2020 IEEE International Conference on Big Data (Big Data) | 978-1-7281-6251-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/BigData50022.2020.9378043

FairFL: A Fair Federated Learning Approach to Reducing Demographic Bias in Privacy-Sensitive Classification Models

Daniel (Yue) Zhang, Ziyi Kou, Dong Wang
Department of Computer Science and Engineering
University of Notre Dame, Notre Dame, IN, USA
{yzhang40, zkou, dwang5}@nd.edu

Abstract—**The recent advance of the federated learning (FL) has brought new opportunities for privacy-aware distributed machine learning (ML) applications to train a powerful ML model without accessing the private training data of the participants. In this paper, we focus on addressing a novel *fair classification* problem in FL where the model trained by FL displays discriminatory bias towards particular demographic groups. Addressing the fairness issue in a FL framework posts three critical challenges: *fairness and performance trade-offs*, *restricted information*, and *constrained coordination*. To address these challenges, we develop FairFL, a fair federated learning framework dedicated to reducing the bias in privacy-sensitive ML applications. It consists of a principled deep multi-agent reinforcement learning framework and a secure information aggregation protocol that optimizes both the accuracy and the fairness of the learned model while respecting the strict privacy constraints of the clients. Evaluation results on real-world applications showed that FairFL can achieve significant performance gains in both fairness and accuracy of the learned model compared to state-of-the-art baselines.**

1. Introduction

The advance of machine learning (ML) has brought new advances in various frontiers such as smart city, medical AI, and autonomous driving [1], [2]. With the proliferation of ML, there exist a growing concern from civil organizations, governments, and researchers about the fairness of ML-driven applications and their potential algorithmic bias towards particular demographic groups [3], [4]. For example, a recent study from IBM has found that current commercial facial recognition services have a much higher error rate for images that involve dark-skinned women than that for light-skinned men [5]. To alleviate the algorithmic bias and discrimination in ML models, an increasing set of fair ML approaches have been developed. Examples include reducing biased training samples with adaptive sampling [6], pre-processing the training data to mitigate the impact of demographic information by transforming the data to a feature space [7], and enforcing fairness constraints with judiciously designed loss functions [8].

Existing fair ML solutions often assume the ML models are trained in a centralized manner with all the training data being directly accessible [9], [10]. Such an assumption prevents those solutions from being applied in a large set of privacy-sensitive distributed ML applications (e.g., smart health, location-based services, and personalized recommendations) where the training data is generated by multiple parties who do not have sufficient trust to share the training data with each other [1], [11]. To address this limitation, we develop a new decentralized and privacy-aware fair ML framework based on federated learning (FL), which enables multiple data owners (referred to as clients) to collaboratively learn a shared ML model without the need to share any private data to the model owner (i.e., the FL server). More specifically, our goal is to provide an accurate and fair FL based classification framework for different demographic groups (e.g., gender, age, and race) in privacy-sensitive ML applications. We refer to this problem as *fair classification in federated learning*. Consider a real-world example of a skin cancer detection application where a classifier is trained to predict the skin cancer risk of a patient based on the image of the skin and the patient's demographic information. Our goal here is to ensure the trained classifier running on the patient's device is both fair and accurate without asking patients to send their private data to a centralized server for model training. Our problem is challenging to solve due to a few technical challenges, which we elaborate below.

Fairness and performance trade-offs across multiple demographic groups: the first challenge lies in balancing the fairness and performance of the classification model learned by FL. It is well observed that reducing the bias (i.e., improving fairness) of the ML models often lead to the model performance degradation [12], [13]. For example, a random binary classifier has perfect fairness on a balanced dataset with a poor accuracy of the classification results (i.e., 50% accuracy for both classes). The fairness and performance trade-offs become much more complicated if the fairness of multiple demographic groups are considered simultaneously. For example, we found improving fairness of one demographic group may lead to the degradation of the fairness and accuracy of another. Consider a simple example where we have much more training data samples from males than females. To achieve fairness for the "female" demographic

978-1-7281-6251-5/20/$31.00 ©2020 IEEE

group, the model should be trained on a more balanced data by reducing the number of male data samples in the training process. However, if it happens that the male data samples are mostly African Americans, the data samples for them will decrease as well, leading to lower accuracy and potentially higher bias for the "African American" demographic group. It remains to be an open challenge on how to achieve a good balance between fairness and accuracy among different demographic groups.

Restricted Information: the second challenge refers to the restricted information that the server is allowed to access due to the strict privacy constraints of our problem [14], [15]. In particular, FL requires the training data to be stored locally and only accessible to the data owners (i.e., the clients) to protect their privacy. However, without accessing the necessary information about the raw training data of the clients, the FL server cannot apply current fair ML solutions to address the fair classification problem. For example, a key assumption of existing fair ML is that the algorithmic bias is introduced by the "tyranny of the majority" - demographic groups with more training data often have lower error rates than demographic groups with fewer data [16]. Therefore, existing fair ML solutions often target at balancing the amount of training data of different demographic groups [6] to eliminate algorithmic bias. These approaches require access to either the raw training data or the underlying data distribution of the demographic groups, which are not allowed in our FL based problem setting.

Constrained Coordination: the third challenge refers to the difficulty in coordinating all clients to collaboratively train a fair and accurate classification model [17], [18]. In the FL setting, the performance of the classification model is directly controlled by the local training process of all clients. Therefore, it is essential to coordinate "when and how" each client performs the local update in order to optimize the fairness of the trained model. However, such a coordination is difficult due to the unique privacy constraints of FL. In particular, from the server's perspective, the clients are blackbox (due to restricted information challenge discussed above) so it is impossible for the server to guide the client's local training process. From the client's perspective, while it is possible to design its own local training strategy, the overall fairness objective will not be achieved without collaborating with other clients [9]. However, such collaborations between clients are constrained in FL as they may lead to various privacy leakages [19].

In this paper, we address the above challenges by developing a *fair* federated learning framework - FairFL that can ensure fairness in the classification outcome across multiple demographic groups in privacy-sensitive ML applications. To address the fairness and performance trade-offs, FairFL employs a carefully engineered reward mechanism to regulate the trade-off between the fairness and accuracy of the trained FL model and facilitates fairness across all demographic groups. FairFL jointly addresses the restricted information and constrained coordination challenges by developing a principled multi-agent reinforcement learning model and a secure aggregation protocol. Our solution allows the clients to individually decide the local updating strategies, and collaboratively optimize the classification model's fairness and accuracy without privacy violations on clients. We evaluate FairFL on two real-world applications: *criminal recidivism* and *income prediction*. We compared FairFL with the state-of-the-art FL and fair ML baselines. The results show that FairFL can train accurate classification models while significantly improve the fairness of trained models.

