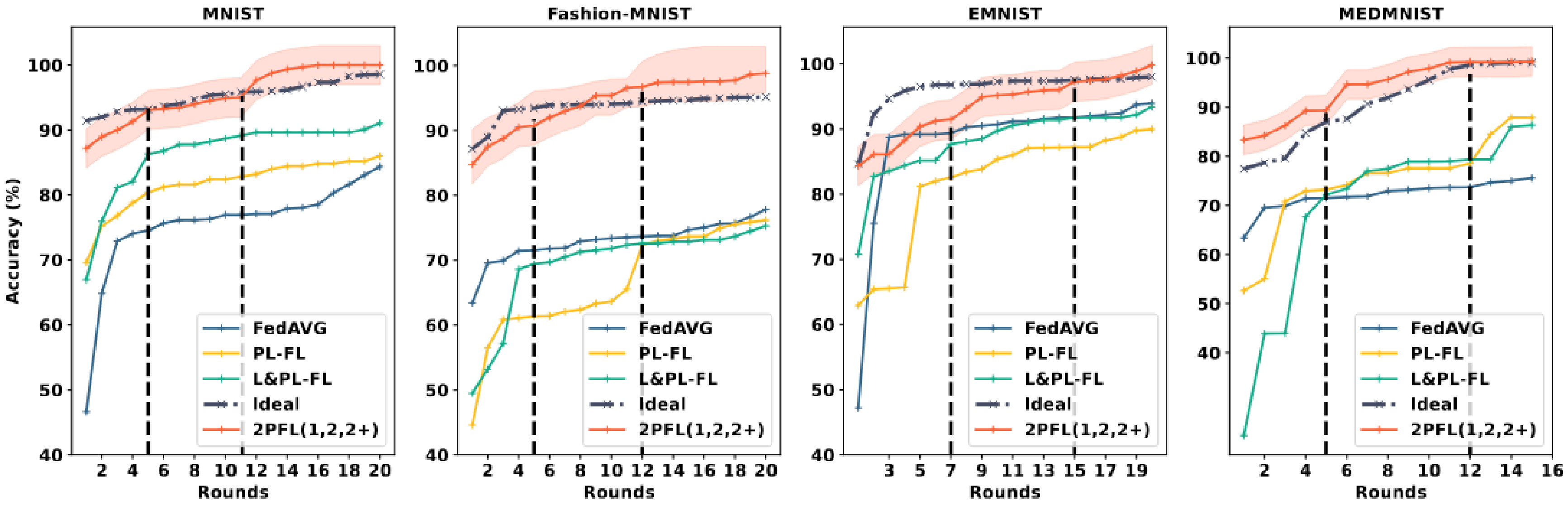
Dataset	Performance	Baselines				2PFL		
		Ideal	FedAvg	PL-FL	L&PL-FL	Phase1	Phase2	Phase2+
MNIST	**Accuracy**	97.92%	88.59%	79.65%	88.67%	96.93%	95.02%	**97.31%**
	LDR, $\phi \in (0.5, 0.9)$	87.08%	35.25%	36.22%	49.31%	80.51%	82.78%	**94.70%**
	Rounds	20	20	32	20	10	11	5
F-MNIST	**Accuracy**	88.76%	79.89%	76.70%	71.43%	86.24%	88.05%	**89.01%**
	LDR, $\phi \in (0.5, 0.7)$	73.26%	20.11%	20.39%	49.31%	63.98%	70.77%	**88.80%**
	Rounds	20	20	20	20	10	7	5
EMNIST	**Accuracy**	96.40%	72.47%	53.30%	84.38%	94.4%	94.80%	**96.00%**
	LDR, $\phi \in (0.5, 0.9)$	66.3%	34.3%	39.37%	24.1%	63.525	67.07%	**76.55%**
	Rounds	20	18	15	20	10	10	8
MedMNIST	**Accuracy**	98.09%	54.69%	49.76%	86.45%	95.38%	98.53%	**98.92%**
	LDR, $\phi \in (0.5, 0.9)$	84.1%	26.53%	31.7%	20.22%	51.02%	60.57%	**82.91%**
	Rounds	30	20	20	20	10	5	7

