ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240762

TrustFed: Navigating Trade-offs Between Performance, Fairness, and Privacy in Federated Learning

Maryam Badar[a,*], **Sandipan Sikdar**[a,**], **Wolfgang Nejdl**[a] **and Marco Fisichella**[a]

[a]L3S Research Center, Leibniz University Hannover, Germany

Abstract. As Federated Learning (FL) gains prominence in secure machine learning applications, achieving trustworthy predictions without compromising predictive performance becomes paramount. While Differential Privacy (DP) is extensively used for its effective privacy protection, yet its application as a lossy protection method can lower the predictive performance of the machine learning model. Also, the data being gathered from distributed clients in an FL environment often leads to class imbalance making traditional accuracy measure less reflective of the true performance of prediction model. In this context, we introduce a fairness-aware FL framework (TrustFed) based on Gaussian differential privacy and Multi-Objective Optimization (MOO), which effectively protects privacy while providing fair and accurate predictions. To the best of our knowledge, this is the first attempt towards achieving Pareto-optimal trade-offs between balanced accuracy and fairness in a federated environment while safeguarding the privacy of individual clients. The framework's flexible design adeptly accommodates both statistical parity and equal opportunity fairness notions, ensuring its applicability in various FL scenarios. We demonstrate our framework's effectiveness through comprehensive experiments on five real-world datasets. TrustFed consistently achieves comparable performance fairness tradeoff to the state-of-the-art (SoTA) baseline models while preserving the anonymization rights of users in FL applications.

1 Introduction

In the realm of machine learning, Federated Learning (FL) has emerged as a revolutionary paradigm that enables collaborative model training [26]. While strides have been made in the domain of FL, the area of mitigating discrimination in the outcomes of an FL system is still underexplored. Recently, few methods [29, 15, 12, 19, 41] have been proposed to address the issue of fairness in FL. However, these fairness aware FL frameworks ignore the key challenges of FL, particularly the assurance of privacy and the equitable distribution of model performance [18].

Privacy preservation is a critical issue in FL, especially when considering the susceptibility of the system to Membership Inference Attacks (MIA), where an adversary may deduce the presence of individual data points in the training dataset. Differential Privacy (DP) stands out as an advanced solution in this regard, outperforming traditional encryption methods. Unlike encryption, which only secures data at rest or in transit, DP provides a quantifiable privacy measure that actively protects the information during the learning process [17]. The integration of DP into FL ensures that the inclusion or exclusion of any individual data point has minimal impact on the global model's output.

Another critical facet of fairness aware FL is the concept of class imbalance. In a distributed setting, data can be inherently imbalanced, leading to skewed performance as the global model tends to favor majority classes [38]. As an illustrative example, consider a dataset that has 70% positive samples and 30% negative samples. Each sample (x) has a sensitive attribute $S \in M, F$. Consider an FL model which always predicts a positive class i.e. $f(x) = 1$, for such an FL model, accuracy is 0.70 and discrimination measured in terms of the difference in probability of being assigned the positive class (aka statistical parity) is 0 as f always predicts 1 irrespective of the sensitive attribute. The low discrimination score achieved here is not at all reflective of the model's discrimination mitigation capability. In fact, 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. While the **existing methods often report high accuracy, our experiments show that the balanced accuracy achieved by them are often low, rendering the achieved fairness score pointless**.

Given these considerations, it is essential to navigate the trade-off between the conflicting objectives: privacy budget provided by DP, fairness, and the equitable performance measured by balanced accuracy. In this context, we propose TrustFed, which utilizes DP along with Multi-Objective Optimization (MOO) to jointly maximize balanced accuracy and minimize discrimination while taking care of the privacy rights of individuals. Extensive experiments on real world datasets show that TrustFed achieves the best performance-fairness trade-off along with privacy guarantees. We also find our method to be more efficient, achieving convergence in fewer communication rounds compared to the baseline methods. It also seamlessly adapts to multiple notions of fairness (e.g., statistical parity and equal opportunity) demonstrating its generalizability.

Key Contributions Central to this work, we delineate the following contributions:

- While existing methods have focused on optimizing privacy, fairness, and accuracy in a federated setting [11], they overlook a critical facet of fairness-aware FL, i.e., class imbalance. The primary contribution of this work lies in the novel formulation of the FL challenge —simultaneously enhancing fairness and balanced accuracy without breaching privacy rights of individuals.
- Catering to the unique complexities of each client's non-Independent and Identically Distributed (non-IID) data distribution we innovatively adapt Multi-objective Bayesian Optimiza-

* Corresponding Author. Email: badar@l3s.de
** Corresponding Author. Email: sandipan.sikdar@l3s.de

tion (MOBO) to identify Pareto optimal trade-offs between balanced accuracy and fairness through a sophisticated consideration of both local and global fairness and balanced accuracy.
- The strategic integration of Differential Privacy (DP) noise makes our framework less susceptible to Membership Inference Attacks (MIA)[17] further safeguarding the data involved.
- The superiority of our method is demonstrated through rigorous experiments involving five benchmark datasets and comparison with several state-of-the-art (SoTA) baseline methods.

