F3: Fair and Federated Face Attribute Classification with Heterogeneous Data

Paper ID 298

Abstract. Fairness across different demographic groups is an essential criterion for face-related tasks, Face Attribute Classification (FAC) being a prominent example. Simultaneously, federated Learning (FL) is gaining traction as a scalable paradigm for distributed training. In FL, client models trained on private datasets get aggregated by a central aggregator. Existing FL approaches require data homogeneity to ensure fairness. However, this assumption is restrictive in real-world settings. E.g., geographically distant or closely associated clients may have heterogeneous data. In this paper, we observe that existing techniques for ensuring fairness are not viable for FL with data heterogeneity. We introduce F3, an FL framework for fair FAC under data heterogeneity. We propose two methodologies in F3, (i) Heuristic-based and (ii) Gradient-based, to improve fairness across demographic groups without requiring data homogeneity assumption. We demonstrate the efficacy of our approaches through empirically observed fairness measures and accuracy guarantees on popular face datasets. Using Mahalanobis distance, we show that F3 obtains a practical balance between accuracy and fairness for FAC.

Keywords: Fairness · Federated Learning · Data Heterogeneity

1 Introduction

Face Attribute Classification (FAC) finds prominence for tasks such as gender classification, face verification, and face identification [28]. Recently, researchers have highlighted a critical issue in FAC: attribute prediction may be biased towards specific demographic groups [17]. E.g., for gender classification, the error rate for 'darker' faces is greater than that on 'lighter' faces. Further, face recognition-based criminal detection systems are prone to classify innocent people with 'darker' faces as suspects. This bias in predictions is *unfairness*. It is often associated with the unavailability of balanced datasets [18]. To overcome this issue, researchers have introduced balanced, large-scale datasets [9].

Federated Learning (FL) has emerged as a popular paradigm for scalable distributed training for large-scale data [11]. FL comprises (i) independent clients that train local models on their private data and (ii) a central aggregator which combines these local models (e.g., through a random weighted average aka `FedAvg` [13]), to derive a generalised global model. Unfortunately, traditional FL models typically focus on standard performance measures (e.g., accuracy) and inherit the fairness related drawbacks of non-FL approaches [7].

To address the unfairness in FL several methods exist [5, 7, 14, 23]. However, these methods inherently assume FL clients with homogeneous data, i.e, they assume that FL clients' data contains samples from all the demographic groups of a particular *sensitive attribute*. For example, with 'age' as the sensitive attribute, the client's local training data would have samples from both 'young' and 'adult' demographic groups. However, clients' data is likely to be heterogeneous in many FL settings. A smartphone belonging to a 'young' user may have content belonging majorly to its peers [16], i.e., inter-client heterogeneity in terms of age. Similarly, geographically separated clients may exhibit inter-client heterogeneity in demographics. Such data heterogeneity may, in turn, reduce fairness for tasks such as FAC.

Our Approach: In this paper, we introduce and study the fair Face Attribute Classification (FAC) problem in FL under data heterogeneity (FL with DH). We first prove that existing approaches to ensure fairness [15, 8] are not applicable in this setting (Proposition 1). Consequently, we introduce *F3*, a FL framework for Fair Face Attribute Classification. Under the F3 framework, we propose two different methodologies (i) *Heuristic-based F3* and (ii) *Gradient-based F3*.

- Heuristic-based F3 includes novel aggregation heuristics: (i) FairBest, (ii) α-FairAvg, and (iii) α-FairAccAvg which prioritize specific local client model(s) to improve the accuracy and fairness trade-off (Sec. 3.1).
- Gradient-based F3 introduces FairGrad, where the client training is modified to include fairness through gradients communicated by the aggregator, to train a fair and accurate global model (Sec. 3.2).
- To validate the efficacy of both our methodologies, we conduct extensive experiments on three popular face datasets, namely FairFace [9], FFHQ [10], and UTK [25] (Sec. 4). Our results highlight that F3, through its aggregation and gradient-based methodologies, outperforms the standard approach, `FedAvg-DH` [13]. More concretely, F3 ensures 25%-82% improvement in terms of fairness with accuracy drop of 0.4%-17% compared to `FedAvg-DH`.

Related Work

We discuss our work w.r.t the existing literature in terms of (i) fairness and (ii) data heterogeneity in FL (see Appendix A in the supplementary for a detailed comparison). Concerning fairness, the de facto aggregation heuristic, `FedAvg` [13], is empirically shown to be more biased than a centrally trained model [21]. To overcome this, several approaches for ensuring fairness in FL exist [5, 7, 12, 14, 23]); with *Lagrangian Multiplier Method* (LMM) [12] being the most prominent. Typically, these approaches add a fairness component to training at the client level. However, we observe that the fairness component in LMM, or other approaches, is not defined for the data heterogeneity (DH) setting. As such, we require different approaches for fairness in FL with DH.

