FairTrade: Achieving Pareto-Optimal Trade-Offs between Balanced Accuracy and Fairness in Federated Learning

Maryam Badar, Sandipan Sikdar, Wolfgang Nejdl, Marco Fisichella

L3S Research Center, Leibniz University, Hannover, Germany
{badar, sandipan.sikdar, nejdl, mfisichella}@l3s.de

Abstract

As Federated Learning (FL) gains prominence in distributed machine learning applications, achieving fairness without compromising predictive performance becomes paramount. The data being gathered from distributed clients in an FL environment often leads to class imbalance. In such scenarios, balanced accuracy rather than accuracy is the true representation of model performance. However, most state-of-the-art fair FL methods report accuracy as the measure of performance, which can lead to misguided interpretations of the model's effectiveness to mitigate discrimination. To the best of our knowledge, this work presents the first attempt towards achieving Pareto-optimal trade-offs between balanced accuracy and fairness in a federated environment (FairTrade). By utilizing multi-objective optimization, the framework negotiates the intricate balance between model's balanced accuracy and fairness. The framework's agnostic design adeptly accommodates both statistical and causal fairness notions, ensuring its adaptability across diverse FL contexts. We provide empirical evidence of our framework's efficacy through extensive experiments on five real-world datasets and comparisons with six baselines. The empirical results underscore the potential of our framework in improving the trade-off between fairness and balanced accuracy in FL applications.

Introduction

Federated learning (FL) has emerged as a transformative strategy for distributed machine learning (ML) systems. By allowing local data storage and global model improvements without compromising data privacy, it presents a sustainable approach for model training and evaluation (Kairouz et al. 2021; Fisichella, Lax, and Russo 2022). However, one critical challenge that often goes unaddressed in such setups is ensuring fair and accurate model outcomes (Emelianov, Gast, and Gummadi 2022).

While there exists several methods in the literature to enhance fairness of ML models in a traditional centralized setup (Mehrabi et al. 2021), they cannot be directly deployed to FL settings owing to its distributed nature. Recently, few methods (Konečnỳ et al. 2016; Du et al. 2021; Cui 2021; Ezzeldin et al. 2023; Zeng, Chen, and Lee 2021) have been proposed to address the issue of fairness in FL. However, most of these methods ignore the presence of class imbalance in the data (Younis and Fisichella 2022), which can be a crucial source of bias (Dullerud et al. 2021). Moreover, in a decentralized set-up as in FL where the training data spans diverse clients, class imbalance is commonplace (Ramaswamy et al. 2019; Wang et al. 2021). Consequently, the performance measured in terms of accuracy as reported by these methods, is not representative of the true performance. As an illustrative example, consider a dataset with 80% positive samples and 20% negative samples. Each sample (x) is associated with a sensitive attribute $S \in M, F$. Consider a classifier f which always predicts a positive class i.e., $f(x) = 1$. For such a classifier, accuracy is 0.80 and discrimination measured in terms of the difference in probability of being assigned the positive class (aka demographic parity) is 0 as f always predicts 1 irrespective of the sensitive attribute. In fact, the low discrimination score achieved here is not at all reflective of the model's discrimination mitigation capability. The true performance of a classifier in such a scenario is revealed through balanced accuracy, which for the classifier in the above example is 0.5. Our experiments on several real-world datasets show that the state-of-the-art (SoTA) discrimination mitigation methods in FL can achieve a balanced accuracy as low as 0.5 rendering the low discrimination score achieved in those cases inconsequential. To address this, we propose FairTrade which utilizes multi-objective optimization (MOO) to jointly maximize balanced accuracy and minimize discrimination. Given these objectives need not necessarily be differentiable or expensive to compute, we further utilize Bayesian optimization to achieve our goal. Experiments on a range of real-world datasets show that FairTrade achieves the best performance-fairness tradeoff. We also find our method to be more efficient, achieving convergence in fewer communication rounds compared to the baseline methods. Our model also seamlessly adapts to other notions of fairness (e.g., causal fairness) demonstrating its generalizability.

To summarize the key contributions, we highlight the importance of evaluating discrimination mitigation methods using balanced accuracy rather than accuracy alone. We further introduce a metric-agnostic design that accommodates a variety of fairness notions, ensuring applicability across different FL contexts. Notably, FairTrade guarantees zero privacy leakage, as global and local updates are exchanged between

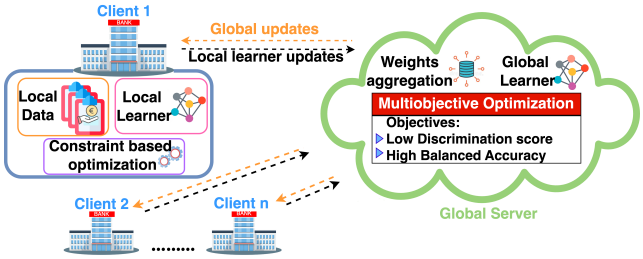

Figure 1: FairTrade: Federated framework with fairness-constrained multi-objective optimization.

server and client models through a secure aggregation process. Our method is rigorously evaluated across a wide range of benchmark datasets and SoTA fairness aware FL methods, demonstrating its efficacy across these settings. For reproducibility, all resources associated with our research, including code and data, are publicly accessible at the provided repository link [1].

Preliminaries

We commence with an overview of the conventional FL framework (FedAvg) as per (McMahan and et al 2017), subsequently, we delineate key concepts central to FairTrade.

