New Metrics to Evaluate the Performance and Fairness of Personalized Federated Learning

Siddharth Divi [1]   Yi-Shan Lin [1]   Habiba Farrukh [1]   Z.Berkay Celik [1]

Abstract

In Federated Learning (FL), the clients learn a single global model (FedAvg) through a central aggregator. In this setting, the non-IID distribution of the data across clients restricts the global FL model from delivering good performance on the local data of each client. Personalized FL aims to address this problem by finding a personalized model for each client. Recent works widely report the average personalized model accuracy on a particular data split of a dataset to evaluate the effectiveness of their methods. However, considering the multitude of personalization approaches proposed, it is critical to study the per-user personalized accuracy and the accuracy improvements among users with an equitable notion of fairness. To address these issues, we present a set of performance and fairness metrics intending to assess the quality of personalized FL methods. We apply these metrics to four recently proposed personalized FL methods, PersFL, Fed-Per, pFedMe, and Per-FedAvg, on three different data splits of the CIFAR-10 dataset. Our evaluations show that the personalized model with the highest average accuracy across users may not necessarily be the fairest. Our code is available at https://tinyurl.com/1hp9ywfa for public use.

1. Introduction

Federated Learning (FL) is a distributed collaborative learning paradigm that does not require centralized data storage in a single location. Instead, a joint global predictor is learned by a network of participating users (McMahan et al., 2016). FL is useful when the clients have sensitive data

that they cannot share with the participating entities due to privacy concerns. Yet, despite its widespread applications, FL faces different challenges, such as expensive communication, systems heterogeneity, statistical heterogeneity, and privacy concerns (Li et al., 2020). Among these, statistical heterogeneity has recently gained attention.

Statistical Heterogeneity Problem. Statistical heterogeneity means that the clients have unbalanced and non-identical and independently distributed (non-IID) data. This causes the global model (FedAvg) trained on non-IID data of clients not to generalize well on the clients' local data. Consider the task of predicting the next word on a smartphone, which enables users to express themselves faster. In such settings, a global model learned collaboratively fails to give personalized suggestions to each user as each user has a unique way of expressing themselves in applications, such as text messaging and writing e-mails. On the other hand, learning a local model without user collaboration might yield large model error due to the lack of data.

Personalized Federated Learning. The personalized learning methods aim to address this problem by learning a personalized model for each client that benefits from the data of the other clients while at the same time overcoming the problem of statistical heterogeneity. These methods learn a personalized model by extending meta-learning, local fine-tuning, multi-task learning, model regularization, contextualization, and model interpolation (See Section 2).

These efforts often solely report the average accuracy of personalized models across all users to measure their effectiveness. Yet, the average accuracy does not capture the notion of per-user personalization, as per-user performance is aggregated into a single (averaged) accuracy metric. Furthermore, they do not measure the fairness of the personalized models from an equitable notion, the concept that the users get similar improvements (Li et al., 2019). The lack of fairness analysis makes it difficult to compare how different personalized models perform on each user. Lastly, these works employ different data-split strategies among users though they often use standard datasets, such as MNIST and CIFAR-100. Overall, these issues make it difficult for a uniform comparison of the effectiveness of each method.

---

[1]Department of Computer Science, Purdue University, USA. Correspondence to: Siddharth Divi <siddhudivi@gmail.com>.

Contributions. We present a set of metrics in two groups, five metrics for performance and four metrics for fairness, to assess the quality of personalized FL methods, supporting the existing evaluation metrics. Performance metrics express how well the personalized model performs over each user's local and global model accuracy. On the other hand, fairness metrics express an equitable notion that quantifies whether the personalized models provide an equal improvement upon each user's local and global models. The metrics allow for quantitatively contrasting the trade-off between fairness and per-user accuracy of the personalized models under different datasets and data splits.

To motivate the need for new metrics for personalized FL, we have surveyed 12 recent works with the goal of studying their datasets, data splitting strategies, and reported evaluation metrics. We found that these works often use a different data splitting strategy on different datasets and solely report the average accuracy improvement of the personalized model over the global model. We evaluate the proposed performance and fairness metrics on four recent personalized FL methods across three different data splits on the CIFAR-10 dataset. Our evaluation results show that the personalized model with the highest average accuracy across users need not necessarily be the fairest.