2. Related Work

Fairness in Machine Learning. Fairness in machine learning has received much attention recently [8], [20], [21]. Three major approaches have been developed to tackle the fairness problem in ML models. The first approach is to remove/obfuscate the sensitive attributes in the raw data before training, making the trained model unaware of the demographic groups of the participants [22]. However, such "unawareness" is not equivalent to "fairness", and often leads to biased models where some demographic groups have better prediction accuracy than others [23]. The second approach is to manipulate the loss function of the model so that the training process of the model satisfies certain fairness constraints [8]. However, these approaches are designed for training a single centralized model and not suitable for a distributed training environment [24]. The third approach is to eliminate the data imbalance to ensure the ML model is trained on balanced data across all demographic groups. There exist many data sampling techniques [6] dedicated to addressing the data imbalance problem. However, they often require the knowledge of either the sensitive attributes of the data or the statistics of the raw data (e.g., distribution of "male" vs "female" data entries), thus violate the privacy constraints of the FL. Our FairFL is one of the first decentralized fairness-aware learning approaches that jointly optimizes the accuracy and fairness of multiple demographic groups, while strictly satisfying the privacy requirements of the clients.

Federated Learning. FL allows multiple clients to collaboratively learn a shared machine learning model while keeping all the private training data of each client locally [24]. FL framework is a nice fit for machine learning applications where data are privacy-sensitive and distributed among different organizations/participants. While recent progress in FL targets at enhancing privacy [25], addressing resource constraints [26], and promoting scalability and communication efficiency [27], few efforts have been made to address the *fairness* issue in FL. Recently, Mohri *et al.*, proposed Agnostic FL (AFL), which defines a "good intent fairness" objective to minimize the maximum loss incurred on all data samples so that the global model is not overfit to any particular data distribution [28]. However, this model does not explicitly define fairness metrics in its learning objective, and only considers the fairness for protected class in a *single* demographic group. Another relevant piece of work is Astraea [6], which develops a self-balancing data sampling scheme to address the class imbalance problem in FL. However, Astraea does not explicitly address the fairness issue. In contrast, FairFL explicitly minimizes the discrimination index of the global model and ensures the trained model is both fair and accurate across *multiple demographic groups* simultaneously.

Multi-agent Reinforcement Learning. Our work is also relevant to the Multi-agent Reinforcement Learning (MARL) framework. We focus on a cooperative MARL setting where the agents collaboratively optimize a shared goal. A naïve approach to solve the cooperative MARL problem is to treat each agent independently such that it considers the rest of the agents as part of the environment [29]. Typical algorithms following this philosophy are independent Q-Learning (IQL) [30] and Nash Q-learning [31]. This naïve approach is problematic as the environment appears non-stationary from the view of any single agent [32]. To address this issue, Lowe *et al.* proposed the MADDPG model, which learns a centralized Q function for each agent that conditions on global information to alleviate the non-stationary problem and stabilize the training in MARL [33]. In this work, we develop a MARL based scheme to solve the fair classification problem in FL by enforcing an optimal client selection policy on each client. In particular, we design a set of novel reward and state functions that guide the clients to collaboratively make decisions of the local updates that optimize the fairness and accuracy of the global model.

3. Problem Formulation

We first provide a brief overview of the key terms and processes in FL and then formally define the FL based fair classification problem we address in this paper.

3.1. Federated Learning

There are two basic entities in a Federated Learning (FL) framework, i.e., the clients and the FL server. Let us assume there are a total of N clients and each client has its private local data $D_i, 1 \leq i \leq N$. The goal of the FL framework is to train a *global model*, denoted as M_G, which resides on the server without access to the client's private data. For example, consider a smart health application where a FL server (i.e., the model owner) targets at building a skin cancer detection model using neural networks (illustrated in Figure 1). The training data is spread across multiple hospitals and clinics (i.e., clients), which contain private health data that cannot be shared with other parties. In this example, the global model is the skin cancer detection model deployed at the server. In this paper, we focus on the classification-based machine learning models.

We assume the federated learning process is composed of a total of T communication rounds. At each communication round $t \in [1, T]$, the FL framework performs two key procedures: 1) the local update process, and 2) the global aggregate process as elaborated below.

Local update process of FL: Each client uses its own private data D_i to training a local model M_i, by minimizing a loss function $\mathcal{L}(w_i^t)$ based on D_i, where w_i^t is the set of trainable parameters (e.g., weights and bias of a neural network) of M_i at the t^{th} communication round. The w_i^t is updated using the gradient descent method as follows:

$$w_i^t = w_i^{t-1} - \eta \times \frac{\partial \mathcal{L}^t(w_i)}{\partial w_i^t} \tag{1}$$

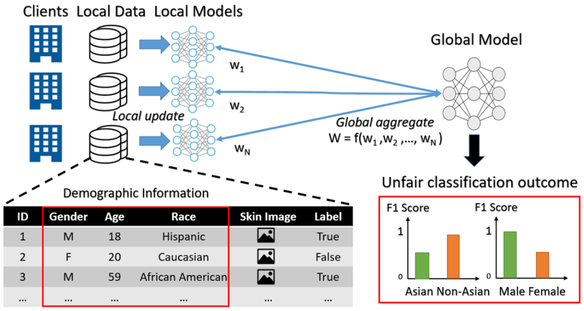


Figure 1. Illustration of Fair Federated Learning

Global aggregate process of FL: The global aggregate process of FL targets at identifying the optimal parameters (w^t) of the global model M_G. The weights w^t are calculated as the weighted sum of the local weight updates of all clients:

$$\Delta w^t = \sum_{i=1}^{N} \Delta w_i^t \times \frac{|D_i|}{\sum_{i=1}^{N} |D_i|} \tag{2}$$

where $|D_i|$ is the size of the local training data on the i^{th} client. The weights of the local models are then updated by obtaining the global weight: $w_i^t = w^t, \forall 1 \leq i \leq N$.