Federated Learning Setup

Consider we have n local clients $(C_1, C_2, ...C_n)$ in an FL environment and a global server G. Each client has its own local dataset D_k with feature space X and output space Y. We consider a binary classification problem, i.e., $Y \epsilon \{0, 1\}$. Each client C_k's dataset D_k has m_k instances where each instance is defined as $I_j^k = \{x_j, y_j\}$, $j \epsilon [1, m_k]$. The global server G learns the predictive function $f(x) = y$ through the collaborative training of the local clients $(C_1, C_2, ...C_n)$. The server aggregates and averages local model updates, weighting them according to the size of each client's dataset. Precisely, the objective is to find parameter vector (ψ) that minimizes the weighted average of the loss across all clients, as presented in Equation (1).

$$\min_{\psi} f(\psi) = w \sum_{k=1}^{n} L_k(\psi) \tag{1}$$

FedAvg ensures security, scalability, and performance, yet its predictions can exhibit demographic biases in datasets. While there exists several notions of fairness, in this paper we consider two, namely (i) statistical group fairness and (ii) causal group fairness. Note that our method could also be deployed in conjunction with other fairness notions.

Fairness Notions

Discrimination refers to the unfair treatment or bias towards certain groups or individuals based on specific characteristics, such as race, gender, or socioeconomic status. These characteristics are often referred to as sensitive attributes.

[1] https://github.com/badarm/FairTrade

We assume that the datasets used to train and test the proposed model have a single sensitive attribute S *(e.g. "gender")* with binary values: s_0 for the protected group (e.g., "female") and s_1 for the non-protected (e.g., "male").

Statistical Group Fairness Notion: We gauge the discriminating behavior of the proposed method by demographic parity (DP) (Equation 2) (Verma and Rubin 2018) which is a statistical group fairness notion.

$$DP = P(f(x) = y^+ \mid S = s_1) - P(f(x) = y^+ \mid S = s_0) \tag{2}$$

Essentially, it represents the difference in mean positive outcomes of protected and non-protected group. $DP = 0$ denotes a perfectly fair classifier, whereas $DP = 1$ or -1 signifies complete unfairness.

Causal Group Fairness Notion: Despite the simplicity and popularity of statistical fairness methods, they might overcorrect, struggle with paradox resolution, and be vulnerable to shifts in data distributions (Makhlouf, Zhioua, and Palamidessi 2020). On the other hand, causal fairness considers underlying causal structures, decoupling predictions from sensitive attributes and providing a deeper insight into data biases. We have utilized the causal group fairness notion average treatment effect (ATE/FACE) (Khademi et al. 2019) to gauge the discrimination embedded in the predictions of the proposed framework as presented in Equation (3). We modified FACE to consider predicted outcomes.

$$\begin{aligned} FACE &= \mathbb{E}(|Y_{pot}^{s_1} - Y_{pred}^{s_1}| - |Y_{pot}^{s_0} - Y_{pred}^{s_0}|) \\ &= \frac{1}{n} \sum_{i=1}^{n} |Y_{pot\_i}^{s_1} - Y_{pred\_i}^{s_1}| - |Y_{pot\_i}^{s_0} - Y_{pred\_i}^{s_0}| \end{aligned} \tag{3}$$

Here $Y_{pot}^{s_1}$ and $Y_{pred}^{s_1}$ represent the potential and predicted outcomes when $S = s_1$. FACE quantifies the difference in the true positive outcomes (observed and potential) between the protected (treated) and non-protected groups (non-treated). $FACE = 1, -1$ indicates complete unfairness, whereas $FACE = 0$ signifies a perfectly fair classifier.

Fairness and Balanced Accuracy for FL

The fairness notions given in Equation (2) and Equation (3) can be employed in any centralized fairness aware learning framework. However, in the context of FL, the non-Independent Identically Distributed (non-IID) data distribution across clients necessitates a distinction between *client-side fairness* and *server-side fairness*.

The client-side fairness can be defined using Equation (2) and Equation (3). For example, DP for client k with local dataset D_k can be defined as:

$$\begin{aligned} disc\_score_k = P(f(x) = y^+ \mid S = s_1, \mathcal{D} = D_k) \\ - P(f(x) = y^+ \mid S = s_0, \mathcal{D} = D_k). \end{aligned} \tag{4}$$

However, the server-side fairness ($disc\_score_g$) considers the complete dataset $D_g = \bigcup_{k \in K} D_k$. If the data distribution among clients is IID then client-side fairness and server-side fairness are identical. For a classifier $f(x)$, the server-side DP can be specified as:

$$\begin{aligned} disc\_score_g = P(f(x) = y^+ \mid S = s_1, \mathcal{D} = D_g) \\ - P(f(x) = y^+ \mid S = s_0, \mathcal{D} = D_g). \end{aligned} \tag{5}$$

The main challenge is computing server-side fairness without access to client data stores. Server-side fairness can be computed through the secure aggregation of client-side fairness measures (Knott et al. 2021). If DP is the fairness metric then server-side fairness can be quantified as:

$$disc\_score_g = \sum_{k=1}^{K} w_k (P(f(x) = y^+ \mid S = s_1, \mathcal{D} = D_k) - P(f(x) = y^+ \mid S = s_0, \mathcal{D} = D_k)). \quad (6)$$

where w_k is the proportion of data points at client k relative to total data points across all clients i.e., $w_k = \frac{|D_k|}{\sum_j |D_j|}$. A similar server-side fairness measure, when the selected fairness metric is $FACE$ is provided in the Appendix.
The server-side balanced accuracy (BA_g) is computed as:

$$BA_g = \sum_{k=1}^{K} w_k BA_k, \quad (7)$$

In fairness-aware FL, the decentralized nature amplifies challenges in finding the optimal trade-off between the two conflicting objectives i.e., balanced accuracy and fairness. By integrating Multi-Objective Optimization (MOO) with Bayesian optimization (BO) we effectively address these challenges. While MOO concurrently optimizes for balanced accuracy and fairness, BO further sharpens this by probabilistically targeting the vast solution space, expediting convergence to the Pareto optimal trade-offs.