2. Related Work

There exist several recent methods proposed for personalization in FL. In local fine-tuning, each user adapts a copy of the global FedAvg model to their local data distribution through gradient-based meta-learning, transfer learning (Pan & Yang, 2010) and domain-adaptation (Mansour et al., 2009). For instance, persFL (Divi et al., 2021) combines the idea of generalized distillation with optimal teacher models for each user to learn more personalized models. Per-FedAvg (Fallah et al., 2020) uses Model-Agnostic Meta-Learning (Finn et al., 2017) to learn a common initialization point for each user during training, which is then subsequently adapted to each user's local data distribution. Fed-Per (Arivazhagan et al., 2019) views the network as a combination of base and personalization layers where the base layers are learned collaboratively, and the personalized layers are specific to each user.

Previous works have also explored contextualization, which aims at learning a model under different contexts. This problem is studied in the next character recognition task (Hard et al., 2018), which needs access to features about the context during the training phase. Local-Global Federated Averaging (LG-FedAvg) (Liang et al., 2020) learns compact local representations on each local and the global model across all users, i.e., an ensemble of local and global models.

Models can also be personalized to each user by regular-

Table 1: Example scenario to motivate the need for alternative metrics in Personalized FL.

	Local Model	FedAvg	Per-1	Per-2	Per-3	Per-4
Datasplits	DS-1	DS-1	DS-1	DS-2	DS-3	DS-1
Users						
User0	73%	78%	82%	79%	76%	75%
User1	71%	75%	82%	74%	72%	72%
User2	61%	69%	82%	75%	68%	68%
User3	55%	71%	75%	79%	82%	96%
User4	69%	74%	74%	78%	85%	97%
User5	65%	77%	75%	89%	87%	75%
User6	74%	80%	77%	74%	78%	76%
User7	68%	82%	77%	76%	79%	78%
User8	75%	85%	78%	79%	79%	77%
Avg. Acc.	67.89%	76.78%	78%	78.11%	78.44%	79.33%

izing the differences between the global and local models. pFedMe (Dinh et al., 2020) uses Moreau envelopes (Moreau, 1963) as a regularization term to learn personalized models and the global FL model parallelly. Federated Mutual Learning (FML) (Shen et al., 2020) uses the non-IID nature of the data as a feature to learn personalized models.

Lastly, model interpolation techniques focus on the mixture of the local and the global models. In Adaptive Personalized FL (APFL) (Deng et al., 2020), an optimal mixing parameter that controls the trade-off between local and global models is integrated into the learning problem. A recent work (Mansour et al., 2020) has proposed the use of user clustering, data interpolation, and model interpolation for personalized models. In LotteryFL (Li et al., 2020), the authors adopt a Lottery Ticket Network through the application of the Lottery Ticket Hypothesis (Frankle & Carbin, 2019) to learn personalized models for each user.

3. Problem Statement

We have studied 12 recent personalized FL methods to identify their datasets, data-split strategies, the metrics other than those commonly used in the FL settings, the approaches they are compared to, and whether they perform a fairness analysis (See Table 2). We observed two main issues in their evaluation of personalized models, which hinder the interpretability of the personalized FL methods. Below we provide an example scenario and present these issues.

Motivating Example. We consider 9 users that collaboratively learn a global model for the next-character prediction on the keypad of their mobile phones. The dataset is distributed to each user to mimic the non-IID nature of real-world data distributions based on a particular data-split strategy (**DS-1**, **DS-2**, and **DS-3**). Each user learns a local model, a global model (FedAvg), and a personalized model using four different personalized FL methods (**Per-1**-**Per-4**). The personalized models are specific to each user and aim to yield better accuracy than local and FedAvg models. Table 1 presents the accuracy of local models, FedAvg model, and four personalized models on different datasplits.

Table 2: The analysis results of studied related personalized FL methods.

#	Method	Only FL metrics[†]	Fairness Analysis	Datasets[‡]	Use of Custom Datasplit[*]	Comparison
	LOCAL FINE-TUNING					
1	**APFL** (Deng et al., 2020)	✓	✗	(1), (2), (3), (6)	✓	FedAvg, SCAFFOLD, Per-FedAvg, pFedMe
2	**pFedMe** (Dinh et al., 2020)	✓	✗	(1), (6)	✓	FedAvg, Per-FedAvg
3	**Per-FedAvg** (Fallah et al., 2020)	✓	✗	(1), (3)	✓	FedAvg
4	**FedPer** (Arivazhagan et al., 2019)	✓	✗	(3), (4), (5)	✓	FedAvg
5	**Three Approaches for Personalization** (Mansour et al., 2020)	✓	✗	(1), (2)	✓	FedAvg, AGNOSTIC
6	**Personalized FedAvg** (Jiang et al., 2019)	✓	✗	(2), (7)	✓	FedAvg
7	**FedMeta** (Chen et al., 2018)	✓	✓	(7), (8), (9), (10)	✓	FedAvg
	MULTI-TASK LEARNING					
8	**MOCHA** (Smith et al., 2017)	✓	✗	(11), (12), (13)	✓	FedAvg
	CONTEXTUALIZATION					
9	**LG-FedAvg** (Liang et al., 2020)	✓	✗	(6), (1), (3), (14)	✓	FedAvg, FEDPROX
	MODEL REGULARIZATION BASED PERSONALIZATION					
10	**FedAMP** (Huang et al., 2021)	✓	✗	(1), (15), (2), (4)	✓	SCAFFOLD, APFL, FedAvg, FEDPROX
11	**FML** (Shen et al., 2020)	✓	✗	(1), (3), (4)	✓	FedAvg, FEDPROX
	MODEL INTERPOLATION BASED PERSONALIZATION					
12	**LotteryFL** (Li et al., 2020)	✓	✗	(1), (3), (2)	✓	FedAvg, LG-FedAvg