3.2. The Fairness Problem in FL

In FL, the client's data often contain certain private demographic information that is subject to discriminatory bias. For example, in the smart health application depicted in Figure 1, the health records in each client's local training data may contain different age groups, genders, and races. Formally, we define a *demographic group* as:

Definition 1. Demographic Group*: a set of data records sharing the same demographic attribute.*

We denote $\mathcal{A}$ to represent all demographic groups, and $\sigma \in A$ to depict a particular group, such as "Female" or "African American". We observe the data imbalance is a common issue in the data records, where there may exist more data records of a demographic group than the other, causing the model to be biased towards the majority group [16]. An example is shown in Figure 1 where the dominant data samples of the male demographic group cause the model to be unfair towards females.

The goal of fair FL, therefore, is to make sure the global model M_G is unbiased across *all* demographic groups, without knowing the private data (including the demographic information) of the clients. To ensure the fairness of FL, we first quantify the demographic bias of the global model and then try to minimize it. The demographic bias for a classification model is often measured as the difference in accuracy when performing classification tasks for different demographic groups [34]. If, for example, the classification accuracy for classifying the male's health records is higher than those of the female's, then the classifier is considered biased towards males. In this paper, we choose F1-score as the accuracy measure due to its robustness against the imbalanced dataset, and define a *discrimination index* to measure the bias.

Definition 2. Discrimination Index (**Φ_σ**): it is a metric to quantify the bias of the global machine learning model w towards a particular demographic group. We use subscript to denote the discrimination index of a demographic group. For example, Φ_{female} and Φ_{asian} depicts the discrimination index for females and Asians, respectively. Formally, we calculate the discrimination index as:

$$\Phi_\sigma = F1(w(X_\sigma^+)) - F1(w(X_\sigma^-)) \tag{3}$$

where $F1(X_\sigma^+)$ is the F1 measure of all the data samples that belong to the demographic group (e.g., "Asian"), and $F1(X_\sigma^-)$ is the F1 measure of all data samples that do not belong to that demographic group (e.g., "non-Asian"). The index Φ_σ falls between [-1,1]. Ideally, the ideal discrimination index should be as close to zero as possible (i.e., achieving the same accuracy for "Asian" and "Non-Asian").

3.3. Privacy Constraints

To capture the unique privacy constraints in the FL, we consider the following information of a participant as private and cannot be shared with other parties [35]: **PR1:** the raw training data samples in all participating clients; **PR2:** the ground truth labels of the raw training data; and **PR3:** the sensitive demographic information of each participant who contributes to the training data. In this paper, we strictly prohibit any violations of the above privacy requirements.

3.4. Objective

The ultimate goal of our problem is to minimize the loss function $\mathcal{L}(w)$ of the global model as well as the discrimination index Φ, while satisfying the privacy requirements defined above. We formulate a multi-objective constrained optimization problem as follows:

$$\begin{aligned} &\text{minimize} \quad \mathcal{L}(w) \\ &\text{minimize} \quad |\Phi_\sigma|, \quad \forall \sigma \in \mathcal{A} \\ &\text{s.t.:} \quad \textbf{PR1, PR2, PR3} \text{ are satisfied} \end{aligned} \tag{4}$$

Note that, our objective does not simply emphasize the fairness requirement (i.e., minimizing Φ), which could be naïvely achieved (e.g., by randomly guessing the output label for all demographic groups). Instead, we force the model to also minimize the global loss so that the final global model is both fair and accurate. The problem is further complicated by the strict privacy constraints, preventing the FL server from acquiring any insights into the private local data. In the next section, we present our framework FairFL to solve the proposed problem.

4. The FairFL Framework

The overview of FairFL is shown in Figure 2. It has two main components: a Team Markov Game for Client Selection (TMGCS) and a Secure Aggregation Protocol (SAP). In particular, TMGCS is a Multi-Agent Reinforcement Learning (MARL) based approach that allows clients to collaboratively decide whether to participate in the local update process. The SAP is designed to address the myopic view issue of the TMGCS component where each client has only access to its own local data due to the privacy concerns of clients. The protocol allows a client to gather the information about all the clients' status without violating their privacy constraints. We present the details of the FairFL framework below.

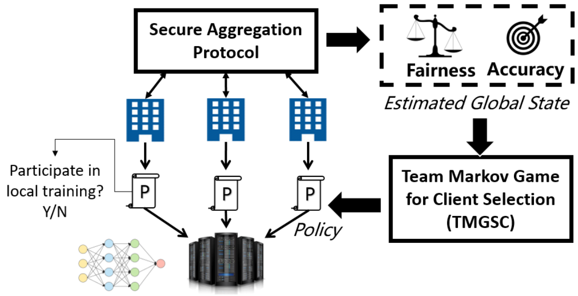

Secure Aggregation Protocol
Fairness
Accuracy
Estimated Global State
Team Markov Game for Client Selection (TMGSC)
Participate in local training? Y/N
P
P
P
Policy

Figure 2. Overview of FairFL

4.1. Team Markov Game for Client Selection (TMGCS)

The core of the FairFL framework is a novel Team Markov Game model for client selection. The client selection targets at identifying the set of clients to participate in the local update process at each communication round, that maximizes the fairness and accuracy of the global model. A major challenge in the client selection process lies in the black-box nature of FL. In particular, it is difficult to pinpoint which exact client can contribute most to the performance of the global model. The reason for such difficulty is two-fold: 1) the FL's training process lacks explainability and the relationship between a client's local update and the global model performance is not easy to predict [36]. 2) The global model is influenced by the joint decision of *all* clients. It is often difficult to attribute the global performance gain to the contribution of a particular client. These challenges motivate us to develop a new MARL based approach that allows multiple agents (i.e., clients) to interact with a blackbox environment and make optimized decisions for a long-term reward (i.e., fairness and accuracy in our problem). Our MARL problem is formulated as a cooperative multi-agent Markov Decision Process (MDP), namely a Team Markov Game as described below.

Definition 3. Team Markov Game is a tuple $\langle \alpha$, S, A, R, T $\rangle$, where α is a set of n agents; S is a finite state space of the environment that is observed by all agents; $A = \{A_1, A_2, ..., A_N\}$ is the joint action space of all n

agents; $R : S \times A \rightarrow \mathcal{R} = \{r_1, r_2, ..., r_N\}$ is the reward of each agent; and $T : S \times A \times S \rightarrow [0, 1]$ is the state transition function.

The goal of each agent in a Team Markov Game is to collaboratively find a deterministic joint policy (or joint strategy profile) $\pi = \{\pi_{i=1,2,..N}\}$ (where $\pi : S \rightarrow A$ and $\pi_i : S \rightarrow A_i$) so as to maximize the expected sum of their discounted rewards:

$$\mathbb{E}\left[\sum_{j=0}^{\infty} \gamma^j r_i^{t+j}\right] \tag{5}$$

where $\gamma \in (0, 1]$ is a discount factor that discounts the future rewards. The discount for future rewards takes an exponential form of γ^j, denoting the reward is discounted by γ every communication round after t [37].