For reproducibility, all resources associated with our research, including code and data, are available at the provided repository link [1].

2 Preliminaries

We begin by providing a summary of the traditional FL framework (FedAvg) as per [30], subsequently, we delineate key concepts central to TrustFed.

2.1 Federated Learning Setup

Consider an FL setting with *n* individual clients $(C_1, C_2, ...C_n)$ and a central global server *G*. Each client possesses its distinct local dataset D_k, characterized by a feature space *X* and an outcome space *Y*. We address a binary classification task, where $Y \in \{0, 1\}$. The dataset D_k of each client C_k contains m_k samples, and each sample is represented as $I_j^k = \{x_j, y_j\},\ j\ \epsilon\ [1, m_k]$. The global server *G* develops a predictive model $f(x) = y$ through the collaborative training of the local clients $(C_1, C_2, ...C_n)$. It does this by aggregating and averaging the updates from each local model, with the weights being proportional to the size of the client's dataset. Specifically, the objective is to determine a set of parameters (ψ) that minimizes the combined average loss across all clients, as detailed in Equation (1).

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

Once the training is complete, the global model parameters are shared with the individual clients. FedAvg ensures scalability and performance, yet its predictions can exhibit demographic biases in datasets along with privacy risks.

While there exists several notions of fairness, in this paper we consider two, namely (i) statistical parity and (ii) equal opportunity. Note that our method could also be deployed with other fairness notions.

2.2 Fairness Notions

Discrimination involves biased or unjust treatment of individuals or groups based on certain traits like race or gender, known as sensitive attributes. We assume that the datasets used in this work have a single sensitive attribute S (e.g., "gender"), which is binary: s_0 representing a protected group (like "female") and s_1 a non-protected group (like "male").

Statistical Parity (Stp): Essentially Stp represents the difference in mean positive outcomes of protected and non-protected group:

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

$Stp = 0$ denotes a perfectly fair classifier, whereas $Stp = 1$ or -1 signifies complete unfairness.

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

Equal Opportunity (Eqop): Eqop focuses on the disparity in true positive rates between the protected and non-protected groups:

$$\begin{aligned} Eqop = P(f(x) = y^+ \mid S = s_1, Y = y^+) \\ -P(f(x) = y^+ \mid S = s_0, Y = y^+). \end{aligned} \tag{3}$$

$Eqop = 0$ signifies a perfectly fair classifier, whereas $Eqop = 1$ or -1 implies complete unfairness.

2.3 Fairness and Balanced Accuracy for FL

The fairness measures specified in Equations (2) and (3) are directly applicable to a centralized setting. However, in FL, due to the non-Independent and Identically Distributed (non-IID) nature of the data across clients, it becomes crucial to distinguish between *client-side fairness* and *server-side fairness*.

The concept of client-side fairness can be understood through minor modifications to Equations (2) and (3). For example, Stp for client k with local dataset D_k can be defined as:

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

Server-side fairness $(disc_g)$ takes into account the entire dataset $D_g = \bigcup_{k \in K} D_k$. When client data is IID, client-side and server-side fairness become identical. The server-side Stp for a classifier $f(x)$ can be specified as:

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

The primary challenge lies in computing server-side fairness without access to client data stores. Server-side fairness can be computed through the aggregation of client-side fairness measures [28]. If Stp is the fairness metric then server-side fairness can be quantified as:

$$disc_g = \sum_{k=1}^{K} w_k\ disc_k. \tag{6}$$

w_k denotes the fraction of data points at client k relative to the total number of data points across all clients i.e., $w_k = \frac{|D_k|}{\sum_j |D_j|}$. The server-side balanced accuracy (BA_g) is computed as:

$$BA_g = \sum_{k=1}^{K} w_k BA_k, \tag{7}$$

2.4 Differential Privacy (DP)

Differential Privacy (DP) provides a robust alternative to encryption for protecting against Membership Inference Attacks (MIA) in FL environments [17]. By introducing noise into data or gradients before uploading them to global server, DP conceals individual data contributions. This not only makes DP effective against personal data inferences in machine learning models but also ensures privacy during data analysis and usage, a gap in protection that encryption alone doesn't address [39].