Multi-objective Optimization (MOO)

In multi-objective optimization (MOO), we aim to optimize a vector-valued objective $\theta(u) : M^d \rightarrow \mathbb{R}^N$ with $\theta(u) = \theta^{(1)}(u), ..., \theta^{(N)}(u)$ over a bounded set of inputs $U \subset \mathbb{R}^d$. The functions $\theta^{(j)}$ are computationally expensive to evaluate black-box functions. The MOO paradigm aims to find a set of Pareto optimal solutions, where improving one objective compromises another, with the overall goal of maximizing all objectives. A solution $\theta(u)$ dominates another solution $\theta(u')$ i.e., $\theta(u) \succ \theta(u')$ if $\theta^{(n)}(u) \geq \theta^{(n)}(u')$ for $n = 1, ..., N$ and $\exists\ n \in \{1, ..., N\}\ s.t.\ \theta^{(n)}(u) > \theta^{(n)}(u')$. The set of Pareto frontier solutions and Pareto frontier inputs can be represented as $\mathcal{P}^* = \{\theta(u)\ s.t.\ \nexists\ u' \in U : \theta(u') > \theta(u)\}$ and $U = \{u \in U\ s.t.\ \theta(u) \in \mathcal{P}^*\}$ respectively. Pareto frontiers are an infinite set of points; the goal is to find a finite approximate frontier. Provided with the approximate Pareto frontiers, the decision maker can select a Pareto optimal trade-off between conflicting objectives according to her preferences.

Bayesian Optimization (BO)

Bayesian optimization (BO) (Jones, Schonlau, and Welch 1998) serves as a robust approach for the optimization of black-box functions that are computationally expensive to evaluate. Utilizing a probabilistic surrogate model, a Gaussian Process (GP) (Rasmussen 2003) and the observed data $D = \{(u_i, y_i) | i = 1, ..., m\}$, BO provides a posterior distribution $\mathbb{P}(f|D)$ over true function values f. An acquisition function $\alpha \in U \mapsto \mathbb{R}$ employs this surrogate model (GP) to assign utility to a set of candidate inputs $U = \{u_i | i = 1, ..., q\}$ for evaluation on the actual function f. This surrogate based acquisition function is much less computationally expensive than the true function f.

Counterfactual Outcomes

To incorporate causal fairness, we calculate potential outcomes using a matching technique. The objective is to compute the potential outcomes by finding the matched neighbors from the opposite group. For instance, in loan approval, the counterfactual outcome for a female x_k as if she were a male is based on similar males' observed outcomes. To determine similarity between individuals x_j and x_k, we use Propensity Score Matching (PSM). PSM is aimed at estimating the effect of a treatment by accounting for the covariates that predict receiving the treatment. The propensity score ($e(x_k)$) is the probability of receiving the treatment given observed covariates. For loan approval example, $S_k = 1$ denotes the individual k who received the treatment i.e., the individual is female and $S_k = 0$ otherwise. Propensity score of x_k derived from observed covariates C_k is:

$$e(x_k) = \mathbb{P}(S_k = 1 | C_k) \quad (8)$$

The similarity between individuals x_j and x_k is determined through their propensity score difference. The logit version of this difference helps in reducing bias (Stuart 2010):

$$Diff(j, k) = |logit\,(e(x_j)) - logit\,(e(x_k))\,|. \quad (9)$$

We match treated (protected) and control (non-protected) individuals using nearest neighbor matching with replacement, based on aforementioned similarity metric.

FairTrade: Fairness constrained MOO framework

This section presents our FL framework that integrates fairness-constrained optimization at the client side and MOO at the server side. Figure 1 represents the conceptual model underpinning FairTrade. Our base learner is a deep neural network with three fully connected layers interspersed with ReLU activation functions, ending with a Sigmoid activation. In every communication round, each client trains a local learner, detects discrimination, and uses fairness constrained optimization for discrimination mitigation. Subsequently, clients send their weights along with local balanced accuracy (BA_k) and local discrimination score ($disc\_score_k$) values to the global server, which applies MOO to optimize the fairness constraint regularization parameter (ζ) and learning rate (lr) to guarantee Pareto optimal trade-offs between balanced accuracy (BA) and discrimination score ($disc\_score$). Details follow in later sections.

Client Side: Fairness Constrained Optimization

In our framework, FairTrade, we employ fairness-constrained optimization at each client (Agarwal et al. 2018) to achieve client-side fairness as detailed in the section "Preliminaries".
Each client possesses a local dataset (D_k) with underlying

demographic biases. The optimization task is to train a predictive model that satisfies specific fairness constraints. The optimization problem at the client side for a local classifier f parameterized by ψ can be formulated as follows:

$$\underset{\psi}{\text{minimize}}\ J(\psi) + \zeta * F(\psi)\ s.t.\ g(\psi) \leq e \tag{10}$$

where $J(\psi)$ is the local predictive loss function, $F(\psi)$ represents the fairness penalty, ζ is a regularization constraint parameter (which will be optimized through MOO at server-side), $g(\psi)$ is selected fairness metric, and e is the fairness budget. For each fairness notion, a set of linear constraints of the following form can be generated:

$$Q\eta(f) \leq e, \tag{11}$$

where Q is a matrix $\mathbb{R}^{|\mathcal{Z}|\times|\mathcal{V}|}$ and e is a vector $\mathbb{R}^{|\mathcal{Z}|}$ that represents the fairness budget allocated for each value of the sensitive attribute (e.g. male and female), and $\eta(f)$ denotes a vector consisting of conditional moments, given by:

$$\eta_v(f) = \mathbb{E}[h_v(X, S, Y, f(X))|\varphi_v]\ for\ v \in \mathcal{V}, \tag{12}$$

where $\mathcal{V} = \mathcal{S} \cup \{\mathcal{X} \backslash \mathcal{S}\}$, $h_v : \mathcal{X} \times \mathcal{S} \times \{0,1\} \times \{0,1\} \rightarrow [0,1]$ captures how the prediction $f(X)$ varies for different subsets of the data (defined by v and conditioned on φ_v), while considering the true labels Y, input data X, and sensitive attributes S. φ_v conditions the data based on a specific criterion; for instance, in the loan approval use case φ might be "the applicant is female". Now we define constraints for the fairness notion demographic parity (DP).