† Whether the personalization method only reports metrics in the FL domain such as training loss, average validation accuracy, prediction error, and the number of communication rounds. ‡ (1) MNIST, (2) EMNIST, (3) CIFAR-10, (4) CIFAR-100, (5) FLICKR-AES, (6) Synthetic, (7) Shakespeare, (8) FEMNIST, (9) Sentiment 140, (10) Industrial recommendation task, (11) Google Glass (GLEAM), (12) Human Activity Recognition (HAR), (13) Vehicle Sensor, (14) Mobile Assessment for Prediction of Suicide (MAPS), (15) FMNIST. ∗ Whether the personalization method uses a custom data split technique.

Missing Per-User Accuracy and Fairness Analysis. The personalized FL methods often solely report the average accuracy of the personalized model across all the users to measure their model effectiveness. In Table 1, we ask the question of *which personalization method yields the best performance in terms of per-user personalized accuracy*. In this example, **Per-4** gives the highest average accuracy of 79.3% on **DS-1**. However, upon closer inspection, we observe that, with respect to the FedAvg model, **Per-4** increases the accuracy of only 2 out of the 9 users. This shows that the average accuracy of the personalized model may fail to fully characterize the quality of a method. Here another observation is that the best performing personalized method, on average, may not necessarily lead to an improvement over the local or global models across all users.

Through surveyed methods in Table 2, we observe that the methods commonly adapt the evaluation metrics from FL, such as the training loss, average validation accuracy, prediction error, number of communication rounds. This means that the methods often do not report the per-user accuracy (Table 2 "Only FL metrics" column). Out of 12 studied methods, only the FedMeta approach performs a fairness analysis among personalized models of users by reporting the per-user accuracy of their method.

Inconsistent Datasets and Data-splits. A data-split strategy is used in personalized FL to split the dataset across users such that each receives a fraction of the non-IID data. The data distribution among users is a crucial feature in personalization because if the data is distributed IID, the

personalized model may not offer any benefits over FedAvg (Deng et al., 2020). The personalized methods often use the same dataset yet the data-splits are different. For instance, in Table 1, **Per-1**, and **Per-4** are trained on **DS-1**, whereas **Per-2** is on **DS-2** while **Per-3** is on **DS-3**. The use of different data splits on the same dataset makes it difficult to interpret the effectiveness of personalized models. When personalized models are compared with each other, each method often reports their results on a new (different) data split than the one used in the compared approach. For example, **Per-1** and **Per-2** cannot be directly compared as they are trained on different data-splits, **DS-1** and **DS-2**.

Out of the 12 studied personalized FL methods, we observe that 9 methods use standard datasets such as MNIST and CIFAR-10 ((1)-(4), (6), and (7), Table 2 "Datasets column"). Other 3 methods use other datasets, such as FEMNIST and Sentiment 140 ((5), and (8)-(15)). Additionally, the methods often use different custom data splits on the datasets. We identified 10 data-splits in 12 methods (Table 2 "Use of Custom Datasplit" column).

4. Evaluation Metrics

We present a set of metrics for the evaluation of personalized FL models from the performance and fairness perspective. To quantify the per-user accuracy improvements gained in terms of personalization, we compute the metrics on the Quantum of Improvement (QoI) as follows:

$$F_i = P_i - \max(G_i, L_i) \qquad (1)$$

where P, G, and L refer to the accuracy of the personalized model, FedAvg and local model of the user i, and F_i refers to the QoI of user i. Henceforth, we will refer to the QoI as F in all the equations.