Definition 1 (**(ε, δ)-DP**): A randomized mechanism $A : X \to R$, with domain X and range R, satisfies (ε, δ)-DP if, for any two adjacent databases $D_j, D_j' \in X$, and for all measurable sets $B \subseteq R$, the probability $\Pr[A(D_j) \in B]$ is bounded by $e^{\varepsilon}\Pr[A(D_j') \in B] + \delta$:

$$\Pr[A(D_j) \in B] \leq e^{\varepsilon}\Pr[A(D_j') \in B] + \delta. \tag{8}$$

To ensure (ε, δ)-DP for numerical data, we utilize the Gaussian mechanism as defined in [16]. This involves the addition of artificial Gaussian noise. To guarantee (ε, δ)-DP with a noise distribution $n \sim \mathcal{N}(0, \sigma^2)$, where $\mathcal{N}$ represents the Gaussian distribution, the noise scale σ should satisfy $\sigma \geq \frac{c\Delta b}{\varepsilon}$, with the constant $c \geq \sqrt{2\ln(1.25/\delta)}$. In this formulation, n represents the additive noise value for data in the dataset, Δb is the sensitivity of the function b, given by $\Delta b = \max_{D_j, D'_j} |b(D_j) - b(D'_j)|$, where b is a real-valued function. **The privacy level is determined by the parameter ε, where a lower ε signifies greater privacy and vice versa.**
In fairness-aware FL, the addition of DP to local learner parameters and local metrics (BA_k, $disc_k$) before sharing them to global server intensifies the difficulty in achieving an optimal trade-off between balanced accuracy and fairness. We effectively address these issues by incorporating Multi-Objective Optimization (MOO) with Bayesian Optimization (BO). While MOO aims to simultaneously optimize for balanced accuracy and fairness, BO further sharpens this approach by probabilistically exploring the extensive solution space, thereby expediting the attainment of Pareto optimal trade-offs.

2.5 Multi-Objective Optimization (MOO)

In Multi-Objective Optimization (MOO), the aim is to optimize a vector-valued objective function $\theta(u) : M^d \rightarrow \mathbb{R}^N$, denoted as $\theta(u) = \{\theta^{(1)}(u), ..., \theta^{(N)}(u)\}$, within a bounded input space $U \subset \mathbb{R}^d$. The functions $\theta^{(j)}$ are complex, requiring intensive computation for evaluation as black-box functions. MOO seeks to identify Pareto optimal solutions, characterized by trade-offs between objectives, aiming to maximize all objectives simultaneously.

A solution $\theta(u)$ is considered to dominate another solution $\theta(u')$, indicated as $\theta(u) \succ \theta(u')$, if $\theta^{(n)}(u) \geq \theta^{(n)}(u')$ for all $n = 1, ..., N$, and there exists at least one n such that $\theta^{(n)}(u) > \theta^{(n)}(u')$. The Pareto frontier, which represents optimal trade-offs, is comprised of solutions $\mathcal{P}^* = \{\theta(u) \text{ s.t. } \nexists\, u' \in U : \theta(u') > \theta(u)\}$ and corresponding inputs $U = \{u \in U \text{ s.t. } \theta(u) \in \mathcal{P}^*\}$. While the Pareto frontiers include an infinite number of points, the objective is to identify a finite, approximate frontier.

2.6 Bayesian Optimization (BO)

Bayesian optimization (BO), as delineated by [25], is a powerful technique for optimizing computationally intensive black-box functions. It utilizes a probabilistic surrogate model, commonly a Gaussian Process (GP) [33], along with observed data $D = \{(u_i, y_i) \mid i = 1, ..., m\}$, to create a posterior distribution $\mathbb{P}(f \mid D)$ over the actual function values f. An acquisition function $\alpha : U \rightarrow \mathbb{R}$, grounded in the GP surrogate model, assesses the utility of a set of prospective input candidates $U = \{u_i \mid i = 1, ..., q\}$ for evaluation on the actual function f. This approach of employing a surrogate-based acquisition function is computationally efficient than direct evaluations of the true function f.

3 TrustFed: Trustworthy Federated MOO framework

3.1 Conceptual Overview

Each client hosts a local data store, a DP noise addition mechanism, a discrimination detection module, and a fairness constrained optimization module. The server hosts a Noisy MOO module which is tasked with maximizing two potentially conflicting objectives: fairness and high balanced accuracy.