The mapping of FairFL to a Team Markov Game problem turns out to be a non-trivial task where action, state, as well as the reward must be carefully designed in order to satisfy the objectives of FairFL. We first list a few key terms of our mapping below.

- **Environment:** it refers to both the global model and local models in the FairFL.
- **Agent:** it refers to the clients that interact with the environment by performing local updates. We use the term "agent" and "client" interchangeably in the rest of the paper.
- **Action** a_i^t**:** a binary value representing the choice of whether to participate (1 = "participate", 0 = "not participate") in a local update at a communication round t.
- **Reward (**r_i^t**):** a feedback (often a score) provided by the environment to the clients when an action is made. It is the objective that the clients are trying to maximize.
- **State (**S^t**):** a feature vector describing the current context of the environment at the communication round t.

The design of the state and reward directly affects how the optimal actions were taken given the current accuracy and fairness of the global model. We present our design of state and rewards in the subsequent sections.

4.2. Global State Approximation with Secure Aggregation Protocol (SAP)

In MARL, each action is chosen based on the current state - a feature vector representing the current status of the environment. In the client selection problem, we are particularly interested in the status of the global model in terms of bias and accuracy. Such status will help guide each client to decide whether it should participate in the local update or not. For a simple example, if the clients observe that the global model is biased towards the female group (i.e., lower accuracy for females), a client with more data samples from females should participate, while another client with more data samples from males should not participate because the extra data imbalance will further exacerbate the bias. With this intuition, we define the state of the TMGCS as follows:

Definition 4. Approximate Global Model State Vector: a vector denoting the current fairness and accuracy of the global model. Formally, it is defined as:

$$\left[\{\overline{\Phi}_\sigma | \sigma \in \mathcal{A}\}, \overline{F1}(M_G)\right] \tag{6}$$

where $\overline{F1}(M_G)$ represents the estimation of the F1 score of the global model, and $\{\overline{\Phi}_\sigma | \sigma \in \mathcal{A}\}$ is a set of the estimations of the discrimination indices for all demographic groups defined in Equation (3). The dimension of the state vector is $|\mathcal{A}| + 1$.

To estimate $\overline{F1}(M_G)$ and $\{\overline{\Phi}_\sigma | \sigma \in \mathcal{A}\}$, we define an aggregation function $g(\cdot)$ that combines the local discrimination indices and F1 scores to estimate those of the global model as follows.

$$\begin{gathered}\overline{F1}(M_G) = g(F1(M_1), ...F1(M_N)) = \sum_{i=1}^{N} \frac{|D_i| \times F1(M_i)}{|D|} \\ \overline{\Phi}_\sigma = g(\overline{\Phi}_{\sigma,1}, ...\overline{\Phi}_{\sigma,N}) = \sum_{i=1}^{N} \frac{|D_i| \times \overline{\Phi}_{\sigma,i}}{|D|}\end{gathered} \tag{7}$$

where $\overline{\Phi}_{\sigma,i}$ and $F1(M_i)$ is the discrimination index of demographic group σ and the F1 score of the i^{th} client, respectively. $|D_i|$ is the total size of training data of the i^{th} client and $|D| = \sum_{i=1}^{N} |D_i|$ is the size of the total amount of training data on all clients.

A special challenge in deriving the above equations lies in the fact that FairFL does not allow the direct share of the local discrimination index of a client to the server, which might reveal private information about the client. For example, if a client has a discrimination index of 0.6 for the "male" group, then the server can infer that the client has more male data samples than females in its private data. Therefore, we propose to develop a secure aggregation procedure that allows the server to calculate $g(\cdot)$ without accessing the local discrimination indices from clients. The intuition of our design is to hide the local discrimination index by applying a polynomial function based on the index. By aggregating the polynomials from all clients, the server is only able to reconstruct the aggregated statistics (i.e., $g(\cdot)$) instead of the original data from clients. We summarize the procedure in Algorithm 1. The privacy properties of the algorithm is based on Polynomial Interpolation, which has been well proved in existing literature [38].

4.3. Reward Design with Shapley Value

Recall that the goal of the TMGCS is to acquire the maximum rewards received from the environment. Therefore, the reward must be aligned with the objective of our problem, which is to maximize the global model accuracy and minimize the discrimination index of demographic groups. We derive the reward r_i for the i^{th} client as follows:

$$r_i = \sum^{\sigma \in \mathcal{A}} \overline{\text{SHAP}}_\sigma \times (|\Phi'_{\sigma,i}| - |\Phi_{\sigma,i}|) + \lambda \times (F1(M'_i) - F1(M_i)) \tag{8}$$

Algorithm 1 Secure Aggregation with Polynomial Interpolation

```
Input: The secrete values $x_1, x_2, ..., x_N$ (e.g., the discrimination indices of the local models).
Output: The sum of the secrete values.
The server generates a set of integers $e_1, e_2, ..., e_N$ for all the clients ($e_i$ for the $i^{th}$ client), such that $\sum_{i=1}^{N} e_i = 0$.
The server sends exponents $e_i$ to the $i^{th}$ client, $\forall 1 \leq i \leq N$.
for each client do
    Calculates a score $C_i = (1+N)^{x_i} \times Rand^{e_i}$.
    Sends $C_i$ to the server.
end for
Server receives $C_1, C_2, ...C_N$ from all clients, and calculates:
    $\prod_{i=1}^{N} C_i = \prod_{i=1}^{N} (1+N)^{x_i} \times Rand^{\sum_{i=1}^{N} e_i} \ mod \ n^2$
    $= (1+N)^{\sum_{i=1}^{N} x_i} \ mod \ N^2$
    $= 1 + (\sum_{i=1}^{N} x_i) \times N$
The server recovers the sum $\sum_{i=1}^{N} x_i = (\prod_{i=1}^{N} C_i - 1)/n$ and broadcasts to all clients.
```

where $|\Phi'_{\sigma,i}| - |\Phi_{\sigma,i}|$ and $F1(M'_i) - F1(M_i)$ denotes the improvements of the discrimination index and the F1-score of the local model M_i, respectively. λ is a tunable weighting factor controlling the relative importance between fairness improvement and accuracy improvement. We evaluate the importance of λ and the accuracy-fairness trade-off in Section 5.4. $\overline{\text{SHAP}}_\sigma$ is the estimated Shapley value of a demographic group. The Shapley value is a metric from machine learning that is used to quantify the marginal contribution of a factor [39]. Here, we adapt the Shapley value to detect the influence of a particular demographic group towards the model accuracy. A demographic group with a high Shapley value is considered as "accuracy sensitive".