Constraint for Demographic Parity: Assuming a binary classification task and a binary sensitive attribute, DP can be expressed as a set of two equality constraints of the form:

$$\mathbb{E}[f(X)|S = s_i] = E(f(X)),\ s_i \in \{s_0, s_1\} \tag{13}$$

Let $h_v(X, S, Y, f(X)) = f(X)$ for all v, $\varphi_S = \{S = s_i\}$, and $\varphi_{\{X\backslash S\}} = \{True\}$ the equality constraints mentioned above can be represented as $\eta_S(f) = \eta_{\{X\backslash S\}}(f)$. Each equality constraint can be formulated as a pair of positive ($\Delta^+ := \eta_S(f) - \eta_{\{X\backslash S\}}(f) \leq 0$) and negative ($\Delta^- := -\eta_S(f) + \eta_{\{X\backslash S\}}(f) \leq 0$) inequality constraints. DP can be expressed as Equation (11), where $\mathcal{Z} = |\mathcal{S}| \times number\ of\ inequality\ constraints$. The elements of Q are initialized to form a set of linear constraints:

$$Q_{(s,\Delta^+),s'} = \begin{cases} 1 & \text{if } s' = s \\ -1 & \text{otherwise} \end{cases},\ Q_{(s,\Delta^-),s'} = \begin{cases} -1 & \text{if } s' = s \\ 1 & \text{otherwise} \end{cases}$$

While computing the fairness loss, we emphasize more on larger errors by taking the L2-norm (Gradshteyn and Ryzhik 2000) of the constraint as follows:

$$F(\psi) = \zeta * ||(ReLU(Q\eta(f))) - e||_2, \tag{14}$$

A similar fairness constraint derivation for the metric $FACE$ is provided in the Appendix. Moreover, the algorithmic detail of the fairness constrained learning process on the client side is also provided in the Appendix.

Server Side: Multiobjective Bayesian Optimization

Server-side fine-tunes the constraint parameter (ζ) and learning rate (lr) through MOO based on Differentiable Expected q-Hypervolume Improvement ($qEHVI$) (Daulton 2020) approach that is exact upto the Monte Carlo (MC) integration error (Robert et al. 1999). This approach outperforms SoTA MOO methods at a fraction of their wall time. Algorithm 1 details this module of FairTrade.

Algorithm 1: FairTrade server side algorithm

Require: Optimization rounds (n_o), Communication rounds (n_c), Initial learning rate (lr'), Initial fairness constraint regularization parameter ζ'.

Ensure: Optimized parameters ψ_g^{l+1} w.r.t. $disc\_score_g$ and BA_g

1: $\psi_g^1 = init()$
2: $global\_model.initialize(\psi_g^l, lr')$
3: **for** $round = 1$ to n_c **do**
4: $\quad \psi_g^{l+1} = sec\_agg\left(\{\psi_k^l\}_{k=1}^N\right)$
5: $\quad disc\_score_g,\ BA_g = sec\_agg\left(\{client\_metrics(k, \psi_g^{l+1})\}_{k=1}^N\right)$
6: $\quad y = \{-disc\_score_g, BA_g\}$ ▷ initial objectives
7: $\quad U = \{lr', \zeta'\}$ ▷ initial inputs
8: $\quad GP.initialize(U, y)$
9: $\quad$ **for** $i = 1$ to n_o **do**
10: $\quad\quad \alpha_{qEHVI}.init(GP, U, y)$
11: $\quad\quad U_{new}[lr_i, \zeta_i] = SAA\_optimize(\alpha_{qEHVI})$
12: $\quad\quad y_{new} = sec\_agg\{client\_metrics(k, \psi_g^{l+1}, lr_i, \zeta_i)\}_{k=1}^N$
13: $\quad\quad U = U \cup U_{new},\ y = y \cup y_{new}$
14: $\quad\quad GP.update(U, y)$
15: $\quad$ **end for**
16: $\quad lr' = lr_{n_o}$ and $\zeta' = \zeta_{n_o}$
17: $\quad send\_global\_updates(\psi_g^{l+1}, lr', \zeta')$
18: **end for**

The algorithm initiates with the Kaiming Uniform initialization (He et al. 2015) of global model parameters followed by the initialization of global model with these parameters (Algorithm 1: lines 1 to 2). In every communication round, the algorithm computes the new global model parameters ψ_g^{l+1}, global discrimination score ($disc\_score_g$: see Equation (6)), and global balanced accuracy (BA_g: see Equation (7)) through secure multi party aggregation and averaging (Knott et al. 2021) of local models' parameters ψ_k^l, local $disc\_score_k$, and local balanced accuracy values (BA_k) from all the clients (Algorithm 1: lines 3 to 5). Next, we initialize a list of GP surrogate models with initial objectives ($-disc\_score_g$, BA_g) and inputs (lr, ζ) (Algorithm 1: lines 6 to 8). We aim to maximize BA and minimize $disc\_score$. However, the Multi-objective Bayesian Optimization (MOBO) method employed here aims at maximizing the conflicting objectives. To fit into this maximization framework, we consider the negative of the discrimination score as our objective. After initializing two GP models for the two objectives, we utilize MOBO to find Pareto optimal trade-offs between BA and $disc\_score$. MOBO initiates with the initialization of the acquisition function (α_{qEHVI}) using the surrogate models (GP), initial inputs, and objectives (Algorithm 1: lines 9 to 10). After computing the ac-

quisition function, we optimize it using the Sample Average Approximation (SAA) method (Balandat et al. 2020) to compute new candidate inputs (U_{new}) (Algorithm 1: line 11). This optimization leverages auto-differentiation to calculate the precise gradient of the MC estimator of $qEHVI$, ensuring faster convergence rates. The new inputs ($U_{new} = lr_i, \zeta_i$) are sent to all the clients and new objectives (y_{new}) are computed through secure aggregation and averaging of BA_k and $-disc\_score_k$ from respective clients (Algorithm 1: line 12). The surrogate GP models are updated to include the new objectives and inputs and the next MOBO round starts (Algorithm 1: lines 13 to 14). At the end of MOBO rounds, the global updates including the global model parameters, learning rate and ζ are sent to all the clients for the next communication round (Algorithm 1: lines 17 to 18).