In each communication round, every client trains its local learner and tries to mitigate discrimination through fairness constrained optimization using the learning rate (lr) and fairness constraint regularization parameter (ζ). ζ controls the trade-off between the local predictive loss function and the fairness loss (for more detail see Section 3.2). It ensures that the local classifier adheres to certain fairness criteria (e.g., statistical parity, equal opportunity) while minimizing its loss. After n epochs, ε-DP noise is added in the local learner weights, local balanced accuracy (BA_k), and local discrimination score ($disc_k$). The privacy level is determined by the parameter ε, where a lower ε signifies greater privacy and vice versa. This additional step ensures that each client's data privacy is preserved, mitigating the risk of sensitive information being inferred during the aggregation process on the server side. At the end of each communication round, every client shares noisy parameters (learner weights, BA_k, $disc_k$), and corresponding lr, ζ with the global server. Global server aggregates and averages noisy local learner weights and computes both global balanced accuracy (BA_g) and global discrimination score ($disc_g$) based on noisy BA_k and noisy $disc_k$ aggregated from all the clients. The server then applies Noisy MOO for fine-tuning ζ and lr, to ensure Pareto optimal trade-offs between balanced accuracy and discrimination score. We fine-tune lr and ζ at the global server to ensure that their optimized values are effective across different local data distributions and fairness constraints. Essentially, through globally optimized lr and ζ we want to ensure Pareto optimal trade-offs between fairness and performance across all clients. The updated global learner weights alongside newly optimized lr' and ζ' are then shared with clients for use in the subsequent communication round. Detailed explanations follow in later sections.

3.2 Client Side: Fairness Constrained Optimization

In TrustFed we employ fairness-constrained optimization [2] at each client to achieve client-side fairness (as detailed in section 2).
Each client has its own dataset (D_k) with inherent demographic biases. The goal is to train a predictive model adhering to specific fairness constraints. The optimization problem at each client side for a local classifier f parameterized by ψ can be formulated as follows:

$$\underset{\psi}{\text{minimize}}\ J(\psi) + \zeta * F(\psi)\ s.t.\ g(\psi) \leq e. \tag{9}$$

Here, $J(\psi)$ is the local loss function, $F(\psi)$ is the fairness penalty, ζ is a regularization parameter (optimized by Noisy MOO at the server), $g(\psi)$ is the chosen fairness metric, and e is the fairness budget. For each fairness notion (g(ψ)), we can derive a set of linear constraints:

$$Q\eta(f) \leq e, \tag{10}$$

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:

$$\eta_v(f) = \mathbb{E}[h_v(X, S, Y, f(X)) | \varphi_v]\ for\ v \in \mathcal{V}. \tag{11}$$

Here $\mathcal{V} = \mathcal{S} \cup \{\mathcal{X} \backslash \mathcal{S}\}$, $h_v : \mathcal{X} \times \mathcal{S} \times \{0,1\} \times \{0,1\} \rightarrow [0,1]$ 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 statistical parity (Stp).

Algorithm 1 TrustFed 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}, lr', ζ') w.r.t. $disc_g$ 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: $\quad \psi_g^{l+1} = \overline{\{\psi_k^l + DPN(\varepsilon)\}_{k=1}^N}$
5: $\quad disc_g,\ BA_g = \{noisy\_client\_metrics(k, \psi_g^{l+1})\}_{k=1}^N$
6: $\quad y = \{-disc_g, BA_g\}$ ▷ 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: $\quad\quad U_{new}[lr_i, \zeta_i] = SAA\_optimize(\alpha_{qNEHVI})$
12: $\quad\quad y_{new} = \{noisy\_client\_metrics(k, \psi_g^{l+1}, lr_i, \zeta_i)\}_{k=1}^N$
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**

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

$$\mathbb{E}[f(X)|S = s_i] = E(f(X)),\ s_i \in \{s_0, s_1\} \tag{12}$$

Let $h_v(X, S, Y, f(X)) = f(X)$ for all v, $\varphi_S = \{S = s_i\}$, and $\varphi_{\{X\setminus S\}} = \{True\}$ the equality constraints mentioned above can be represented as $\eta_S(f) = \eta_{\{X\setminus S\}}(f)$. Each equality constraint can be formulated as a pair of positive ($\Delta^+ := \eta_S(f) - \eta_{\{X\setminus S\}}(f) \leq 0$) and negative ($\Delta^- := -\eta_S(f) + \eta_{\{X\setminus S\}}(f) \leq 0$) inequality constraints. Stp can be expressed as Equation (10), 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 [21] of the constraint as follows:

$$F(\psi) = \zeta * ||(ReLU(Q\eta(f))) - e||_2. \tag{13}$$

3.3 *Server Side: Multiobjective Noisy Bayesian Optimization*

Server-side fine-tunes the constraint parameter (ζ) and learning rate (lr) through MOO based on Differentiable Noisy Expected q-Hypervolume Improvement ($qNEHVI$) [13] approach that is exact upto the Monte Carlo (MC) integration error [34]. This approach outperforms SoTA MOO methods at a fraction of their wall time. Algorithm 1 details this module of TrustFed.