To calculate $\overline{\text{SHAP}}_\sigma$, we first derived a local Shapley value on the i^{th} client as:

$$\text{SHAP}_{\sigma,i} = \sum^{Z \subseteq \mathcal{A} \setminus \{\sigma\}} \frac{|Z|!\,(|\mathcal{A}| - |Z| - 1)!}{\mathcal{A}!} \left(f\left(Z \cup \{\sigma\}\right) - f(Z)\right)$$
$$\overline{\text{SHAP}}_\sigma = g(\text{SHAP}_{\sigma,1}, \text{SHAP}_{\sigma,2}, ..., \text{SHAP}_{\sigma,N}) \tag{9}$$

where $\mathcal{A}$ is the set of all the demographic groups in the training data. Z is a subset of the demographic groups, σ is demographic group to be examined. $f(\cdot)$ is the prediction outcome of the local model, and $f\left(Z \cup \{\sigma\}\right) - f(Z)$ is the marginal contribution the model makes when data samples from demographic group σ is used in training the local model. The $\overline{\text{SHAP}}_\sigma$ of the global model is estimated using the secure aggregation protocol defined in Algorithm 1.

The key intuition of leveraging the Shapley value is to further balance the fairness improvements across different demographic groups. In particular, to maximize the fairness, it is essential to ensure the model is fair to all the demographic groups (i.e., minimizing Φ_σ for all $\sigma \in \mathcal{A}$). However, we observe that taming the fairness of some "accuracy sensitive" demographic groups (i.e., groups with high Shapley values) can sometimes negatively affect the model accuracy of other groups. Consider an example where the "female" group is a group that is more prone to model accuracy degradation than the "African American" group when we try to reduce the bias of both groups. Then a reward design that treats all demographic groups the same will guide the clients to only focus on reducing the bias of the "African American" group while ignoring the "female" group. To address this issue, we adapt the Shapley value as a weighting factor to motivate the clients to balance the fairness improvement across all demographic groups. The Shapley value is normalized to a range of [0,1] with standard min-max normalization.

4.4. Policy Gradient with Partial Information

Given the above definitions of action, state, and reward, our goal now is to find the optimal joint policy of all clients that maximizes the expected reward defined in Equation (5). In this paper, we adopt a multi-agent version of policy gradient method - MADDPG [33] to solve the proposed TMGCS problem. In the MADDPG model, each client is treated as an "actor" who receives the advice from a "critic". The job of the critic is to predict the value of an action in a particular state, which is then used by the actor to update its policy. Formally, let $\pi = \{\pi_1, \pi_2, ..., \pi_N\}$ be the set of policies for all clients that are parameterized by $\Theta = \{\Theta_1, \Theta_2, ...\Theta_N\}$. The policy gradient of the expected reward from each client $J(\Theta_i) = \mathbb{E}[\sum r_i]$ is derived as:

$$\nabla_{\Theta_i} J(\Theta_i) = \mathbb{E}_{S \sim p(\mu), a_i \sim \pi_i}[\nabla_{\Theta_i} log \pi_i(a_i|S) \times \Upsilon] \tag{10}$$

where μ is the policies chosen based on the parameter Θ, and $p(\mu)$ is the policy distribution of Θ. $\Upsilon = Q_i^\pi(S, a_1, a_2, ..., a_N)$ is an action-value function that takes as input the actions of all clients, as well as the state S, and outputs the Q-value for the i^{th} client. The Q-value can be learned by minimizing the loss function defined below:

$$\mathcal{L}(\Theta_i) = \mathbb{E}_{S,a_i,r_i,S'}[Q_i^\pi(S, a_1, ..., a_N) - y^2]$$
$$y = r_i + \gamma \times Q_i^{\mu'}(S', a'_1, a'_2, ..., a'_N) \tag{11}$$

where μ' denotes the policy obtained from updated parameter Θ'. To ensure all the clients act in a globally-coordinated way, we allow critics (deployed at the server) to access the states and actions of all the clients. Note that this step does not violate the privacy constraints of the clients. In deployment phase (i.e., when the optimal policies are learned and the actual federate learning process starts), a client will only need the access to the global state vector to derive its optimal action and does not need to know the information (e.g., the action and local data) from any other clients. Figure 3 illustrates the pre-training and deployment of the MARL process.

5. Evaluation

In this section, we conduct extensive experiments on two real-world datasets to answer the following questions about FairFL. **Q1:** Can FairFL effectively reduce bias and provide fairer classification outcomes than the state-of-the-art baselines? **Q2:** Can FairFL achieve better overall classification accuracy than the state-of-the-art baselines? **Q3:**

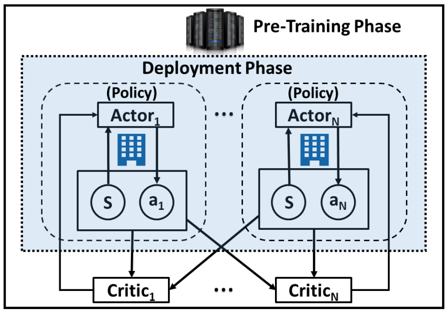


Figure 3. Illustration of The MARL for Solving TMGCS

What is the accuracy and fairness trade-off of FairFL? **Q4:** How does each component of FairFL contribute to its overall performance?

5.1. Experiment Setup

5.1.1. Dataset. We evaluate the FairFL on the two real-world datasets. These datasets were commonly used for fair machine learning literature and both of them are known to contain significant demographic bias.

COMPAS Recidivism Racial Bias dataset [1]**:** This dataset contains features and labels used by the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm in predicting recidivism of defendants within 2 years of the decision, for over 10,000 criminal defendants. The output of the algorithm is the prediction of "recidivism" (positive class) or not.

Adult Census Income dataset [2]**:** This dataset was extracted from the United States Census Database. The dataset consists of anonymous information such as occupation, age, native country, race, capital gain, capital loss, education, work class, etc. Each entry is predicted as either having a salary greater than "50K" (positive class) or not.