The QoI can result in negative values. This means that the personalized method decreases a user's personalized model accuracy rather than the expected increase over the local or global models. In such cases, the direct application of evaluation metrics may misguide the interpretation of results. Therefore, we split the QoI into two sets that contain the absolute QoI values, i.e., a set of users (U^+) who have positive QoI, and a set of users (U^-) who have negative QoI. We then apply the introduced metrics below to both sets and interpret accordingly.

4.1. Performance Metrics

We introduce five performance metrics to express how well the personalized model performs over each user's local and global model.

Percentage of User-models Improved (PUI). PUI is the percentage of users who experience an improvement over their local and global models. Ideally, a personalized model is expected to improve the per-user accuracy of a maximal set of users.

$$PUI = \frac{COUNT(F_i > 0)}{COUNT(U)} \times 100, i \in U(U : Users) \quad (2)$$

In a normal distribution, the mean is the best measure of the central tendency. However, the median might be a better measure of central tendency when this is not the case. We define median and average percentage of improvement since the QoI distribution is not known apriori.

Median Percentage of Improvement (MPI). MPI is computed as $Median(U^+)$ where $Median()$ function returns the median of its input, and U^+ is the QoI of the set of users who obtained an increase in their performance. A personalized model is expected to have a high median of the QoI values among the users who experience an improvement.

Average Percentage of Improvement (API). API is the average percentage improvement among the users who obtained an increase in their performance (U^+).

$$API = \frac{\sum_{i \in U^+} F_i}{len(U^+)} \quad (3)$$

We observe that, in some scenarios, a personalization method does not yield an improvement over users' local and global accuracy. Thus,in such cases, it is crucial to report the per-user accuracy decrease of the personalized model. Because this decrease cannot be derived from the improvement metrics (MPI and API), we define two metrics to quantify the decreased accuracy.

Median Percentage of Decrease (MPD). Similar to MPI, MPD is computed as $Median(U^-)$.

Average Percentage of Decrease (APD). Similar to API, APD is the average percentage decrease among the users whose performance is decreased (U^-).

4.2. Fairness Metrics

We extend four metrics to evaluate personalization methods that yield better results from a fairness perspective. For two personalization methods t and t', the QoI distribution among K users $\{F_1(t), \ldots, F_K(t)\}$ is more fair (uniform) under technique t than t' based on the relation captured by the fairness metric.

Average Variance (AV). Average Variance (AV) (Li et al., 2019) is a measure of the spread of data. For AV, the relation is extended to personalized models as follows:

$$AV(F_1(t), \ldots F_K(t)) < AV(F_1(t'), \ldots F_K(t')) \quad (4)$$

where, $AV(F_1(t), \ldots F_K(t))$ is computed as,

$$AV = \frac{1}{K} \sum_{i=1}^{K} (F_i(t) - \bar{F}(t))^2 \quad (5)$$

$\bar{F}(t)$ in Equation 5 refers to the average QoI across all users and is computed as,

$$\bar{F}(t) = \frac{1}{K} \sum_{i=1}^{K} (F_i(t)) \quad (6)$$

A lower AV means a higher fairness capability for a personalized method.

Cosine Similarity (CS). One of the drawbacks of AV is that the outliers may cause skewing of the data. We measure Cosine Similarity (CS) (Li et al., 2019) for two personalized methods to quantify the similarity between their QoI distributions. For CS, the relation is computed as:

$$CS[(F_1(t), \ldots F_K(t)), 1] \geq CS[(F_1(t'), \ldots F_K(t')), 1] \quad (7)$$

where, $CS[(F_1(t), \ldots F_K(t)), 1]$ is computed as follows.

$$CS = \frac{\frac{1}{K} \sum_{i=1}^{K} F_i(t)}{\sqrt{\frac{1}{K} \sum_{i=1}^{K} F_i^2(t)}} \quad (8)$$

A higher CS means a higher fairness capability for a personalized method.

Entropy. One of the drawbacks of CS is that the magnitude of the QoI values is not taken into consideration, yet only

their orientation is considered. For this reason, we extend Entropy (Li et al., 2019) as follows:

$$\texttt{Entropy}(\texttt{F}_1(\texttt{t}), ... \texttt{F}_K(\texttt{t})) \geq \texttt{Entropy}(\texttt{F}_1(\texttt{t}'), ... \texttt{F}_K(\texttt{t}')) \tag{9}$$

where Entropy is defined as follows.

$$\texttt{Entropy} = -\sum_{\texttt{i}=1}^{\texttt{K}} \frac{\texttt{F}_\texttt{i}(\texttt{t})}{\sum_{\texttt{i}=1}^{\texttt{K}} \texttt{F}_\texttt{i}(\texttt{t})} \log(\frac{\texttt{F}_\texttt{i}(\texttt{t})}{\sum_{\texttt{i}=1}^{\texttt{K}} \texttt{F}_\texttt{i}(\texttt{t})}) \tag{10}$$

A higher Entropy means a higher fairness capability for a personalized method.