FL approaches typically consider data heterogeneity w.r.t. target labels where each client has access only to a specific set of target labels, i.e., non-i.i.d. data [19, 27]. In contrast, we consider data heterogeneity w.r.t. demographic groups.

2 Preliminaries

We consider the Face Attribute Classification (FAC) task, where $\mathcal{X}$ is the universal set of face images, with binary labels from $\mathcal{Y} = \{0, 1\}$ (e.g., male or female), and sensitive attribute $A \in \mathcal{A}$. Here, A can be age, race, or gender. The sensitive attribute takes a finite set of values, $A = \{a_1, \ldots, a_s\}$. E.g., age can take values such as 'young' or 'adult'. We next describe our FL setting for federated FAC.

2.1 Federated Learning (FL) Setting

In FL, the data is distributed across multiple parties referred to as clients. Let $\mathcal{C} = \{C_1, \ldots, C_m\}$ represent the set of clients; each C_i owns the set $D_i \subset \mathcal{X} \times \mathcal{Y} \times A$ containing n_i samples. Each C_i trains its local model $h_{\theta_{i,t}} : \mathcal{X} \to \mathcal{Y}$ parameterized by $\theta_{i,t}$ at round t. The aggregator holds small amount of data not enough to train a good model. Generally, the data is split into a test set and a validation set (D_a) [24]. We assume that D_a comprises samples from each demographic group.

At each round t, a random subset of clients $S_t \subseteq \mathcal{C}$ communicate their locally updated model parameters $\Theta_t = \{\theta_{i,t} \mid C_i \in S_t\}$ to the aggregator. The aggregator combines all communicated model parameters to obtain the global parameters at round t, ϕ_t, using a heuristic choice function $\mu : \Theta_t \to \phi_t$. The Weighted average (or `FedAvg`) [13] is the most used heuristic, defined as: $\mu_{\texttt{FedAvg}}(\Theta_t) \triangleq \phi_t = \sum_{i \in S_t} \frac{n_i}{\sum_j n_j} \theta_{i,t}$. The aggregator then communicates the model parameters back to the clients. Then clients initialise their local model with these parameters and train further. This back and forth process is repeated multiple times till convergence.

2.2 Fairness Notions

The standard notions for fair classification depend on the error rates: False negative rate (FNR) and False positive rate (FPR). For a face attribute classifier h, given a face image x with true label y and sensitive attribute $a \in A$, we have $FNR = \Pr(h(x) \neq y | y = 1)$ and $FNR_a = \Pr(h(x) \neq y | A = a, y = 1), \forall a \in A$. Likewise, $FPR = \Pr(h(x) \neq y | y = 0)$, and $FPR_a = \Pr(h(x) \neq y | A = a, y = 0), \forall a \in A$. FNR_a and FPR_a are the error rates observed on the data samples belonging to a particular demographic group with sensitive attribute $a \in A$. E.g., consider an FAC task for 'gender' classification with 'age' as the sensitive attribute. The attribute comprises {'young', 'adult'} as the demographic groups. Now, consider the following group-fairness notions.

Equality of Opportunity (EOpp) [3]: A classifier h satisfies EOpp for a distribution over $(\mathcal{X}, \mathcal{Y}, A)$ if: $FNR_a = FNR, \forall a$. We denote the violation in EOpp as $\Delta_{EOpp} = \max(\{FNR_a - FNR | \forall a \in A\})$. That is, Δ_{EOpp} is the maximum disparity in FNR across the groups. Intuitively, EOpp ensures that the probability of predicting a 'male' face as 'female' is the same across age groups.

Equalized Odds (EO) [6]: A classifier h satisfies EO over $(\mathcal{X}, \mathcal{Y}, A)$ if: $FNR_a = FNR$ and $FPR_a = FPR\ \forall a$. Let $\Delta_{EO} = \max(\max(\{FPR_a - FPR | \forall a \in A\}), \max(\{FNR_a - FNR | \forall a \in A\}))$ denote the violation in EO. EO states that the probability of mis-predicting the gender must be independent of age.

Accuracy Parity (AP) [26]: A classifier h satisfies AP for a distribution over $(\mathcal{X}, \mathcal{Y}, A)$ if: $FPR + FNR = FPR_a + FNR_a, \forall a$. Let $\Delta_{AP} = \max(\{FPR_a - FPR | \forall a \in A\}) + \max(\{FNR_a - FNR | \forall a \in A\})$ denote violation in AP. AP ensures that the overall classification error is equal across the age groups.

