Predictive Diagnostics for Federated Learning:
A Privacy-Preserving Toolbox Towards Successful Federated Learning

Universität Bielefeld
Faculty of Technology

Christoph Düsing
CITEC, Bielefeld University, Germany

Motivation

FL suffers from data imbalance among clients, causing the performance of the jointly trained model to decrease [6]. More importantly, depending on the data provided, clients were shown to **vary greatly in terms of their benefit from participation and contribution** to the federation [2, 1], where certain clients contribute more towards the success of the federation without benefiting to the same extent [1]. Consequently, FL becomes both **less fair and reliable** when data is imbalanced. Existing studies require full access to data distributions of clients and measure benefit and contribution only retrospectively, i.e., after training the federated model [3]. Both of these constrains severely **limit real-world applicability**, as (1) granting full access to clients' data undermines the benefit of FL and (2) requiring all computations prior to measurement significantly increases computational costs for involved clients. To alleviate these drawbacks, in my dissertation, I will introduce **Predictive Diagnostics for FL (PDFL)**, a toolset utilizing federated analytics and secure aggregation to **identify determinants of successful FL and client participation**.

PDFL Toolbox

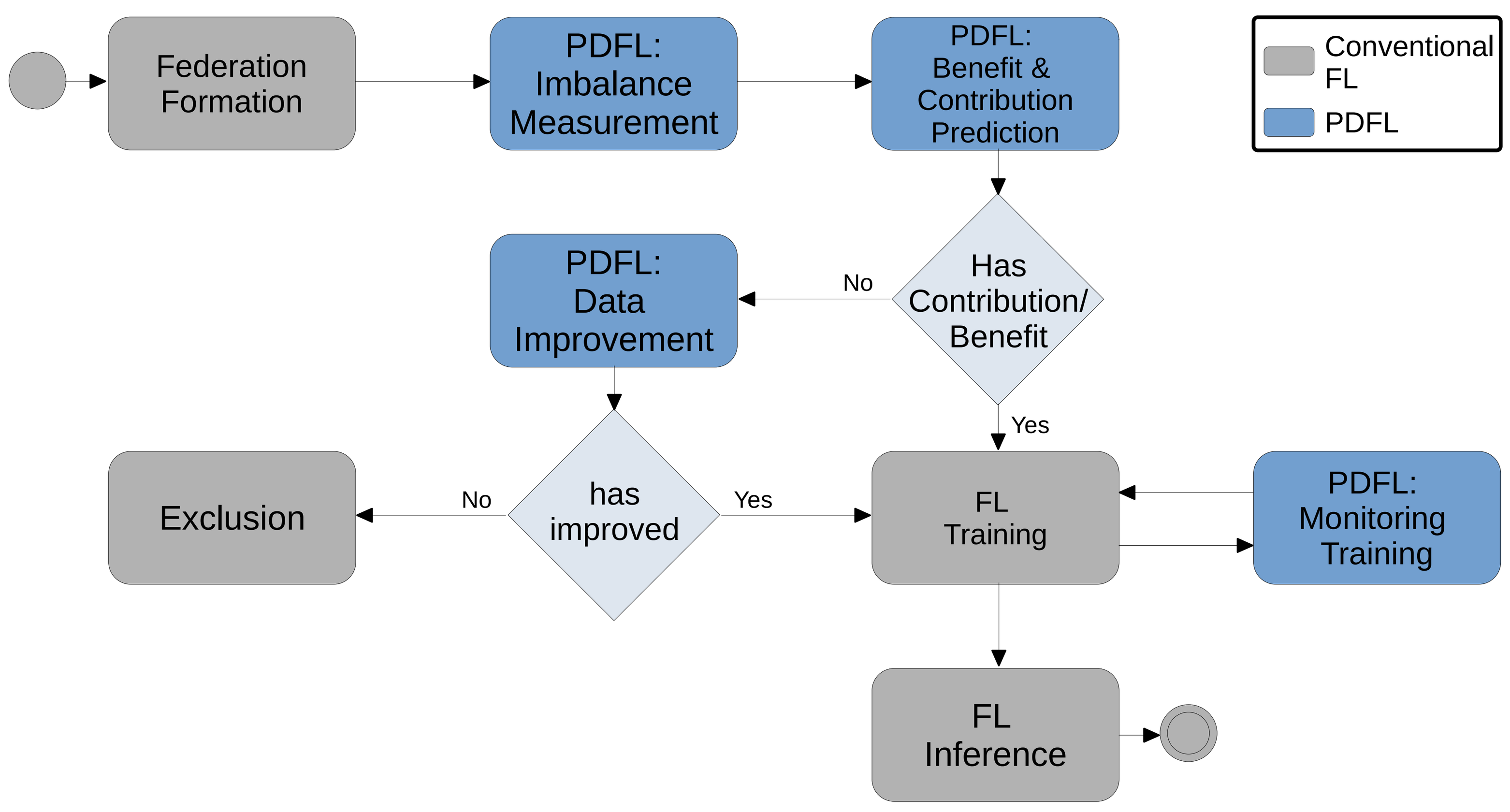

Fig. 1: Process of FL with different PDFL methods applied

Imbalance Measurement

Input: Data held by each client without having to share it with others.

