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

KDES: Knowledge & Data Engineering Systems

Tahani Aladwani, Christos Anagnostopoulos, Shameem Parambath, Fani Deligianni

European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2024)

Introduction & Overview

Federated Learning (FL) is a distributed learning paradigm that allows multiple clients to collaboratively train Deep Learning (DL) models without sharing their private raw data.

Ideal Assumptions in Federated Learning (FL):

- **Supervised Learning**: All clients possess training data with corresponding ground-truth labels.
- **Semi-Supervised Learning**: A subset of clients have access to adequately labeled data.
- **High-Quality Pseudo-Labels**: The model generates pseudo-labels for unlabeled data using only labeled data available during training.

DID YOU KNOW?

In real-world FL senarios:

- Data can be **non-IID**.
- Data across clients can be unlabeled , due to e.g., limited resources, labeling costs, human errors, etc.

Problem Fundamentals

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

Overview of our Idea

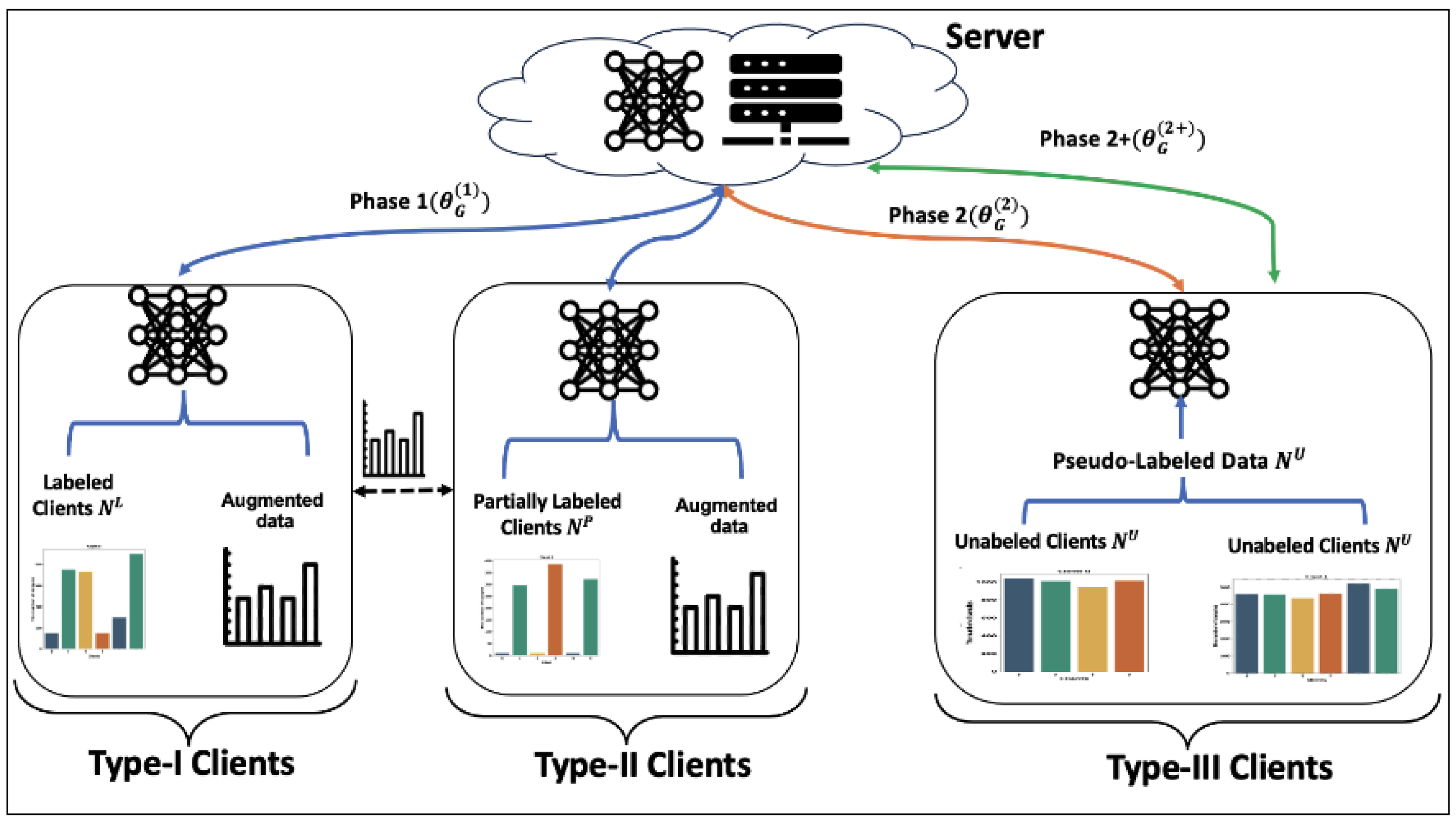
A set $\mathcal{N} = \{n_1, \dots, n_N\}$ of distributed clients categorized into:

– **Type I** clients (**labelled clients**) having all data labeled.