Having detailed the server-side algorithm, the rest of this section elucidates the underlying mathematical framework that guides the trade-offs between balanced accuracy and fairness in our optimization strategy. Essentially, we demonstrate how the acquisition function (α_{qEHVI}) is defined for MOBO and how it can be computed efficiently.

The Pareto front represents the set of optimal trade-offs between the two objectives (BA, $disc\_score$): each point on the Pareto front signifies a unique balance between the balanced accuracy and fairness. Some points may have high fairness but lower balanced accuracy, and others may have high balanced accuracy but lower fairness.

Hypervolume (HV) is a metric that quantifies the coverage of the "fairness-balanced accuracy" space by the Pareto front with the aim to maximize this coverage. HV is calculated by measuring the volume of the region in our dual-objective space—balanced accuracy and fairness—that is dominated by the Pareto front ($\mathcal{P}^*$), with the reference point $r = (BA_{min}, -disc\_score_{max})$ as the lower bound:

$$HV(\mathcal{P}^*, r) = \lambda_N(\cup_{j=1}^{|\mathcal{P}^*|}[r, y_j]). \tag{15}$$

HV is the N-dimensional Lebesgue measure $\lambda_N(\cdot)$ (Bartle 2014) of the region dominated by the Pareto front. $[r, y_j]$ denotes the hyper rectangle which is bounded by vertices y_j and r, while y_j is the j^{th} solution in the Pareto set.

Hypervolume Improvement (HVI) is the difference in HV before and after a new set of candidate solutions ($\mathcal{Y} : \{y_1, ..., y_q\}$) is considered as shown in Equation (16). For our case the new set of candidate solutions corresponds to potential solutions offering varying trade-offs between BA and ($-disc\_score$). HVI indicates the enhanced trade-off between fairness and balanced accuracy that the new set of solutions provides.

$$HVI(\mathcal{Y}, \mathcal{P}^*, r) = HV(\mathcal{P}^* \cup \mathcal{Y}, r) - HV(\mathcal{P}^*, r) \tag{16}$$

The non-rectangular shape of the region $\mathcal{P}^* \cup \mathcal{Y}$ necessitates its division into hyper-rectangles to calculate HVI.

Expected Hypervolume Improvement ($EHVI$) serves as the acquisition function for MOBO specifically tailored for our dual objectives: balanced accuracy and fairness. $EHVI$ guides the selection of solutions offering potential trade-offs between BA and $-disc\_score$. It is quantified as the expectation of HVI computed by the posterior distribution ($\theta(U) = \mathbb{P}(f|D)$) provided by the surrogate models (GP):

$$\alpha_{qEHVI}(U) = \mathbb{E}[HVI(\theta(U))] = \int_{-\infty}^{+\infty} HVI(\theta(U))d\theta, \tag{17}$$

where U is the candidate input set, each member of which signifies a potential trade-off between the BA and $-disc\_score$ in the objective space and q denotes the number of candidate points considered. For our case, $U := \{lr, \zeta\}$ where lr is the learning rate and ζ is the regularization parameter of fairness constraints as discussed in the above section . The integral sign in the Equation (17) denotes the expectation operation, computing the average HVI over all possible outcomes of $\theta(U)$. The limit of the integral depends on the range of the two objectives i.e., $[0, 1]$ for BA and $[-1, 1]$ for $-disc\_score$.

Next, we explain how to compute this integral through MC sampling. The high level idea is to sum the HV of all the partitions of the non-dominated objective space. Each partition $\{A\}_{w=1}^{W}$ is a disjoint hyper-rectangle bounded by a lower bound vertex $l_w \in \mathbb{R}^N$ and an upper bound vertex $t_w \in \mathbb{R}^N \cup \{\infty\}$. The acquisition function α_{qEHVI} in Equation (17) is estimated using MC integration given the posterior distribution $\{\theta_r(u_j)\}_{j=1}^{q} \sim \mathbb{P}(\theta(u_1), ..., \theta(u_q)|D), r = 1, ..., M$ obtained by the surrogate GP model. Equation (18) depicts the formulation for computation of acquisition function of $qEHVI$, where $z^n_{w,\mathcal{U}_i,r} := min[t_w, min_{u' \in \mathcal{U}_j}\theta_r(u')]$, $\mathcal{U}_i \subset U$, W is the number of hyper-rectangles, and M is the number of MC samples (the number of samples drawn from $\theta_r(U)$).

$$\begin{aligned} \alpha^M_{qEHVI}(U) &= \frac{1}{M}\sum_{r=1}^{M} HVI(\theta_r(U)) \\ &= \frac{1}{M}\sum_{r=1}^{M}\sum_{w=1}^{W}\sum_{i=1}^{q}\sum_{U_i \in \mathcal{U}_i}(-1)^{i+1}\prod_{n=1}^{N}[z^{(n)}_{r,U_i,r} - l^w_{(n)}]_+ \end{aligned} \tag{18}$$

Comprehensive insights on this MOBO approach are available in the original paper and the attached Appendix.