Relying on Secure Aggregation to compute global label distribution $\vec{V} = [\sum_j N_j^1, ..., \sum_j N_j^Q]$ and number of samples N. Afterwards, each client computes local data imbalances [2], namely:

Label Imbalance $LI_j = \frac{max_p\{N_j^p\}}{min_p\{N_j^p\}}$

Label Distribution Imbalance $LDI_j = 1 - \frac{\vec{v}_j \cdot \vec{V}}{||\vec{v}_j|| \cdot ||\vec{V}||}$

Quantity Imbalance $QI_j = N_j / \frac{\sum_{l=1}^{J} N_l}{J}$.

Proposal to also utilize federated clustering to compute feature imbalance [1]:

Feature Imbalance $FI_j = \frac{|C_j \cap D_j|}{|C_j|}$

Output: Measurements of different types of data imbalance.

Predicting Benefit & Contribution

Input: Measurements of global and local data imbalance.

Previous imbalance measurements demonstrate predictive potential for both client benefit and contribution [2, 1]. In turn, I train classifiers and regressors to predict whether and to what extent clients benefit from and contribute to FL [2]. Here, clients' imbalance measurements serve as feature representation, e.g., $\vec{x_j} = [LI_j, LDI_j, QI_j]$.

Output: Predictions of client benefit and contribution; binary or numerical.

Data Improvement

Input: Each client's local dataset.

Applying different approaches for local data sampling, which serve to decrease data imbalance. Ultimately, this improves the performance and convergence of FL models. Among others, I apply the following local data sampling strategies:

Undersampling. Random undersampling of the majority classes at each client to match the size of their respective minority classes.

Oversampling. It focuses on adding minority class samples to match the size of the majority class using SMOTE.

Hybrid sampling. Hybrid data sampling combines undersampling the majority classes and oversampling minority classes in order to balance the dataset.

Output: Balanced local dataset that improve FL performance.

Monitoring Training

Input: Train performance (loss) and client contribution.

Relying on data- and algorithm-based approaches to detect changing data imbalance and concept drift in FL applications with dynamic client participation. The goal is to identify concept drift as soon as possible.

Output: Information about training performance and concept drift.

Background: Federated Learning

A **distributed learning paradigm** allowing mutually distrustful clients to **jointly train a machine learning model** while **maintaining data privacy** [4, 5]. A joint model is trained during **several rounds**. Each round consists of:

1. A central server sending a global model to all clients
2. Each client fitting the model to their respective data
3. Clients sending local model updates back to the server
4. The central server aggregating all model updates to a new global model

Preliminary Results

Imbalance Measurement

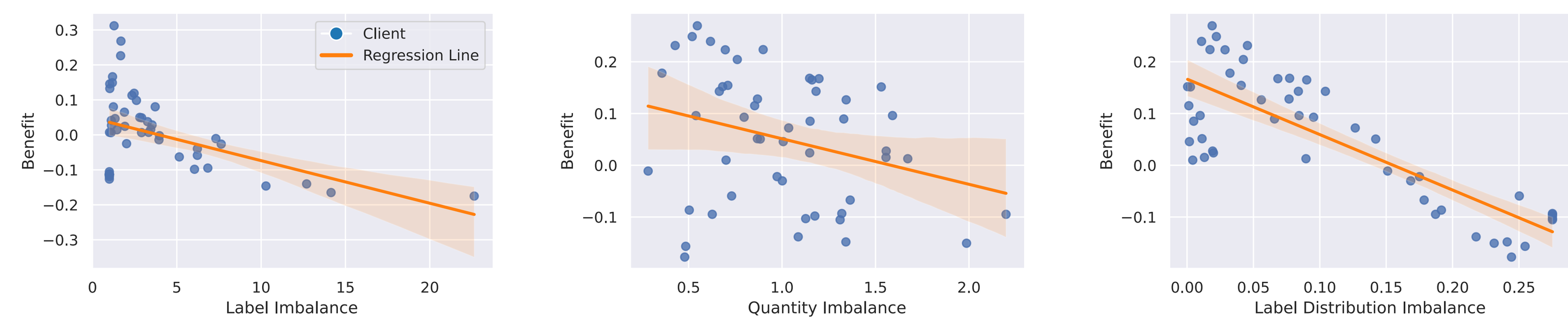

Fig. 2: Effects of Data Imbalance on Benefit [2]

Predicting Benefit & Contribution

	Benefit		Contribution	
Name	Acc	Std. Dev.	Acc	Std. Dev
Covtype	**0.8928**	±0.0253	0.6005	±0.0822
Adult	0.6806	±0.0634	0.6870	±0.0798
Diabetes	0.7395	±0.0646	**0.7673**	±0.0370
Postures	0.6075	±0.0456	0.7378	±0.0866
MNIST	0.7344	±0.0720	0.6364	±0.0787
CIFAR10	0.5804	±0.0610	0.4980	±0.0445
Mean	0.7059	-	0.6545	-