For both the COMPAS and Adult Income datasets, we use Multi-layer Perception (MLP) and support vector machine (SVM) as classification algorithms. The choice of these two algorithms is based on two factors: 1) the algorithm is effective when applying to the two datasets; 2) the algorithm has to be applicable to the FL framework (i.e., requiring a gradient descent approach for parameter learning). Therefore, we skip algorithms such as decision tree and random forest that can not be easily adapted to the FL framework. For both datasets, we focus on three categories of demographic groups - age, race, and gender.

5.1.2. Parameter Settings. We provide the critical parameter settings for FairFL. For the pre-training phase of the TMGCS, we set: policy learning rate = 0.0001, value learning rate = 0.001, $\gamma = 0.9$, $\lambda = 0.4$. For the FL phase, we set the number of clients as 10, B = 100 (local batch size) , and E = 1 (local epoch number).

5.1.3. Baseline. We found no existing approach that can be directly applied to solve our fairness problem in federated learning. Therefore, we adapt some of the state-of-the-art solutions from both fair machine learning and federated learning as our baselines.

- **AdaFair [40]**: a centralized fairness-aware classifier based on AdaBoost that updates the weights of the instances in boosting based on a fairness constraint.
- **Astraea [6]**: a self-balancing FL framework that reduces the classification bias via an adaptive sampling technique that regulates the class imbalance of the training data.
- **FedAvg [24]**: the original federated learning algorithm for distributed training of private data. It does not consider the fairness of different demographic groups.
- **Centralized:** an ideal centralized training scenario where the server has access to all the raw data from the clients.

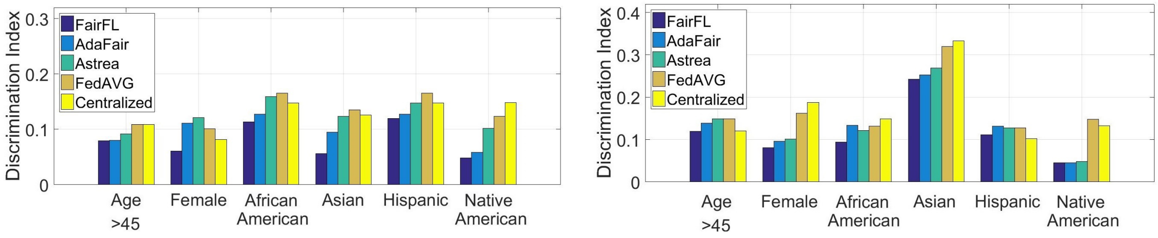
5.2. Q1: Evaluating Fairness

In the first set of experiments, we evaluate the fairness of all schemes across the two datasets. The results are presented in Figures 4 and 5. We can clearly observe that FairFL significantly reduces the discrimination index of all demographic groups. In particular, compared to the original FL algorithm (FedAVG), FairFL is able to reduce the discrimination index by 13.2% to 69.4% in the COMPAS dataset, and by 21.0% to 68.2% in the Adult Income dataset across all demographic groups, respectively. To provide a more comprehensive understanding of the fairness improvement achieved by FairFL, we further evaluate the fairness of all schemes using additional fairness metrics from existing literature. We choose three popular metrics - Equalized Odds (Eq. Odds), Difference in True Positive Rates (Diff. TPR), and Difference in True Negative Rates (Diff. TNR) [23]. Due to the space limit, we only present the average results across all demographic groups in Table 1. We again observe that FairFL achieves the best fairness performance across all metrics in both datasets. Compared to the best-performing baseline AdaFair, FairFL is able to achieve a performance gain as high as 31.3%, 28.3%, and 36.4% in the COMPAS dataset on Eq. Odds, Diff. TPR, and Diff. TNR metrics, and 45.6%, 20%, and 15.6% in the Adult Income dataset. This further demonstrates that FairFL can achieve significantly better fairness for demographic groups compared to all baselines.

5.3. Q2: Evaluating Classification Accuracy

In the second set of experiments, we evaluate the classification accuracy of the compared schemes. The results are presented in Table 2. We have observed that FairFL is able to outperform all baselines in terms of F1 scores except the Centralized scheme, which represents the upper bound of the accuracy by assuming the server has access to all raw data of clients, which unfortunately violates the privacy preservation objective of our problem as discussed in Section 3. The results suggest that FairFL cannot only significantly improve the fairness of the model, but also achieve

[1] https://www.kaggle.com/danofer/compass
[2] https://archive.ics.uci.edu/ml/datasets/census+income

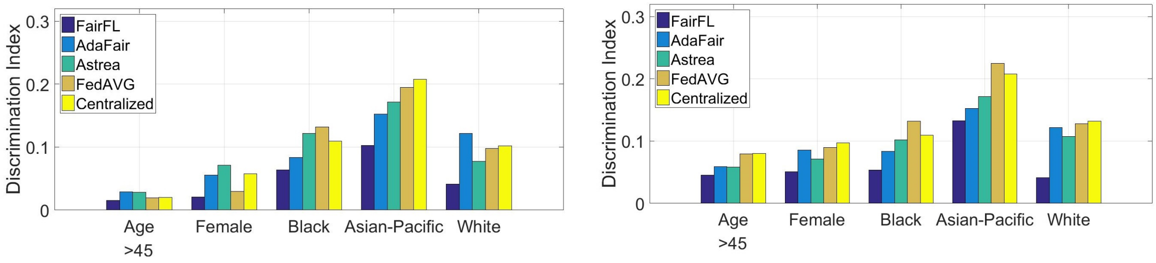


(a) SVM Algorithm

(b) MLP Algorithm

*The figures show the discrimination indices of "Age >45 vs. Age ≤45", "female vs. male", "African American vs. Non-African American", "Asian vs. Non-Asian", "Hispanic vs. Non-Hispanic", and "Native American vs. Non-Native American"

Figure 4. Fine-grained Discrimination Index for COMPAS Recidivism Dataset



(a) SVM Algorithm

(b) MLP Algorithm

*The figures show the discrimination indices of "Age > 45 vs. Age ≤45", "female vs. male", "Black vs. Non-Black", "Asian-Pacific vs. Non-Asian-Pacific", , and "White vs. Non-White"