Experimental Setup

Benchmark Datasets

We evaluate FairTrade with five real-world datasets: (1) Bank (Bache and Lichman 2013), (2) Default (Bache and Lichman 2013), (3) Adult (Bache and Lichman 2013), (4) Law (Wightman 1998), and (5) KDD (Bache and Lichman 2013). The datasets vary in their number of attributes, number of instances, sensitive attribute and class imbalance ratio. Specifically, the positive to negative class ratios for Adult, KDD, Bank, Default, and Law are $1:3.0$, $1:15.11$, $1:7.87$, $1:3.52$, $1:3.50$ respectively. Further details about the datasets are available in the Appendix. To mimic FL setup, each dataset is distributed among a specified set of clients - (i) randomly or (ii) based on specific attributes (age for Bank, Default, Adult, and KDD; income for Law) to mirror more realistic scenarios. Note that the baseline methods also deploy some of these datasets but ignore the class imbalance and report accuracy as the measure of performance.

Dataset	FedAvg			FF-SMOTE			Agnostic-Fair			FCFL			FedFB			FairFed			FairTrade		
	Acc	BA	DP	Acc	BA	DP	Acc	BA	DP	Acc	BA	DP	Acc	BA	DP	Acc	BA	DP	Acc	BA	DP
Adult	0.77	0.79	0.409	0.82	0.67	0.07	0.77	0.58	0.067	0.80	0.62	-0.092	0.76	0.50	0.0005	0.76	0.50	0.0008	0.76	0.77	0.001
Bank	0.87	0.79	0.076	0.87	0.51	-0.0006	0.86	0.53	0.064	0.88	0.55	-0.003	0.89	0.56	0.0006	0.88	0.56	0.030	0.88	0.78	0.019
Default	0.78	0.69	0.053	0.80	0.63	0.043	0.78	0.52	0.006	0.80	0.57	0.089	0.80	0.58	0.008	0.79	0.58	0.026	0.77	0.69	-0.010
Law	0.88	0.72	0.037	0.90	0.61	0.051	0.93	0.50	0.001	0.91	0.61	-0.029	0.91	0.57	0.004	0.91	0.59	0.021	0.88	0.68	-0.008
KDD	0.94	0.59	0.012	0.91	0.48	0.055	0.93	0.50	0	0.93	0.52	0.003	0.93	0.50	0	0.93	0.50	0	0.87	0.82	0.002

Table 1: Acc, BA, and DP achieved by FairTrade and competitors across all datasets with R3C data split.

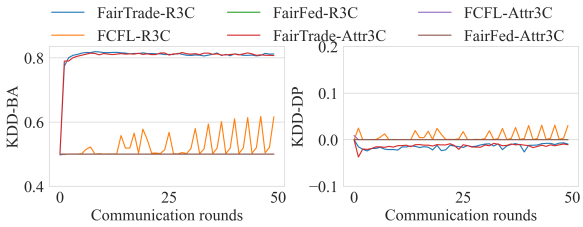

Figure 2: Comparison between BA and DP values achieved by FairTrade, FCFL, and FairFed for KDD dataset with R3C and Attr3C data split over different communication rounds.

Dataset	R3C			Attr3C		
	BA	DP	FACE	BA	DP	FACE
Adult	0.7945	0.0913	-0.0094	0.744	0.0636	0.0035
Bank	0.806	0.0314	0.009	0.8055	0.0167	0.0117
Default	0.6981	0.0227	0.0117	0.68	-0.008	0.0138
Law	0.7467	0.0333	0.01	0.7266	-0.0033	-0.0194
KDD	0.8196	0.0558	0.011	0.796	0.0408	0.0024

Table 2: Causal fairness (FACE) and BA achieved by FairTrade for all datasets with R3C and Attr3C data splits.

Benchmark Baselines

In this section, we introduce the baseline methods that we employ to compare the performance of FairTrade. For further details, see *section: "Related Work"*.

FedAvg (Konečnỳ et al. 2016) is the conventional FL framework without fairness interventions.

FF-SMOTE debiases predictions locally at each client using Fair-SMOTE (Chakraborty and Majumder 2021).

Agnostic-Fair (Du et al. 2021) removes discrimination by adding regularization terms to reweight the training samples.

FCFL (Cui 2021) is a gradient-based approach that provides consistent Pareto utility (accuracy and fairness) distribution across all clients.

FairFed (Ezzeldin et al. 2023) is a discrimination-aware weights aggregation method in an FL setup.

FedFB (Zeng, Chen, and Lee 2021) debiases predictions locally at each client using Fair-Batch. (Roh et al. 2021).

FairTrade is the proposed MOO based FL framework.

Experimental Evaluation and Discussion

The evaluation of FairTrade is performed across three different facets - (i) comparison with existing baselines across different datasets, (ii) generalizability across metrics and application scenarios and (iii) sensitivity to hyperparameters.

Comparison with Benchmark Baselines:

We compare FairTrade with six baseline methods across five datasets. For fair comparison, we follow the experimental setup proposed in (Ezzeldin et al. 2023). We consider demographic parity (DP) as the fairness metric and randomly distribute the data among 3 clients. The results in Table 1 show that FairTrade reports the best trade-off between balanced accuracy (BA) and DP compared to the baselines.

FairTrade's BA marginally declines compared to that obtained by FedAvg (FL without fairness interventions) for some datasets. However, this slight degradation in BA is a calculated trade-off, which allows for a much-needed reduction in $disc\_score$. While some baselines may offer slightly better Accuracy (Acc) on some datasets, they often lag in BA which is crucial for skewed datasets. Notably, fairness-aware SoTA FL methods, Agnostic-Fair, FCFL, FedFB, and FairFed show high Acc with the corresponding BA values nearing $\sim$ 0.5 and DP scores close to 0. Such low BA values categorize these FL models as random classifiers, rendering their low DP scores insignificant. In contrast, FairTrade achieved significantly lower DP with remarkably higher BA for all datasets. Figure 2 shows a comparison between BA and DP values achieved by FairTrade, FCFL, and FairFed across KDD dataset over 50 communication rounds. The figure attests to the efficacy of FairTrade in achieving fairness without compromising the model's BA.