Lagrangian Multiplier Method (LMM) [12]: To incorporate these fairness notions in FAC, the standard technique is to train a model that maximises accuracy while minimising the violation in these fairness notions. LMM adopts a loss function that simultaneously incorporates cross-entropy loss l_{CE} and the violation in fairness constraint $(\Delta_{EOpp}, \Delta_{EO}, \Delta_{AP})$, weighted by the lagrangian multiplier $\lambda \in \mathbb{R}^+$. Formally, in LMM, the loss function $L_{LMM}(h(X), Y, A)$ for a classifier h, for $k \in \{EOpp, EO, AP\}$ and $(X, Y) \subseteq \mathcal{X} \times \mathcal{Y}$, is as follows.

$$L_{LMM}(\cdot) = \mathbb{E}_{(x,y)\sim(X,Y)}[l_{CE}(h(x), y)] + \lambda \Delta_k. \tag{1}$$

3 Methodology

We first motivate the problem by showing that existing fair FL approaches are not applicable in the data heterogenous setting. We then introduce our novel methodologies, (i) Heuristic-based F3 and (ii) Gradient-based F3.

Motivation: In a practical FL setting, each client might only possess samples from an individual demographic group. E.g., samples belonging only to the 'young' age group when age is the sensitive attribute. We refer to this scenario as Federated Learning with Data Heterogeneity (FL with DH). The existing approaches for fair FL typically compute fairness violation locally [12, 15]. Proposition 1 shows that with DH, this fairness violation component for the demographic groups not present in a particular client's data is not defined.

Proposition 1. *In the Lagrangian Multiplier Method, the loss L_{LMM} (Eq. 1) is not defined for FL with Data Heterogeneity (FL with DH).*

The formal proof is available in the accompanying supplementary (Appendix B). Proposition 1 holds for any fairness violation function (such as Δ_{EO}, Δ_{AP}) that requires samples belonging to all the demographic groups. E.g., the loss functions defined in [1, 20, 22]. Thus, we cannot use these functions to train for fairness in FL with DH. Additionally, training only for accuracy compromises fairness [4], implying that standard approaches such as `FedAvg` may not suffice. As a result, we next propose novel methodologies curated for FL with DH.

3.1 Heuristic-based F3

Observe that as `FedAvg` aggregates a random (sub)set of models at each round, it fails to ensure fairness as the local models for aggregation may potentially

Algorithm 1 Heuristic-based F3

Input: (1) Each client $C_k \in \mathcal{C}$ with D_k (2) Hyperparameters: maximum number of communication rounds T, number of local epochs E, learning rate η, accuracy threshold a, epoch threshold τ (3) A heuristic choice function $\mu_\varsigma(\Theta_t)$ s.t. $\varsigma = \{\texttt{FedAvg}, \text{FairBest}, \alpha\text{-FairAvg}, \alpha\text{-FairAccAvg}\}$
Output: Model ϕ

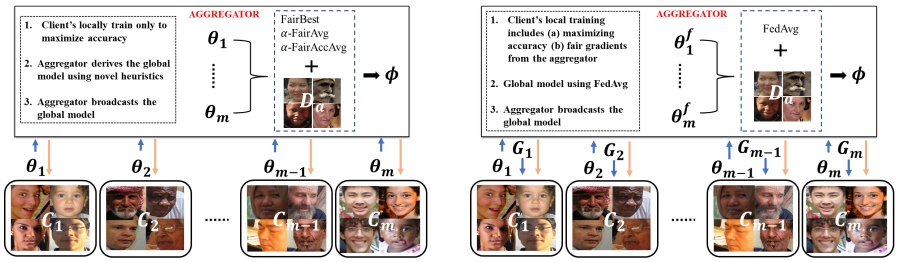
$\phi_0 \leftarrow$ randomly initialized weights
for each round $t = 0, 1 \ldots, T-1$ **do**
  for each client $k \in S_t$ (in parallel) **do**
    (Client k) $\theta_{k,t} \leftarrow \text{LocalTraining}(k, \phi_t)$
  end for
  (Aggregator) $\phi_{t+1} \leftarrow \mu_\varsigma(\Theta_t)$; $\Theta_t = \{\theta_{k,t} \mid k \in S_t\}$
  $\phi_{best} \leftarrow \mathsf{StoppingCondition}(t+1, \phi_{t+1}, \phi_{best}, a, \tau)$
end for
return ϕ_{best}
procedure LocalTraining(k, ϕ_t)
  $\theta_{k,t} \leftarrow \phi_t$
  for each local epoch $i = 1, 2, \ldots, E$ **do**
    (Per Batch) $\theta_{k,t} \leftarrow \theta_{k,t} - \eta \cdot \nabla_{\theta_{k,t}} L_k \left(h_{\theta_{k,t}}(\cdot), D_k\right)$
  end for
  return $\theta_{k,t}$
end procedure

be biased. In turn, they may amplify the unfairness of the global model. In Heuristic-based F3, we propose novel aggregation heuristics that prioritize the local client models, which perform desirably in terms of fairness. The method comprises the following steps.