Tab. 1: Predicting Benefit and Contribution (Classification) [2]

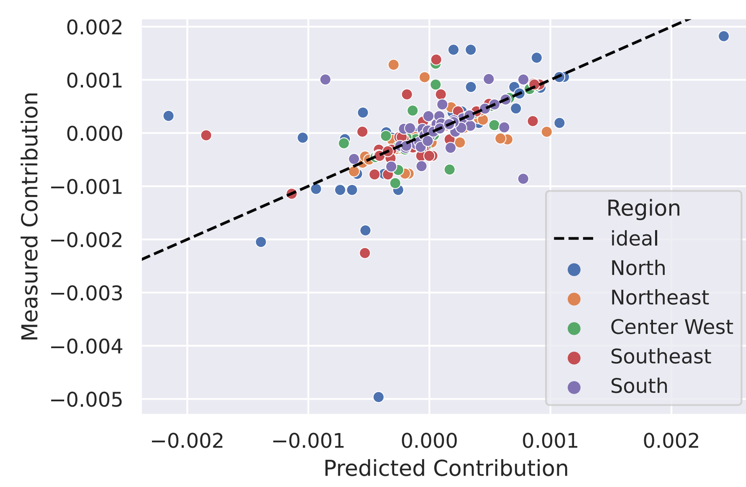

Fig. 4: Predicting Client Contribution (Regression)

Data Improvement

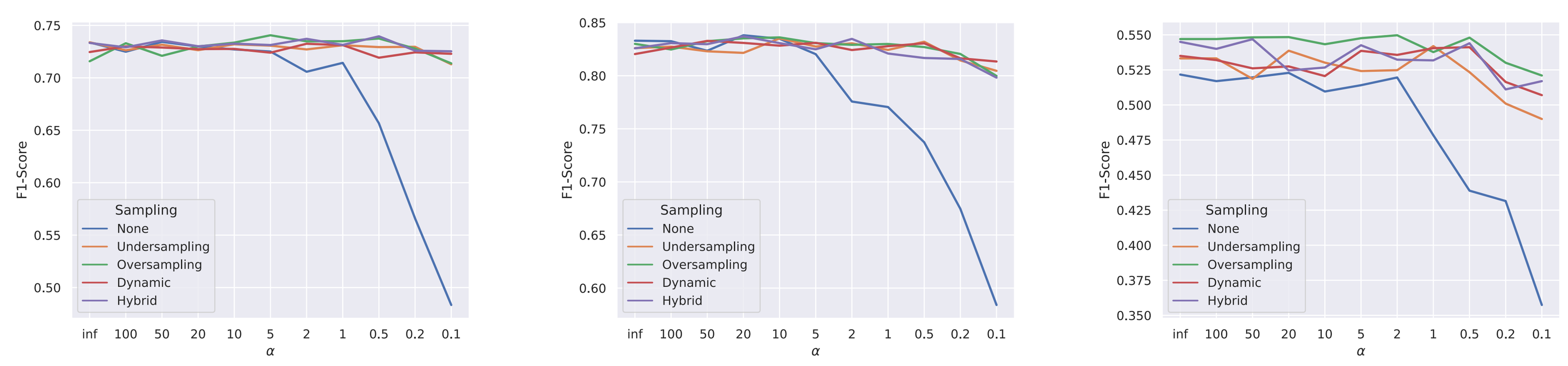

Fig. 3: Improvements of FL Performance through Local Data Sampling among three Different Datasets

Next Steps

1. Analyzing and extending data improvement strategies
2. Visualizing measured imbalances, especially to monitor dynamic FL
3. Integration of different methods into a holistic framework
4. Publishing an open-source library for straight-forward deployment

References

[1] Christoph Düsing and Philipp Cimiano. "On the Trade-off Between Benefit and Contribution for Clients in Federated Learning in Healthcare". In: *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2022, pp. 1672–1678.

[2] Christoph Düsing and Philipp Cimiano. "Towards predicting client benefit and contribution in federated learning from data imbalance". In: *Proceedings of the 3rd International Workshop on Distributed Machine Learning*. 2022, pp. 23–29.

[3] Jiyue Huang et al. "An exploratory analysis on users' contributions in federated learning". In: *2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*. IEEE. 2020, pp. 20–29.

[4] Peter Kairouz et al. "Advances and open problems in federated learning". In: *Foundations and Trends® in Machine Learning* 14.1–2 (2021), pp. 1–210.

[5] Brendan McMahan et al. "Communication-efficient learning of deep networks from decentralized data". In: *Artificial intelligence and statistics*. PMLR. 2017, pp. 1273–1282.

[6] Yue Zhao et al. "Federated learning with non-iid data". In: *arXiv preprint arXiv:1806.00582* (2018).