Figure 5. Fine-grained Discrimination Index for Adult Income Dataset

Table 1. Overall Fairness using Other Metrics from SoTA Literature

	COMPAS						Adult Income					
	SVM			MLP			SVM			MLP		
Schemes	Eq. Odds	Diff. TPR	Diff. TNR	Eq. Odds	Diff. TPR	Diff. TNR	Eq. Odds	Diff. TPR	Diff. TNR	Eq. Odds	Diff. TPR	Diff. TNR
AdaFair	0.118	0.145	0.137	0.132	0.148	0.095	0.092	0.122	0.107	0.112	0.130	0.077
Astrea	0.218	0.170	0.192	0.155	0.198	0.091	0.103	0.111	0.091	0.103	0.125	0.993
FedAvg	0.240	0.258	0.132	0.148	0.172	0.129	0.099	0.132	0.152	0.157	0.182	0.141
Centralized	0.138	0.237	0.233	0.192	0.177	0.212	0.143	0.179	0.133	0.152	0.155	0.131
FairFL	**0.081**	**0.103**	**0.087**	**0.089**	**0.095**	**0.082**	**0.050**	**0.088**	**0.046**	**0.073**	**0.102**	**0.065**

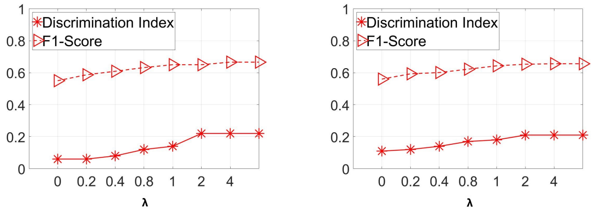
Table 2. Classification Accuracy

	COMPAS						Adult Income					
	SVM			MLP			SVM			MLP		
Schemes	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
AdaFair	0.473	0.929	0.627	0.454	0.987	0.622	0.652	0.877	0.748	0.553	0.895	0.684
Astrea	0.448	1.000	0.619	0.448	0.998	0.619	0.614	0.895	0.728	0.588	0.862	0.699
FedAvg	0.474	**0.958**	0.635	0.448	**1.000**	0.618	0.582	0.851	0.691	0.608	**0.877**	0.718
Centralized	0.562	0.837	0.673	0.589	0.770	0.668	0.699	0.893	0.784	0.643	0.895	0.748
FairFL	**0.504**	0.926	**0.652**	**0.483**	0.964	**0.643**	**0.667**	**0.872**	**0.756**	**0.609**	**0.877**	**0.719**

a better classification accuracy compared to the state-of-the-art baselines. We attribute such a performance gain to the objective function in Equation (5), which explicitly targets at maximizing both model fairness and accuracy of the FairFL framework.

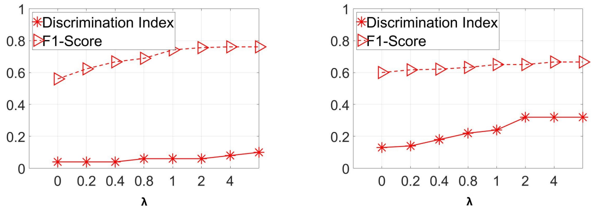
5.4. Q3: Accuracy and Fairness Trade-off

In the third set of experiments, we further investigate the accuracy and fairness trade-off of FairFL. Recall that FairFL achieves the accuracy and fairness trade-off via the control of λ defined in Equation 8. To study this important tradeoff, we tune the value of λ from 0 (i.e., model accuracy is not considered) to 4 (i.e., the accuracy is 4 times more important than the fairness) and present the results in Figure 6 and 7. We can clearly observe that when λ increases the model accuracy (as captured by F1 score) improves, while the fairness of the model degrades (i.e., the discrimination index increases). This is because when increasing λ, the model will focus more on the accuracy improvement rather than fairness.



(a) SVM (b) MLP

Figure 6. Accuracy-Fairness Trade-off of COMPAS Dataset



(a) SVM (b) MLP

Figure 7. Accuracy-Fairness Trade-off of Income Dataset

5.5. Q4: Ablation Study

Finally, we perform an ablation study to carefully examine the contribution of each component of the FairFL scheme. In particular, we compare the original FairFL with the following three baselines: FairFL without the Shapley value defined in Equation (8), FairFL without the Shapley value aggregation process described in Algorithm 1, and FairFL with only static state vectors. The results for the two datasets are presented in Tables 3 and 4. We observe that by completely removing the Shapley value, the model accuracy of FairFL almost stays the same, but the average discrimination index significantly increases (i.e., fairness decreases). This is due to the fact that the Shapley is only designed to balance the fairness improvements of the demographic groups. We also observe that using local Shapley value without performing global aggregation will result in 15.2% and 37.5% increase in discrimination index for the COMPAS and Adult Income dataset, respectively. The performance deficiency highlights the necessity of the secure aggregation protocol we designed that allows the clients to share important information without sacrificing their privacy. The state vector turns out to be a critical design of the FairFL framework. By replacing the dynamic state vector with a static vector, both fairness and model accuracy degrades (as high as 32.6% increase in discrimination index, and 5.4% decrease in F1 score). This justifies our design of the state vector for the proposed Markov game in FairFL.

Table 3. Ablation Study - Discrimination Index

	COMPAS		Income	
Schemes	SVM	MLP	SVM	MLP
FairFL	**0.092**	**0.103**	**0.040**	**0.061**
FairFL w/o Shap. Value	0.114	0.117	0.047	0.078
FairFL w/o Shap. Aggregation	0.106	0.109	0.055	0.072
FairFL w/ Static State Vector	0.122	0.137	0.069	0.104

Table 4. Ablation Study - F1

	COMPAS		Income	
Schemes	SVM	MLP	SVM	MLP
FairFL	**0.652**	**0.643**	**0.756**	**0.719**
FairFL w/o Shap. Value	0.648	0.642	0.755	0.715
FairFL w/o Shap. Aggregation	0.650	0.644	0.755	0.718
FairFL w/ Static State Vector	0.631	0.613	0.722	0.680

6. Conclusion

This paper presents the FairFL, a fairness-aware federated learning framework to perform privacy-aware training for classification tasks. FairFL ensures the trained model to be unbiased towards multiple demographic groups simultaneously. We develop a novel MARL-based client selection scheme to jointly maximize model accuracy and minimize the discrimination index of the classification outcome. A secure aggregation protocol has been developed to allow clients to coordinate and share information without privacy violation. Evaluation results through two real-world datasets demonstrate that FairFL achieves more favorable fairness performance compared to the state-of-the-art baselines.

Acknowledgment

This research is supported in part by the National Science Foundation under Grant No. CNS-1845639, CNS-1831669, Army Research Office under Grant W911NF-17-1-0409. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[2] D. Wang, T. Abdelzaher, and L. Kaplan, *Social sensing: building reliable systems on unreliable data*. Morgan Kaufmann, 2015.