1. Local Training. Each client C_i trains its model h_{θ_i} only for maximising accuracy, i.e., minimising $L_i(h_{\theta_i}, D_i) = \mathbb{E}_{(x,y)\sim D_i}[l_{CE}(h_{\theta_i}(x), y)]$. At each round t, a random subset of clients $S_t \subseteq \mathcal{C}$ communicate their model parameters to the aggregator.
2. Model Aggregation. To better control the accuracy and fairness trade-off, we propose novel heuristics for aggregation. These heuristics derive the global model based on accuracy and fairness values for the models in S_t computed on the aggregator's set D_a.
3. Model Communication. The aggregator then communicates the global model parameters to each client. The clients adopt these parameters and further train on them to maximise accuracy.

Fig. 1a depicts Heuristic-based F3 and Algorithm 1 provides a procedural outline.

Stopping Criteria. In Algorithm 1, StoppingCondition controls the training's stoppage and model updates (see Appendix C). The procedure records the improvement in accuracy and fairness values across epochs. It updates the "best" model observed so far if: (i) the current epoch is less than a threshold τ, or (ii) ϕ produces a lesser fairness violation than the previous best model. The training stops if the change in accuracy does not exceed a threshold "a" across τ rounds.



(a) Heuristic-based F3 (b) Gradient-based F3

Fig. 1: Overview of our Methodologies.

Heuristics for Fair FL: With our novel heuristics, we aim to ensure fairness in FL with DH by deliberately aggregating only the subset of local models that perform desirably w.r.t. to fairness and accuracy. The aggregator quantifies the performance of local client models based on their empirical fairness violation and accuracy computed on the aggregator set D_a. Specifically, let $\Delta_{loss}(h_i(\theta_{i,t}))$ denote any fairness violation for client's $C_i \in S_t$ model on D_a at any round t. Denote $Acc(h_i(\theta_{i,t}))$ as the accuracy of C_i's model on D_a at t. With this, consider the following novel heuristics that aim to strike a practical balance between fairness and accuracy.

1. FairBest: Aggregator selects a specific model among local models that provides the least fairness violation on D_a. Formally, the global model parameters ϕ_t at round t are: $\mu_{\text{FairBest}}(\Theta_t) \triangleq \phi_t = \theta_{i^*,t}$ s.t. $i^* = \arg\min_i \{\Delta_{loss}(h_i(\theta_{i,t}))\}$.
2. α-FairAvg: This heuristic generalizes FairBest by selecting the top $\alpha\%$ of local models followed by their weighted average. Formally, consider the set F_t, at a round t, which comprises the top-$\alpha\%$ of clients in *increasing* order of the ratio $\Delta_{loss}(h_i(\theta_{i,t}))$. Now,

$$\mu_{\alpha\text{-FairAvg}}(\Theta_t) \triangleq \phi_t = \sum_{i \in F_t} \frac{n_i}{\sum_{j \in F_t} n_j} \theta_{i,t} \tag{2}$$

3. α-FairAccAvg: Aggregator selects the top-$\alpha\%$ of local model parameters that give the best ratio of accuracy with fairness violation on D_a and take their weighted average. Consider the set F_t, at a round t, comprising the top-$\alpha\%$ of clients in *decreasing* order of the ratio $\frac{Acc(h_i(\theta_{i,t}))}{\Delta_{loss}(h_i(\theta_{i,t}))}$.

$$\mu_{\alpha\text{-FairAccAvg}}(\Theta_t) \triangleq \phi_t = \sum_{i \in F_t} \frac{n_i}{\sum_{j \in F_t} n_j} \theta_{i,t} \tag{3}$$

As α increases, more local models get aggregated akin to `FedAvg` with heterogeneous data. That is, with an increase in α, Eq. 2 and Eq. 3 tend to $\sum_{i \in \mathcal{C}} \frac{n_i}{\sum_j n_j} \theta_{i,t}$ (the standard `FedAvg` aggregation), such that α-FairAvg and α-FairAccAvg tend to mimic `FedAvg`. We also show this behavior empirically in the supplement (Appendix D).

Algorithm 2 FairGrad: Gradient-based F3

Input: (1) Each client $C_k \in \mathcal{C}$ with D_k (2) Hyperparameters: maximum number of communication rounds T, number of local epochs E, learning rate η, accuracy threshold a, epoch threshold τ and $\beta \in [0, 1]$.
Output: Model ϕ