University of Glasgow

School *of* Computing Science
Knowledge & Data Engineering Systems

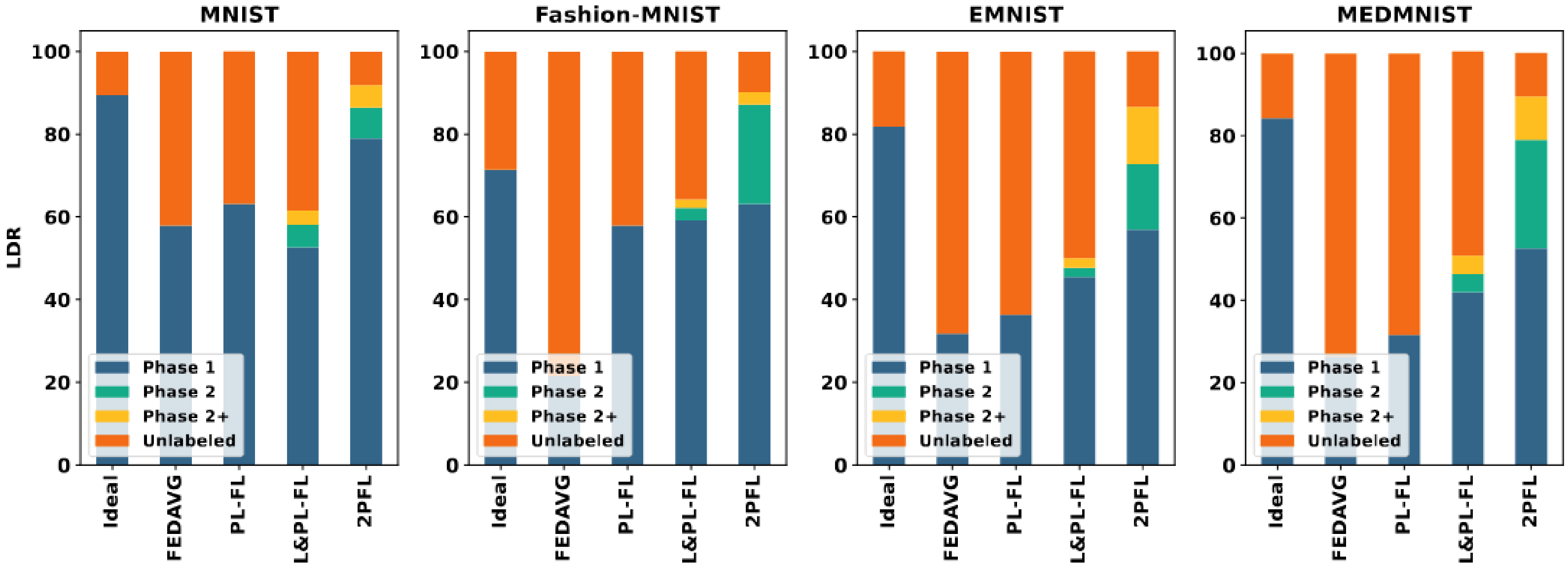
Experimental Results

Comparison assessment with baselines (across datasets)



Experimental Results

Impact of phases on model convergence & pseudo-labeling efficiency



Conclusions

- Our **2PFL** framework addresses the challenge of training FL models across different **types of clients** with limited and skewed labeled and unlabelled data.
- By leveraging data augmentation, 2PFL leads to improved model performance and accelerates convergence by progressive pseudo-labelling.
- Our experiments highlight that 2PFL consistently outperforms baselines across various performance metrics and datasets.

The price for learning a global model with skewed and unlabeled data is <u>minimal</u> with 2PFL

University of Glasgow

School *of* Computing Science
Knowledge & Data Engineering Systems

Thank you!

Tahani Aladwani