– **Type II** clients (**partially labelled clients**) having labelled & unlabelled data

– **Type III** clients (**unlabelled clients**) where all data are unlabelled

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



Idea 1: Local Data Augmentation

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

In labelled/partially labelled client: for any two inputs x_k and x_ℓ with labels y_k and y_ℓ, MixUp synthesizes the sample (x', y'):

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

Idea 2: 2PFL Training Phases

2PFL exploits labelled, partially labelled and unlabelled data across 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 & unlabelled clients:

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

Phase 1: Engagement of Labelled & Partially Labelled Clients

Phase 1 trains a global pseudo-labeling model $\boldsymbol{\theta_G^{(1)}}$ from labelled data, using the ground-truth labels optimizing the loss:

$$\boldsymbol{\theta_G^{(1)}} = \boldsymbol{min}\, [\tfrac{1}{N^L}\textstyle\sum_{\ell=1}^{N^L}\mathcal{L}_{CE}\left(x_\ell; (\boldsymbol{\theta_G^{(1)}}), y_\ell\right)]$$

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

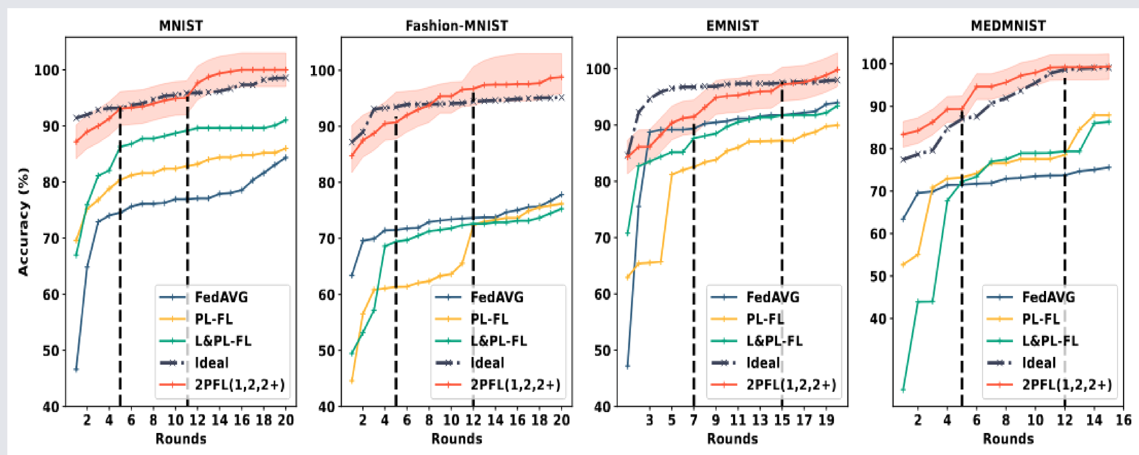
The unlabelled clients (along with the rest) are engaged in Phase 2 to enhance the robustness of the global $\boldsymbol{\theta_G^{(2)}}$.

Experimens

❖ **Impact** of pseudo-labeling confidence on training phases

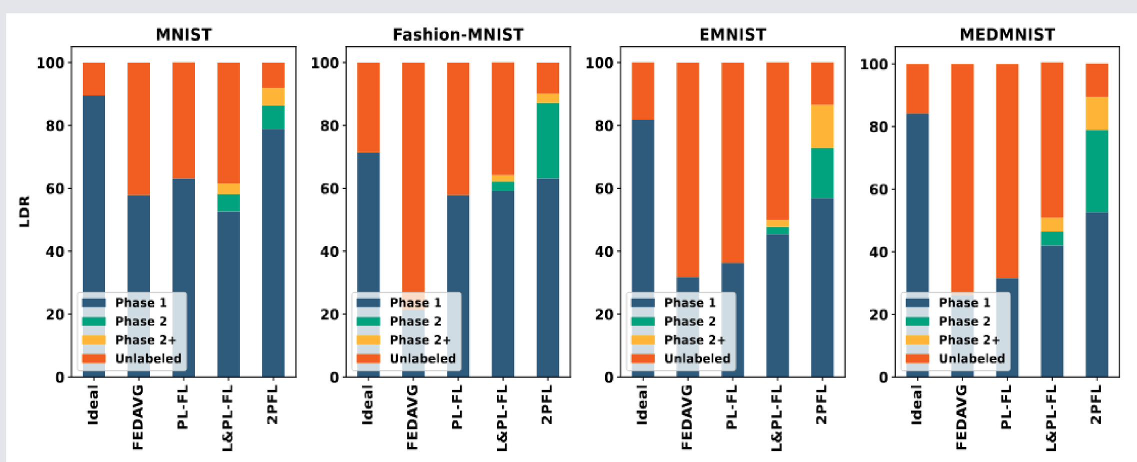
Dataset	Performance	Baselines: Ideal	FedAvg	PL-FL	L&PL-FL	2PFL: Phase1	Phase2	Phase2+
MNIST	Accuracy	97.92%	88.59%	79.65%	88.67%	96.93%	95.02%	97.31%
	LDR, $\phi \in (0.5, 0.9)$	87.08%	35.25%	36.22%	49.31%	80.51%	82.78%	**94.70%**
	Rounds	20	20	32	20	10	11	5
F-MNIST	Accuracy	88.76%	79.89%	76.70%	71.43%	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

❖ **Comparison across datasets**



- Effectiveness and efficiency of 2PFL against baselines w.r.t **test accuracy, Labelled Data Ratio (LDR), number of training rounds.**
- Accuracy vs. training rounds over all datasets (vertical dotted lines correspond to **T1, T1 + T2** rounds of 2PFL's phases).

❖ **Impact** of phases on model convergence & pseudo-labeling efficiency



- **Pseudo-labelling ratio** of unlabelled samples across datasets and phases.

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 minimal with 2PFL

This work is partially funded by the EU Horizon Grant `Integration and Harmonization of Logistics Operations' TRACE (\#101104278).

Horizon Europe