$\phi_0 \leftarrow$ randomly initialized weights
for each round $t = 0, 1 \ldots, T-1$ **do**
  for each client $k \in S_t$ (in parallel) **do**
    $\theta_{k,t} \leftarrow \phi_t$
    for each local epoch $i = 1, \ldots, E$ **do**
      (Aggregator) $G_{k,t} \leftarrow \nabla_{\theta_{k,t}} \Delta_{EO} \left(h_{\theta_{k,t}}(\cdot), D_a\right)$
      For a fixed $G_{k,t}$, the client updates for all batches,
      (Client k) $\begin{cases} g_{k,t} & \leftarrow \nabla_{\theta_{k,t}} L_k \left(h_{\theta_{k,t}}(\cdot), B\right); \; g^*_{k,t} \leftarrow \beta \cdot g_{k,t} + (1-\beta) \cdot G_{k,t} \\ \theta_{k,t} & \leftarrow \theta_{k,t} - \eta g^*_{k,t} \end{cases}$
    end for
  end for
  (Aggregator) $\phi_{t+1} \leftarrow \mu_{\texttt{FedAvg}}(\Theta_t)$, $\Theta_t = \{\theta_{k,t} \mid k \in S_t\}$
  $\phi_{best} \leftarrow \mathsf{StoppingCondition}(t+1, \phi_{t+1}, \phi_{best}, a, \tau)$
end for
return ϕ_{best}

3.2 FairGrad: A Gradient-based F3

In Heuristic-based F3, it is only possible to explore a limited set of models that provide different trade-offs between accuracy and fairness. While each client trains to maximize accuracy, the client models may diverge from the "fair" aggregated model. Hence aggregation of these individual client models may not always provide a good trade-off.

Based on these observations, and motivated from [2], we now propose a Gradient-based approach curated for the FL with DH setting, namely FairGrad. Informally, in FairGrad we train the individual client models for accuracy and w.r.t. fair gradients obtained from the aggregator. As a result, the local client models comprise fairness information and subsequent aggregation through `FedAvg` provides a balance between accuracy and fairness. Note that clients communicate with the aggregator even during their local training. Formally, FairGrad comprises the following steps.

1. Local Training. At every epoch, we compute the following gradients.
   (i) Client Level: For each client C_i, we compute the gradients $g_i(\theta_i, D_i)$ w.r.t. client weights θ_i and local dataset D_i for maximizing accuracy. That is, minimizing the cross-entropy loss l_{CE}, $L_i(h_{\theta_i}, D_i) = \mathbb{E}_{(x,y)\sim D_i}[l_{CE}(h_{\theta_i}(x), y)]$.
   (ii) Aggregator Level: For each client C_i, the aggregator computes the gradients $G_i(\theta_i, D_a)$ w.r.t. client weights θ_i and aggregator dataset D_a for minimizing fairness, i.e. Δ_k, for $k \in \{EOpp, EO, AP\}$. Aggregator communicates G_i to each C_i.

Upon receiving G_i, each client C_i must now judiciously aggregate the two gradients. To this end, we fix G_i for an entire epoch while updating g_i for each batch-wise training. Further, C_i performs weighted aggregation with weight $\beta \in (0, 1)$ to determine the aggregated gradients g_i^* as follows: $g_i^* = \beta g_i + (1 - \beta)G_i$. Next, C_i performs the SGD update: $\theta_i \leftarrow \theta_i - \eta g_i^*$.

2. Model Aggregation. After a few epochs of local training, a random subset of clients send their models to the aggregator, who then aggregates them using `FedAvg` to derive the global model ϕ.
3. Model Communication. The aggregator communicates the aggregated global model parameters to each client. The clients adopt these parameters to perform further local training.

Fig. 1b depicts Heuristic-based F3 and Algorithm 2 provides a procedural outline. Given our novel heuristics and FairGrad, we next conduct experiments to compare their empirical performance and highlight the efficacy of F3 for FL under data heterogeneity.

4 Experiments

We conduct our experiments on the following face datasets: FairFace [9], FFHQ [10], and UTK [25]. In this section, we first define our baseline `FedAvg-DH` for an appropriate comparison. Then, we provide our FL setup, training details, and network model. Finally, we present our results and the key takeaways.

Baseline: To compare our methods' performance, we create the baseline `FedAvg-DH`. `FedAvg-DH` is simply `FedAvg` [13] for our FL setting with Data Heterogeneity.

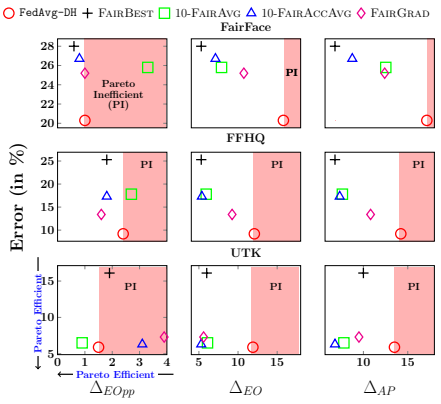
FL Setup: We consider 50 clients, i.e., $\mathcal{C} = \{C_1, \ldots, C_{50}\}$. We randomly distribute the training data so that each client has samples of only a particular demographic group to ensure data heterogeneity. Each client locally trains its model on its private data. The global model aggregation is performed periodically till convergence. At each aggregation round t, we let $S_t = \mathcal{C}$. The training details specific to each dataset follow next.