Jain's index (JI). JI is widely studied fairness measure in computer networks and resource allocation to identify the underutilized channels (Jain et al., 1998). We extend it for personalization as follows:

$$\texttt{JI} = \frac{[\sum_{\texttt{i}=1}^{\texttt{K}} \texttt{F}_\texttt{i}(\texttt{t})]^2}{\texttt{K} \sum_{\texttt{i}=1}^{\texttt{K}} \texttt{F}_\texttt{i}(\texttt{t})^2} \tag{11}$$

A higher JI means a higher fairness capability for a personalized method.

5. Experimental Results

We evaluate four recently introduced personalization methods, PersFL (Divi et al., 2021), FedPer (Arivazhagan et al., 2019), pFedMe (Dinh et al., 2020), and Per-FedAvg (Fallah et al., 2020), to evaluate their performance and fairness through introduced metrics.

5.1. Experimental Setting

We use CIFAR-10 dataset on three data splits with a total of 10 clients. CIFAR-10 includes 32×32 color images with 10 classes and $60,000$ instances. We use a CNN-based model with two 2-D convolutional layers separated by a MaxPool layer between them and followed by three fully connected (FC) layers. The fully connected layers have 400, 120, and 84 hidden neurons. We use ReLu activations after each layer except the last FC layer.

We make three assumptions in line with the assumptions made in personalized FL literature. First, we assume that all clients are active during the entire training phase to speed up the model convergence. Second, each client's data does not change between the global aggregations. Lastly, the hyper-parameters, batch-size (B), and local epochs (E), are invariant among the simulated clients. We conduct all experiments with a 60%-20%-20% train-validation-test splits.

The experiments are run with Python 3.7, and a PyTorch version of v1.3.1 on an NVIDIA Tesla T4 GPU with 16GB memory with a CUDA version of 10.0.

Data-splitting Strategies. We split the data among users following three different strategies used in the literature.

In **DS-1**, each user has the same total number of samples but may have different classes and a different number of samples per class. The statistical heterogeneity is varied by tuning the parameter k, which controls the number of overlapping classes between each user (Arivazhagan et al., 2019). For example, $k = 4$ corresponds to a highly non-identical data partition, whereas $k = 10$ corresponds to a highly identical data partition across the participating users. We set k to 4 to have non-IID data across users.

In **DS-2** (Yu et al., 2020), all users have samples from all classes, but the number of samples per class they have is different, and hence the total number of samples per user is also different across users. In order to simulate a non-IID distribution, we assign samples from each class to the users using a Dirichlet distribution with $\alpha = 0.9$, following the previous work (Hsu et al., 2019). Each class is parameterized by a vector q where $q \geq 0, i \in [1, N]$ is sampled from a Dirichlet distribution with parameters α and p. The parameter p is the prior class distribution over the classes, and α is the concentration parameter that controls the data similarity among the users. If $\alpha \to \infty$, all users have an identical distribution to the prior. If $\alpha \to 0$, each user only has samples from one class randomly chosen.

In **DS-3**, each user has two of the ten class labels. Additionally, the total number of samples per user is different, i.e., the users do not have the same number of total samples. The samples assigned to users are drawn from a log-normal distribution with the parameters $\mu = 0$ and $\sigma = 2$ (Dinh et al., 2020). The parameters μ and σ correspond to the underlying normal distribution from which we draw samples.

5.2. Experimental Results

We evaluate four personalized learning methods to answer the following two questions:

1. Which personalization method performs the best in terms of per-user personalized accuracy across all users?

2. Which algorithm is the fairest?

Table 3 presents the per-user accuracy of personalized models and Fed-Avg across different data splits of CIFAR-10.

5.2.1. PERFORMANCE METRICS

We use the average accuracy (avg-acc) in conjunction with introduced metrics to make a more informed decision on evaluating personalized approaches instead of solely using avg-acc. Table 4 shows the performance metrics applied to the QoI of the different personalization methods across different data splits of CIFAR-10. We report a subset of the performance metrics (PUI, MPI, and API) since all the

Table 3: Per-user accuracy for the different personalized FL methods on the different data splits of the CIFAR-10 dataset.