Generalizability

We now demonstrate FairTrade's generalizability by incorporating a different fairness notion and deploying it in settings with attribute-based client data distributions.

Causal fairness notions: Table 2 shows FairTrade's BA, DP, and FACE results for all datasets with random and attribute-based distribution among 3 clients. The table shows that FairTrade consistently maintains high BA values while achieving low FACE scores across all datasets. This suggests that FairTrade is agnostic with respect to the chosen fairness metric. Furthermore, the table highlights an intriguing observation: even as we optimize FairTrade to minimize FACE, the resulting DP values remain notably low. This indicates that DP and FACE are not conflicting fairness notions.

Attribute-based Data Distribution: In real-world FL environments, data may often be partitioned based on inherent attributes rather than being distributed randomly across clients. For example, in a hospital network, each healthcare facility collects patients' data based on local demo-

Split	FedAvg		FF-SMOTE		Agnostic-Fair		FCFL		FedFB		FairFed		FairTrade	
	BA	DP	BA	DP	BA	DP	BA	DP	BA	DP	BA	DP	BA	DP
Attr3C	0.744	0.2686	0.6	-0.013	0.5	0	0.507	0.0422	0.499	0.0006	0.5	0.0006	0.748	0.0054
R3C	0.794	0.4096	0.675	0.07	0.58	0.067	0.626	-0.0928	0.501	0.0005	0.5006	0.0008	0.776	0.0012
R5C	0.788	0.4349	0.57	0.021	0.59	0.047	0.545	-0.0056	0.500	0.0008	0.501	0.0008	0.779	0.0167
R10C	0.785	0.4284	0.621	0.0389	0.58	0.061	0.541	-0.0026	0.501	0.0009	0.502	0.0007	0.781	0.0141
R15C	0.780	0.4122	0.622	0.106	0.58	0.08	0.576	-0.06	0.501	0.0011	0.502	0.0006	0.775	-0.0044

Table 3: BA and DP obtained by FairTrade and competitors for the Adult dataset with random and attribute-based distribution. Note that RnC/Attr**n**C denotes random/attribute-based distribution of data among n clients.

graphics rather than a random assortment of patients from across the region. To replicate such scenarios, we employ an attribute-based splitting of data among the clients. The details about data splits can be found in the section "Benchmark Datasets". Table 3 (row: Attr3C) provides a detailed comparison of FairTrade's performance with other baselines for Adult dataset with attribute based data distribution among 3 clients. The BA values of competing baselines degrade with Attr3C distribution compared to what they achieved with random distributions. FairTrade efficiently manages such distribution complexities, ensuring fairness and high predictive performance in real-world scenarios.

Sensitivity

To gain a deeper understanding of our proposed method, we also perform a set of sensitivity experiments.

Number of clients: We systematically vary the number of clients to investigate the robustness of FairTrade under different client settings. The performance of FairTrade remains remarkably consistent across the variations, as presented in Table 3. Competing fairness-aware FL methods show low DP along with very low BA values, indicating that they achieve fairness at the expense of predictive performance. While FCFL reports the best BA among the baselines, its performance drops with more clients, showing its limited adaptability. In contrast, the consistent superior predictive and fairness performance of FairTrade proves its adaptability and effectiveness regardless of the client count.

Number of Communication rounds: We also examine the performance of FairTrade by varying the number of communication rounds. Figure 2 presents the BA and DP values achieved by FairTrade over 50 communication rounds across KDD dataset with R3C and Attr3C data distribution. The figure shows that FairTrade achieves an optimal trade-off between BA and DP within the initial 10 rounds and maintains this, whereas competitors consistently post a BA near 50% (akin to a random classifier) with DP close to 0 (similar trend is observed for other datasets). These observations further highlight FairTrade's efficiency in achieving and sustaining optimal trade-offs between BA and DP.

Related Work

Fairness-aware Learning: In recent years, ML methods that aim at detecting and then mitigating bias have received a great deal of attention (refer to (Mehrabi et al. 2021) for a detailed survey). These techniques can be broadly classified into three categories: Pre-processing, In-processing and Post-processing. **Pre-processing methods** tailor the training data to make it bias-free before feeding it to the learner, such as (Iosifidis and Ntoutsi 2018). Fair-SMOTE (Chakraborty and Majumder 2021) is another SoTA method which deals with discrimination. **In-processing methods** adapt the classification model itself to generate fair outcomes. Some examples include adaptation of optimization objective (Padala and Gujar 2020) and adaptive reweighting (Iosifidis and et al 2019). **Post-processing methods** modify the classifier decisions to mitigate bias, such as (Hajian and et al 2015).

Fairness-aware Federated Learning: Recently, there have been a few attempts at developing bias mitigation methods in an FL setting. Agnostic-Fair (Du et al. 2021) is an adaptive instance re-weighting based fairness aware FL framework. FedFB (Zeng, Chen, and Lee 2021) is another FL framework where each client locally debiases its predictions using Fair-Batch (Roh et al. 2021). In (Ezzeldin et al. 2023) the authors introduce a discrimination-aware weights aggregation method in an FL setup (FairFed) where the local fairness of each client dictates the weight assigned to its contribution in the global parameter aggregation, subsequently promoting global fairness. A gradient-based approach, FCFL, is presented by (Cui 2021) that provides consistent Pareto utility (accuracy and fairness) distribution across all clients. (Badar, Nejdl, and Fisichella 2023) is another data augmentation based method for fairness aware FL in a streaming environment. These methods overlook the key issue of class imbalance in discrimination-aware FL.