Training Details: We focus on popular face datasets: FairFace [9], FFHQ [10], and UTK [25]. For each of these, we consider 'age' as the sensitive attribute and 'gender' as the predicting label. Further, we divide the samples into two age groups, ≤ 30 and > 30 years. We distribute the data among the clients such that 50% of the clients have access to data samples belonging to the age group ≤ 30 and others have access to samples belonging to age group > 30. Each client's local data comprises $\approx$ 1K training samples for all three datasets.

For FairFace [9] and FFHQ [10], we use a batch size of 256 and train the models for $T = 50$ communication rounds with clients training their local models for $E = 4$ epochs (per round). We use learning rates of $\eta = 0.05$ and $\eta = 0.01$ for FairFace and FFHQ, respectively. We set the accuracy tolerance at $a = 1\%$ and threshold round at $\tau = 20$. For UTK [25], we train using the learning rate $\eta = 0.01$ and batch size 64. We also set $T = 80$, $E = 2$, $a = 1\%$ and $\tau = 20$.

Heuristic/ Method	Accuracy (↑)	Reduction in Fairness Violation (↓)		
		Δ_{EOpp}	Δ_{EO}	Δ_{AP}
FairFace [9]				
`FedAvg-DH`	79.7% ± 0.8	1.0 ± 1.4	16.3 ± 3.3	17.3 ± 2.5
FairBest	72.0% ± 0.8	0.6 ± 0.6	8.2 ± 1.1	8.8 ± 1.1
10-*FairAvg*	74.2% ± 1.1	3.3 ± 1.2	10.2 ± 1.6	13.5 ± 2.5
10-*FairAccAvg*	73.3% ± 1.2	0.8 ± 0.7	9.6 ± 0.9	10.4 ± 1.6
FairGrad	74.4% ± 0.4	1.0 ± 0.2	9.2 ± 3.1	10.2 ± 3.2
FFHQ [10]				
`FedAvg-DH`	90.8% ± 0.5	2.4 ± 1.3	10.5 ± 1.3	13.0 ± 0.5
FairBest	74.7% ± 2.0	1.8 ± 0.4	1.9 ± 0.3	3.2 ± 0.8
10-*FairAvg*	82.2% ± 0.5	2.7 ± 0.3	2.7 ± 0.3	4.3 ± 0.6
10-*FairAccAvg*	82.7% ± 0.5	1.8 ± 1.0	2.0 ± 0.5	3.9 ± 1.1
FairGrad	86.6% ± 1.2	1.6 ± 0.3	6.9 ± 1.1	8.5 ± 1.6
UTK [25]				
`FedAvg-DH`	94.1% ± 0.5	1.5 ± 1.2	11.9 ± 1.6	13.5 ± 2.7
FairBest	83.9% ± 0.5	1.9 ± 4.6	6.0 ± 4.3	10.0 ± 6.9
10-*FairAvg*	93.5% ± 2.9	0.9 ± 4.3	6.1 ± 2.9	7.8 ± 4.1
10-*FairAccAvg*	93.7% ± 1.1	3.1 ± 3.1	5.3 ± 1.0	6.8 ± 2.1
FairGrad	92.7% ± 0.2	3.9 ± 0.7	5.6 ± 0.4	9.5 ± 1.0

(a)



(b)

Fig. 2: Accuracy and Fairness Violation (Δ_k, $\forall k \in \{EOpp, EO, AP\}$) values for the baseline `FedAvg-DH` [13] and our approaches. In Fig. 2a, the numbers highlighted in green represent the least value of Δ_k. The numbers highlighted in magenta provide the highest accuracy out of our proposed approaches. In Fig. 2b, the optimum point is bottom left, i.e., low %-Error and low Δ_k.

Model: We adopt PyTorch's implementation of the standard ResNet-18 architecture for the base model [12]. We run our experiments on 8 NVIDIA GeForce GTX 1080 with 10 GB RAM.

Method: We run every experiment 5 times and report the approaches' average and standard deviation. For each instance, we randomly generate an aggregator set D_a with samples between 10%-20% of the overall dataset size. For α-FairAvg and α-FairAccAvg, we choose $\alpha = 10$, i.e., 10% of the total local models.

We restate that fairness guarantees often come at the cost of accuracy [4]. However, our methodologies aim to strike an effective balance between fairness improvements and accuracy.

4.1 Results

First, we observe a maximum Coefficient of Variation (CoV) of 0.96 across all experiments. For 70% of our experiments, we observe a CoV < 0.2 indicating the stability of our approaches and the results presented. We next discuss accuracy and fairness violations of our approaches for the three datasets.