Users	FedAvg			PersFL			FedPer			pFedMe			Per-FedAvg		
	DS-1	DS-2	DS-3	DS-1	DS-2	DS-3	DS-1	DS-2	DS-3	DS-1	DS-2	DS-3	DS-1	DS-2	DS-3
User 0	43.6	50.8	48.2	85.5	61.3	94.5	83.2	57.2	93.1	74.3	61.7	94.2	69.2	58.2	92.5
User 1	50.9	45.3	40.8	78.2	56.9	79.9	74.5	51.4	77	64.1	57.3	79.2	65	56.1	73.7
User 2	44.5	49.4	31.2	82.2	57.3	68.9	78.4	53.2	64.7	69.6	57.3	64.4	67.9	57.2	64.6
User 3	51.3	46.5	31.5	82.1	60.1	82.5	77.9	55.4	77.5	69.4	58.9	72.5	67.2	58.8	77
User 4	45.3	50.8	49.4	79.4	59.1	82.5	76.1	54.4	78.6	67.2	59.5	80	65.8	59.4	82.5
User 5	44.2	50.7	47.8	77.1	61.9	79.9	72.1	57.6	76.9	62.1	60.6	77.5	62.7	59.3	77.9
User 6	35.8	46.5	56.8	75.6	58.9	90.3	70.9	53.3	88.5	59.9	58.6	88.3	58.2	57.7	89.1
User 7	37.9	49.5	58.1	79.7	61.2	87.6	75.6	56.9	84.6	65.8	60.2	84	64.3	58	83.7
User 8	47.7	48.7	49	87.7	60	76.7	84.4	57.5	73.5	75.5	59.6	66.9	72.5	55.8	64.1
User 9	48.6	49.1	53.5	91	58.8	80.3	88.5	54.3	77.9	81.7	58.3	73.8	76.6	55.3	72.7
Avg. Acc.	45	48.7	46.6	81.9	59.6	82.3	78.2	55.1	79.2	69	59.2	78.1	66.9	57.6	77.8
Std Dev	5.1	2	9.4	4.9	1.7	7.2	5.6	2.1	7.9	6.7	1.4	9.2	5.1	1.5	9.5

Table 4: The performance metrics applied to the QoI of different personalized FL methods on CIFAR-10 dataset.

Metrics	PersFL			FedPer			pFedMe			PerFed		
	DS-1	DS-2	DS-3	DS-1	DS-2	DS-3	DS-1	DS-2	DS-3	DS-1	DS-2	DS-3
PUI	100	100	100	100	100	100	100	100	100	100	100	100
MPI	38.74	11.23	33.29	34.53	6.59	30.43	24.58	10.81	31.06	22.9	8.55	32.65
API	36.87	10.83	35.67	33.18	6.41	32.59	23.98	10.47	31.44	21.95	8.85	31.17
AA†	81.85	59.56	82.29	78.16	55.13	79.21	68.95	59.19	78.07	66.93	57.57	77.79

AA† is the average accuracy across all users.

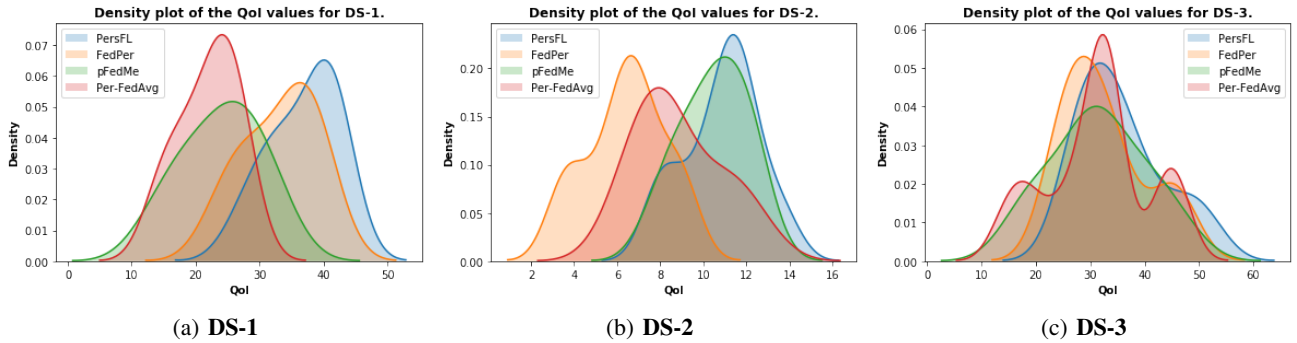(a) **DS-1** (b) **DS-2** (c) **DS-3**

Figure 1: Density plots of QoI across all data splits of CIFAR-10 for each personalized FL method.

personalization methods that we have surveyed lead to an increase in the personalized per-user accuracy in our experiments. This means that the PUI for each personalization method yields 100% with none of the users experiencing a decrease over their local and global models.