[3] D. Roselli, J. Matthews, and N. Talagala, "Managing bias in ai," in *Companion Proceedings of The 2019 World Wide Web Conference*, 2019, pp. 539–544.

[4] M. T. Rashid and D. Wang, "Covidsens: a vision on reliable social sensing for covid-19," *Artificial Intelligence Review*, pp. 1–25, 2020.

[5] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic *et al.*, "Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *arXiv preprint arXiv:1810.01943*, 2018.

[6] M. Duan, "Astraea: Self-balancing federated learning for improving classification accuracy of mobile deep learning applications," *arXiv preprint arXiv:1907.01132*, 2019.

[7] F. Kamiran and T. Calders, "Classification with no discrimination by preferential sampling," in *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*. Citeseer, 2010, pp. 1–6.

[8] W. Zhang and E. Ntoutsi, "Faht: an adaptive fairness-aware decision tree classifier," *arXiv preprint arXiv:1907.07237*, 2019.

[9] H. Hu, Y. Liu, Z. Wang, and C. Lan, "A distributed fair machine learning framework with private demographic data protection," *arXiv preprint arXiv:1909.08081*, 2019.

[10] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher, "On truth discovery in social sensing: A maximum likelihood estimation approach," in *Proceedings of the 11th international conference on Information Processing in Sensor Networks*, 2012, pp. 233–244.

[11] D. Wang, M. T. Amin, S. Li, T. Abdelzaher, L. Kaplan, S. Gu, C. Pan, H. Liu, C. C. Aggarwal, R. Ganti *et al.*, "Using humans as sensors: an estimation-theoretic perspective," in *IPSN-14 proceedings of the 13th international symposium on information processing in sensor networks*. IEEE, 2014, pp. 35–46.

[12] D. Wang, B. K. Szymanski, T. Abdelzaher, H. Ji, and L. Kaplan, "The age of social sensing," *Computer*, vol. 52, no. 1, pp. 36–45, 2019.

[13] M. T. Al Amin, T. Abdelzaher, D. Wang, and B. Szymanski, "Crowdsensing with polarized sources," in *2014 IEEE International Conference on Distributed Computing in Sensor Systems*. IEEE, 2014, pp. 67–74.

[14] D. Zhang, Y. Ma, X. S. Hu, and D. Wang, "Towards privacy-aware task allocation in social sensing based edge computing systems," *arXiv preprint arXiv:2006.03178*, 2020.

[15] D. Wang, D. Zhang, Y. Zhang, M. T. Rashid, L. Shang, and N. Wei, "Social edge intelligence: Integrating human and artificial intelligence at the edge," in *2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE, 2019, pp. 194–201.

[16] B. Ustun, Y. Liu, and D. Parkes, "Fairness without harm: Decoupled classifiers with preference guarantees," in *International Conference on Machine Learning*, 2019, pp. 6373–6382.

[17] D. Zhang, Y. Ma, Y. Zhang, S. Lin, X. S. Hu, and D. Wang, "A real-time and non-cooperative task allocation framework for social sensing applications in edge computing systems," in *2018 IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*. IEEE, 2018, pp. 316–326.

[18] D. Zhang, Y. Ma, C. Zheng, Y. Zhang, X. S. Hu, and D. Wang, "Cooperative-competitive task allocation in edge computing for delay-sensitive social sensing," in *2018 IEEE/ACM Symposium on Edge Computing (SEC)*. IEEE, 2018, pp. 243–259.

[19] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," *arXiv preprint arXiv:1712.07557*, 2017.

[20] D. Y. Zhang, Y. Huang, Y. Zhang, and D. Wang, "Crowd-assisted disaster scene assessment with human-ai interactive attention." in *AAAI*, 2020, pp. 2717–2724.

[21] D. Zhang, Y. Zhang, Q. Li, T. Plummer, and D. Wang, "Crowdlearn: A crowd-ai hybrid system for deep learning-based damage assessment applications," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2019, pp. 1221–1232.

[22] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[23] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," *arXiv preprint arXiv:1507.05259*, 2015.

[24] H. B. McMahan, E. Moore, D. Ramage, S. Hampson *et al.*, "Communication-efficient learning of deep networks from decentralized data," *arXiv preprint arXiv:1602.05629*, 2016.

[25] S. Hardy, W. Henecka, H. Ivey-Law, R. Nock, G. Patrini, G. Smith, and B. Thorne, "Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption," *arXiv preprint arXiv:1711.10677*, 2017.

[26] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1205–1221, 2019.

[27] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[28] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," *arXiv preprint arXiv:1902.00146*, 2019.

[29] M. T. Rashid, D. Y. Zhang, and D. Wang, "Socialdrone: An integrated social media and drone sensing system for reliable disaster response," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2020, pp. 218–227.

[30] M. Tan, "Multi-agent reinforcement learning: Independent vs. cooperative agents," in *Proceedings of the tenth international conference on machine learning*, 1993, pp. 330–337.

[31] J. Hu and M. P. Wellman, "Nash q-learning for general-sum stochastic games," *Journal of machine learning research*, vol. 4, no. Nov, pp. 1039–1069, 2003.

[32] M. T. Rashid, D. Y. Zhang, and D. Wang, "Dasc: Towards a road damage-aware social-media-driven car sensing framework for disaster response applications," *Pervasive and Mobile Computing*, vol. 67, p. 101207, 2020.

[33] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Advances in neural information processing systems*, 2017, pp. 6379–6390.

[34] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, 2016, pp. 3315–3323.

[35] N. Vance, D. Y. Zhang, Y. Zhang, and D. Wang, "Privacy-aware edge computing in social sensing applications using ring signatures," in *2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS)*. IEEE, 2018, pp. 755–762.

[36] G. Wang, "Interpret federated learning with shapley values," *arXiv preprint arXiv:1905.04519*, 2019.

[37] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[38] E. Karnin, J. Greene, and M. Hellman, "On secret sharing systems," *IEEE Transactions on Information Theory*, vol. 29, no. 1, pp. 35–41, 1983.

[39] S. B. Cohen, E. Ruppin, and G. Dror, "Feature selection based on the shapley value." in *IJCAI*, vol. 5, 2005, pp. 665–670.

[40] V. Iosifidis and E. Ntoutsi, "Adafair: Cumulative fairness adaptive boosting," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 781–790.