Fairness Improvements: Fig. 2a provides the accuracy and fairness violation (Δ_k for $k \in \{EOpp, EO, AP\}$) of our approaches compared to the baseline `FedAvg-DH`. Recall that lower Δ_k implies lesser fairness violation. In Fig. 2a, the numbers highlighted in green represent the least value of Δ_k (obtained often by one of our approaches) for each dataset. The numbers highlighted in magenta represent the highest accuracy out of our proposed approaches (i.e., excluding the

baseline `FedAvg-DH`). In general, our methodologies provide significant fairness improvements for a marginal drop in accuracy. Details follow.
FairFace [9]: With FairGrad, we observe fairness improvement up to 22% with an accuracy drop of only 4% compared to `FedAvg-DH`. 10-FairAccAvg also shows fairness improvements up to 42% with an accuracy drop of 7%. FairBest provides the least fairness violation for each of the three fairness notions.
FFHQ [10]: FairBest provides the least reduction in fairness violation: 82% and 75% reduction in Δ_{EO} and Δ_{AP} respectively, compared to `FedAvg-DH`. However, FairGrad provides a desirable trade-off with fairness improvement upto 20% with a marginal accuracy drop of 4% compared to `FedAvg-DH`.
UTK [25]: For UTK, 10-FairAvg, 10-FairAccAvg and FairGrad outperform `FedAvg-DH`. 10-FairAccAvg improves fairness by 40% (Δ_{EOpp}), 48% (Δ_{EO}) and 42% (Δ_{AP}) for an accuracy drop of only 0.6%. 10-FairAvg improves fairness by 55% in Δ_{EO} and 49% in Δ_{AP} for an accuracy drop of 0.4%. Similarly, FairGrad provides improvement in fairness by 53% in Δ_{EO} and 30% in Δ_{AP} with 1.7% accuracy drop.

Visualizing Accuracy and Fairness Trade-off: Fig. 2b depicts the accuracy and fairness trade-offs of our heuristics with the baseline `FedAvg-DH`. Note that the optimum point is bottom left, i.e., low %-Error and low fairness violation (Δ_k). The red circle marker for `FedAvg-DH` appears at the bottom right on most of the plots, showing low error (higher accuracy) at the cost of fairness. Markers in the bottom-left corner of the plot assure the least fairness violation while maintaining high accuracy. FairBest, 10-FairAvg and 10-FairAccAvg provide lower fairness violations, for a marginal decrease in accuracy. FairGrad also shows a better accuracy and fairness trade-off by obtaining accuracy values approximately equivalent to `FedAvg-DH` and fairness violations closer to our heuristics. The highlighted "Pareto Inefficient" region is the area which is *pareto dominated* by the baseline `FedAvg-DH`. Observe that, in general, our approaches lie outside the Pareto inefficient region.

Additional Experiments: In the supplement (Appendix F), we provide additional results, including (i) an ablation study for the hyper parameters α, a and τ, (ii) different network architecture, and (iii) experiments on accuracy and fairness trade-off for a different FAC task and sensitive attribute.

4.2 Discussion

We see that both Heuristic and Gradient-based F3 perform significantly better in fairness while maintaining competitive accuracy compared to `FedAvg-DH`.

Pareto-optimality: To quantify the improved accuracy and fairness trade-off using our approaches, we use Mahalanobis distance (MD). More concretely, we derive the distance between our approaches' performance, i.e., % Error and Δ_k, $k = \{EOpp, EO, AP\}$, from the origin in Fig. 2b. We compute the distance as $MD(\overline{x}) = \sqrt{(\overline{x} - \overline{\mu})^T S^{-1} (\overline{x} - \overline{\mu})}$. Here, $\overline{x} = (Error, \Delta_k)$ is the vector with observed % Error and Fairness Violation, $\overline{\mu}$ is a vector with mean values, and S the covariance matrix of $\overline{x}$.

An approach's lesser MD implies a better trade-off. We provide the specific distance values of our approaches and `FedAvg-DH` in the supplementary (Appendix E). Our results highlight that for FairFace, FairGrad provides the best accuracy and trade-off, i.e., least MD). For FFHQ and UTK, FairGrad and 10-FairAvg provide the least distance for Δ_{EOpp}, respectively. For Δ_{EO} and Δ_{AP}, 10-FairAccAvg outperforms others for both datasets. These distances quantitatively show that FairGrad and 10-FairAccAvg significantly improve the accuracy and fairness trade-off over `FedAvg-DH`.

Heuristic-based F3 vs. Gradient-based F3: Overall, FairBest provides the least fairness violation, and the baseline `FedAvg-DH` provides the highest accuracy. Ranking the approaches using MD shows that 10-FairAccAvg and FairGrad provide an improved accuracy and fairness trade-off than `FedAvg-DH`.