In terms of MPI and API on **DS-1**, PersFLperforms the best at 38.74% and 36.87%. On **DS-2**, PersFLand pFedMe yield the highest MPI at 11.23% and 10.81%. Similarly, these methods lead to an API of 10.83% and 10.47%. We observe that on **DS-3**, PersFLhas the highest MPI and API, at 33.29% and 35.67%. In terms of avg-acc, PersFLachieves the highest accuracy 81.85% on **DS-1**.

On **DS-2**, PersFLand pFedMe are the best performing at 59.56%, and 59.19%. Lastly, on **DS-3**, PersFLis the top-performing method at 82.29%.

Figure 1 shows the density plots of the QoI for all users across data splits of the CIFAR-10 dataset. The peaks in a density plot show the values concentrated over an interval where the x-axes of the plots show the QoI intervals and the y-axes show the density. We observe on **DS-1** that the most values for PersFLare concentrated around a higher value in the range of QoI values. On **DS-2**, the peak of PersFLis associated with a higher QoI interval compared to pFedMe. However, the distribution of pFedMe is more nor-
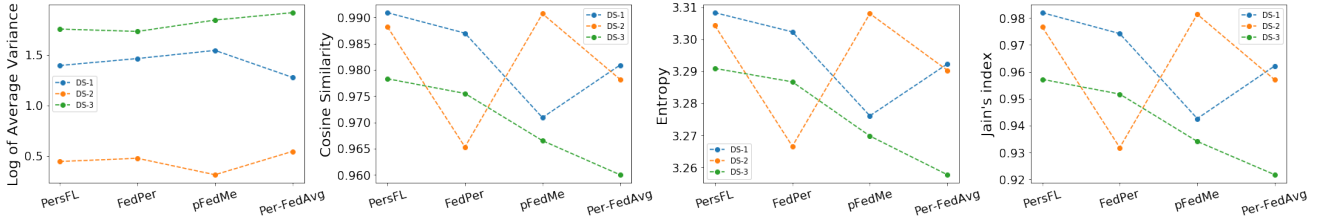
Figure 2: Fairness metrics applied to the personalization methods across all data splits of the CIFAR-10 dataset.

mal compared to PersFLas PersFLhas an additional peak corresponding to the QoI interval of 8. On **DS-3**, the QoI interval corresponding to the peaks of PersFL, pFedMe, and Per-FedAvg gives almost the same QoI value, where their QoI is concentrated.

5.2.2. FAIRNESS METRICS

We apply fairness metrics across all the data splits of CIFAR-10, as shown in Figure 2. The x-axis represents the different personalization methods, and the y-axis is the different fairness metrics (AV, CS, Entropy, and JI) applied to these methods. We observe in Figure 2 that the trends in metrics are similar across the different data splits. The main reason is that QoI distribution of personalized methods for the CIFAR-10 dataset is identical to each other. However, different trends can be observed with the application of methods to different datasets. We note that for a method to be fair, it also needs to give a reasonable performance in terms of the per-user personalization accuracy.

In Figure 2, Per-FedAvg gives the lowest AV amongst all methods. A lower AV means better per-user personalization accuracy distribution as the method yields more uniform accuracy than the other algorithms. For other three metrics (CS, Entropy, and JI), PersFLperforms the best. The higher the value of these metrics, the fairer QoI distribution. Therefore, on **DS-1**, PersFLis the fairest amongst the evaluated personalization methods.

On **DS-2**, pFedMe has the least AV, and the highest value for the other three metrics. This confirms that it generalizes per-user personalized accuracy across all the users. On **DS-2** in Table 4, the per-user personalization accuracy of pFedMe is similar to that of PersFLfrom a performance perspective. At the same time, pFedMe is relatively fairer than PersFL, therefore pFedMe is the fairest algorithm on **DS-2**.

Lastly, on **DS-3**, from Figure 2, FedPer has the lowest AV. However, it is not the fairest algorithm according to the other fairness metrics. Among the other three metrics, PersFLgives the highest values.

6. Conclusions

We introduce and adapted new metrics for performance and fairness, complementing the widely reported average personalized model accuracy to evaluate the personalization methods. We employed these metrics on four recent personalized FL methods across three different data splits on the CIFAR-10 dataset. Our evaluation results show that the personalized model that gives the highest average accuracy across users is not necessarily the fairest.

References

Arivazhagan, M. G., Aggarwal, V., Singh, A. K., and Choudhary, S. Federated learning with personalization layers. *CoRR*, abs/1912.00818, 2019. URL http://arxiv.org/abs/1912.00818.