Conclusion

We proposed a novel Federated learning (FL) framework, FairTrade, aimed at achieving a Pareto-optimal trade-off between balanced accuracy and fairness in FL applications. Our methodology, employing multi-objective optimization, presents a major leap forward from traditional SoTA frameworks that primarily focus on accuracy. Merely focusing on accuracy creates a misleading impression about the classifier's ability to mitigate discrimination. The efficacy of FairTrade is further demonstrated through experiments across several benchmark datasets and fair FL methods, with FairTrade consistently achieving better fairness-balanced accuracy trade-off. It is agnostic to the fairness metric and effectively generalizes to diverse client data distributions and varying numbers of clients.

Acknowledgements

This research was partially funded by the European Commission through the xAIM project, agreement No. INEA/CEF/ICT/A2020/2276680.

References

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. In *International conference on machine learning*, 60–69. PMLR.

Bache, K.; and Lichman, M. 2013. UCI ML repository.

Badar, M.; Nejdl, W.; and Fisichella, M. 2023. FAC-fed: Federated adaptation for fairness and concept drift aware stream classification. *Machine Learning*, 112: 2761–2786.

Balandat, M.; Karrer, B.; Jiang, D.; Daulton, S.; Letham, B.; Wilson, A. G.; and Bakshy, E. 2020. BoTorch: A framework for efficient Monte-Carlo Bayesian optimization. *Advances in neural information processing systems*, 33: 21524–21538.

Bartle, R. G. 2014. *The elements of integration and Lebesgue measure*. John Wiley & Sons.

Chakraborty, J.; and Majumder, S. 2021. Bias in machine learning software: why? how? what to do? In *29th ESEC/FSE*, 429–440.

Cui, S. 2021. Addressing algorithmic disparity and performance inconsistency in federated learning. *NeurIPS*.

Daulton, S. 2020. Differentiable expected hypervolume improvement for parallel MOBO. *NeurIPS*.

Du, W.; Xu, D.; Wu, X.; and Tong, H. 2021. Fairness-aware agnostic federated learning. In *SDM*, 181–189. SIAM.

Dullerud, N.; Roth, K.; Hamidieh, K.; Papernot, N.; and Ghassemi, M. 2021. Is Fairness Only Metric Deep? Evaluating and Addressing Subgroup Gaps in Deep Metric Learning. In *International Conference on Learning Representations*.

Emelianov, V.; Gast, N.; and Gummadi, K. P. 2022. On fair selection in the presence of implicit and differential variance. *Artificial Intelligence*, 302: 103609.

Ezzeldin, Y. H.; Yan, S.; He, C.; Ferrara, E.; and Avestimehr, A. S. 2023. Fairfed: Enabling group fairness in federated learning. In *AAAI*, volume 37, 7494–7502.

Fisichella, M.; Lax, G.; and Russo, A. 2022. Partially-federated learning: A new approach to achieving privacy and effectiveness. *Inf. Sci.*, 614: 534–547.

Gradshteyn, I.; and Ryzhik, I. 2000. Table of Integrals, Series, and Products 6th edn (New York: Academic).

Hajian, S.; and et al. 2015. Discrimination-and privacy-aware patterns. *Data Mining and Knowledge Discovery*, 29(6): 1733–1782.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, 1026–1034.

Iosifidis, V.; and et al. 2019. Adafair: Cumulative fairness adaptive boosting. In *CIKM*.

Iosifidis, V.; and Ntoutsi, E. 2018. Dealing with bias via data augmentation in supervised learning scenarios. *Jo Bates Paul D. Clough Robert Jäschke*, 24.

Jones, D. R.; Schonlau, M.; and Welch, W. J. 1998. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13: 455–492.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. 2021. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2): 1–210.

Khademi, A.; Lee, S.; Foley, D.; and Honavar, V. 2019. Fairness in algorithmic decision making: An excursion through the lens of causality. In *The World Wide Web Conference*, 2907–2914.

Knott, B.; Venkataraman, S.; Hannun, A.; Sengupta, S.; Ibrahim, M.; and van der Maaten, L. 2021. Crypten: Secure multi-party computation meets machine learning. *Advances in Neural Information Processing Systems*, 34: 4961–4973.

Konečnỳ, J.; McMahan, H. B.; Yu, F. X.; Richtárik, P.; Suresh, A. T.; and Bacon, D. 2016. Federated learning: Strategies for improving communication efficiency. *NIPS Workshop on Private Multi-Party Machine Learning*.

Makhlouf, K.; Zhioua, S.; and Palamidessi, C. 2020. Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553*.

McMahan, B.; and et al. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Padala, M.; and Gujar, S. 2020. FNNC: achieving fairness through neural networks. In *IJCAI*.

Ramaswamy, S.; Mathews, R.; Rao, K.; and Beaufays, F. 2019. Federated learning for emoji prediction in a mobile keyboard. *arXiv preprint arXiv:1906.04329*.

Rasmussen, C. E. 2003. Gaussian processes in machine learning. In *Summer school on machine learning*, 63–71. Springer.

Robert, C. P.; Casella, G.; Robert, C. P.; and Casella, G. 1999. Monte carlo integration. *Monte Carlo statistical methods*, 71–138.

Roh, Y.; Lee, K.; Whang, S. E.; and Suh, C. 2021. Fairbatch: Batch selection for model fairness. In *ICLR*.

Stuart, E. A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1): 1.

Verma, S.; and Rubin, J. 2018. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, 1–7. IEEE.

Wang, L.; Xu, S.; Wang, X.; and Zhu, Q. 2021. Addressing class imbalance in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 10165–10173.

Wightman, L. F. 1998. LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series. *ERIC*.

Younis, R.; and Fisichella, M. 2022. FLY-SMOTE: Re-Balancing the Non-IID IoT Edge Devices Data in Federated Learning System. *IEEE Access*, 10: 65092–65102.

Zeng, Y.; Chen, H.; and Lee, K. 2021. Improving fairness via federated learning. *arXiv preprint arXiv:2110.15545*.