Of the two, FairGrad requires comparatively higher communication overhead between the clients and the aggregator. On the other hand, finding an optimal α for 10-FairAccAvg that obtains a desirable trade-off may also be challenging. As α increases, the increase in accuracy also increases the fairness violation (Appendix D). As a result, a practitioner can appropriately decide between the approaches to achieve the desired accuracy and fairness trade-off.

5 Conclusion

In this paper, we focus on Fair Attribute Classification (FAC) in FL setting with data heterogeneity. We observe that existing approaches to ensure fairness in FL do not work in a heterogeneous setting due to the unavailability of demographic-specific data samples across clients. To address this, we propose F3, a novel FL framework to achieve fairness in FAC. With F3, we introduce (i) Heuristic-based F3, which includes three aggregation heuristics that ensure fairness while simultaneously maximizing the model's accuracy (ii) Gradient-based F3 to ensure clients are trained for fairness and accuracy. Experimentally, our approaches outperform the default counterpart in FL on challenging benchmark face datasets. The results suggest that F3 helps strike a practical balance between fairness and accuracy for FAC.

References

1. Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., Wallach, H.: A reductions approach to fair classification. In: ICML. pp. 60–69 (2018)
2. Augenstein, S., Hard, A., Partridge, K., Mathews, R.: Jointly learning from decentralized (federated) and centralized data to mitigate distribution shift. arXiv preprint arXiv:2111.12150 (2021)
3. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data pp. 153–163 (2017)
4. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data pp. 153–163 (2017)
5. Ezzeldin, Y.H., Yan, S., He, C., Ferrara, E., Avestimehr, S.: Fairfed: Enabling group fairness in federated learning. arXiv preprint arXiv:2110.00857 (2021)

6. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: NeurIPS. vol. 29, pp. 3315–3323 (2016)
7. Hu, S., Wu, Z.S., Smith, V.: Provably fair federated learning via bounded group loss. arXiv preprint arXiv:2203.10190 (2022)
8. Jung, S., Chun, S., Moon, T.: Learning fair classifiers with partially annotated group labels. In: CVPR. pp. 10348–10357 (2022)
9. Karkkainen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: WACV. pp. 1548–1558 (2021)
10. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR. pp. 4401–4410 (2019)
11. Konečnỳ, J., McMahan, H.B., Ramage, D., Richtárik, P.: Federated optimization: Distributed machine learning for on-device intelligence. arXiv preprint arXiv:1610.02527 (2016)
12. Lokhande, V.S., Akash, A.K., Ravi, S.N., Singh, V.: Fairalm: Augmented lagrangian method for training fair models with little regret. In: ECCV. pp. 365–381 (2020)
13. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: AISTATS (2017)
14. Padala, M., Damle, S., Gujar, S.: Federated learning meets fairness and differential privacy. In: ICONIP. pp. 692–699 (2021)
15. Padala, M., Gujar, S.: Fnnc: Achieving fairness through neural networks. In: IJCAI. pp. 2277—2283 (2020)
16. Ruan, Y., Joe-Wong, C.: Fedsoft: Soft clustered federated learning with proximal local updating. arXiv preprint arXiv:2112.06053 (2021)
17. Terhörst, P., Kolf, J.N., Huber, M., Kirchbuchner, F., Damer, N., Moreno, A.M., Fierrez, J., Kuijper, A.: A comprehensive study on face recognition biases beyond demographics. IEEE Transactions on Technology and Society **3**(1), 16–30 (2021)
18. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR. pp. 1521–1528 (2011)
19. Wang, H., Kaplan, Z., Niu, D., Li, B.: Optimizing federated learning on non-iid data with reinforcement learning. In: IEEE INFOCOM. pp. 1698–1707 (2020)
20. Zafar, M.B., Valera, I., Rogriguez, M.G., Gummadi, K.P.: Fairness constraints: Mechanisms for fair classification. In: AISTATS. pp. 962–970 (2017)
21. Zeng, Y., Chen, H., Lee, K.: Improving fairness via federated learning. arXiv preprint arXiv:2110.15545 (2021)
22. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: AIES. pp. 335–340 (2018)
23. Zhang, D.Y., Kou, Z., Wang, D.: Fairfl: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models. In: IEEE Big Data. pp. 1051–1060 (2020)
24. Zhang, J., Wu, Y., Pan, R.: Incentive mechanism for horizontal federated learning based on reputation and reverse auction. In: WWW. pp. 947–956 (2021)
25. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: CVPR. pp. 5810–5818 (2017)
26. Zhao, H., Gordon, G.: Inherent tradeoffs in learning fair representations. In: NeurIPS. vol. 32, pp. 15675–15685 (2019)
27. Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., Chandra, V.: Federated learning with non-iid data. arXiv preprint arXiv:1806.00582 (2018)
28. Zheng, X., Guo, Y., Huang, H., Li, Y., He, R.: A survey of deep facial attribute analysis. International Journal of Computer Vision **128**(8), 2002–2034 (2020)