Chen, F., Dong, Z., Li, Z., and He, X. Federated meta-learning for recommendation. *CoRR*, abs/1802.07876, 2018. URL http://arxiv.org/abs/1802.07876.

Deng, Y., Mahdi Kamani, M., and Mahdavi, M. Adaptive Personalized Federated Learning. *arXiv e-prints*, art. arXiv:2003.13461, March 2020.

Dinh, C. T., Tran, N. H., and Dung Nguyen, T. Personalized Federated Learning with Moreau Envelopes. *arXiv e-prints*, art. arXiv:2006.08848, June 2020.

Divi, S., Farrukh, H., and Celik, B. Unifying distillation with personalization in federated learning, 2021.

Fallah, A., Mokhtari, A., and Ozdaglar, A. Personalized Federated Learning: A Meta-Learning Approach. *arXiv e-prints*, art. arXiv:2002.07948, February 2020.

Finn, C., Abbeel, P., and Levine, S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *arXiv e-prints*, art. arXiv:1703.03400, March 2017.

Frankle, J. and Carbin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJl-b3RcF7.

Hard, A., Rao, K., Mathews, R., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., and Ramage, D. Federated learning for mobile keyboard prediction. *CoRR*, abs/1811.03604, 2018. URL http://arxiv.org/abs/1811.03604.

Hsu, T.-M. H., Qi, H., and Brown, M. Measuring the Effects of Non-Identical Data Distribution for Federated Visual Classification. *arXiv e-prints*, art. arXiv:1909.06335, September 2019.

Huang, Y., Chu, L., Zhou, Z., Wang, L., Liu, J., Pei, J., and Zhang, Y. Personalized cross-silo federated learning on non-iid data, 2021.

Jain, R., Chiu, D., and Hawe, W. A quantitative measure of fairness and discrimination for resource allocation in shared computer systems. *CoRR*, cs.NI/9809099, 1998. URL https://arxiv.org/abs/cs/9809099.

Jiang, Y., Konecný, J., Rush, K., and Kannan, S. Improving federated learning personalization via model agnostic meta learning. *CoRR*, abs/1909.12488, 2019. URL http://arxiv.org/abs/1909.12488.

Li, A., Sun, J., Wang, B., Duan, L., Li, S., Chen, Y., and Li, H. LotteryFL: Personalized and Communication-Efficient Federated Learning with Lottery Ticket Hypothesis on Non-IID Datasets. *arXiv e-prints*, art. arXiv:2008.03371, August 2020.

Li, T., Sanjabi, M., and Smith, V. Fair resource allocation in federated learning. *CoRR*, abs/1905.10497, 2019. URL http://arxiv.org/abs/1905.10497.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, May 2020. ISSN 1558-0792. doi: 10.1109/msp.2020.2975749. URL http://dx.doi.org/10.1109/MSP.2020.2975749.

Liang, P. P., Liu, T., Liu, Z., Salakhutdinov, R., and Morency, L. Think locally, act globally: Federated learning with local and global representations. *CoRR*, abs/2001.01523, 2020. URL http://arxiv.org/abs/2001.01523.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. *CoRR*, abs/0902.3430, 2009. URL http://arxiv.org/abs/0902.3430.

Mansour, Y., Mohri, M., Ro, J., and Theertha Suresh, A. Three Approaches for Personalization with Applications to Federated Learning. *arXiv e-prints*, art. arXiv:2002.10619, February 2020.

McMahan, H. B., Moore, E., Ramage, D., and y Arcas, B. A. Federated learning of deep networks using model averaging. *CoRR*, abs/1602.05629, 2016. URL http://arxiv.org/abs/1602.05629.

Moreau, J.-J. Propriétés des applications 'prox'. *Compte Rendus Acad. Sci.*, (256):1069–1071, 1963.

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng.*, 22(10):1345–1359, October 2010. ISSN 1041-4347. doi: 10.1109/TKDE.2009.191. URL https://doi.org/10.1109/TKDE.2009.191.

Shen, T., Zhang, J., Jia, X., Zhang, F., Huang, G., Zhou, P., Kuang, K., Wu, F., and Wu, C. Federated mutual learning, 2020.

Smith, V., Chiang, C.-K., Sanjabi, M., and Talwalkar, A. S. Federated multi-task learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30, pp. 4424–4434. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/6211080fa89981f66b1a0c9d55c61d0f-Paper.pdf.

Yu, T., Bagdasaryan, E., and Shmatikov, V. Salvaging federated learning by local adaptation. *CoRR*, abs/2002.04758, 2020. URL https://arxiv.org/abs/2002.04758.