The algorithm initiates with the Kaiming Uniform initialization [23] of global model 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 ($disc_g$: see Equation (6)), and global balanced accuracy (BA_g: see Equation (7)) through aggregation and averaging of local models' noisy parameters $\psi_k^l + DPN(\varepsilon)$, noisy local discrimination scores ($disc_k + DPN(\varepsilon)$), and noisy local balanced accuracy values ($BA_k + DPN(\varepsilon)$) from all the clients (Algorithm 1: lines 3 to 5). Next, we initialize a list of GP surrogate models with initial objectives ($-disc_g$, BA_g) and inputs (lr, ζ) (Algorithm 1: lines 6 to 8). We aim to maximize BA_g and minimize $disc_g$. 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 Noisy MOBO to find Pareto optimal trade-offs between BA_g and $disc_g$. MOBO initiates with the initialization of the acquisition function (α_{qNEHVI}) using the surrogate models (GP), initial inputs, and objectives (Algorithm 1: lines 9 to 10). After computing the acquisition function, we optimize it using the Sample Average Approximation (SAA) method [7] 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 $qNEHVI$, 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 aggregation and averaging of ($BA_k + DPN(\varepsilon)$) and ($disc_k + DPN(\varepsilon)$) from all clients (Algorithm 1: line 12). The surrogate GP models are updated to include the new objectives and inputs and the next Noisy MOBO round starts (Algorithm 1: lines 13 to 14). At the end of Noisy MOBO rounds, the global updates including the global model parameters, learning rate (lr') 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, rest of this section elucidates the underlying mathematical framework that guides the trade-offs between balanced accuracy and fairness in our optimization strategy. We demonstrate how the acquisition function (α_{qNEHVI}) 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, $disc$): 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 $r = (BA_{min}, -disc_{max})$ as the lower bound:

$$HV(\mathcal{P}^*, r) = \lambda_N(\cup_{j=1}^{|\mathcal{P}^*|}[r, y_j]). \tag{14}$$

HV is the N-dimensional Lebesgue measure $\lambda_N(\cdot)$ [8] of the region dominated by the Pareto front. $[r, y_j]$ is the hyper rectangle 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 (15). For our case the new set of candidate solutions corresponds to potential solutions offering varying trade-offs between BA and ($-disc$). HVI indicates the enhanced trade-off between fairness and balanced accuracy that the new set of solutions provides.

$$HVI(\mathcal{Y}, \mathcal{P}^*, r) = HV(\mathcal{P}^* \cup \mathcal{Y}, r) - HV(\mathcal{P}^*, r) \tag{15}$$

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

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

$$\alpha_{qEHVI}(U) = \mathbb{E}[HVI(\theta(U))] = \int_{-\infty}^{+\infty} HVI(\theta(U))d\theta, \tag{16}$$

For our case, $U := \{lr, \zeta\}$ where lr is the learning rate and ζ is the regularization parameter of fairness constraints as discussed in section 3.2 and q denotes the number of candidate points considered. The integral sign in the Equation (16) 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 $-disc$.
Equation (17) depicts the formulation for computation of acquisition function of $qEHVI$ through MC sampling, 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 $\theta_r(U)$).

$$\alpha^M_{qEHVI}(U) = \frac{1}{M}\sum_{r=1}^{M} HVI(\theta_r(U)) \\ = \frac{1}{M}\sum_{r=1}^{M}\sum_{w=1}^{W}\sum_{i=1}^{q}\sum_{U_i \in \mathcal{U}_i}(-1)^{i+1}\prod_{n=1}^{N}[z^{(n)}_{r,U_i,r} - l^w_{(n)}]+ \tag{17}$$

Noisy Expected Hypervolume Improvement ($NEHVI$): $EHVI$ is a Bayes optimal algorithm for one-step maximization of hypervolume in MOO. $EHVI$ operates under the assumption that the observations are noise-free. However, in our case, we cater for the privacy rights of individual clients by adding DP noise in the local observations (balanced accuracy, fairness) and local learner parameters before sharing them with the global server. In this context, we employ $NEHVI$ which serves as the acquisition function for MOBO specifically tailored for our multiple objectives: balanced accuracy, fairness, and privacy. $NEHVI$ is a novel one step Bayes optimal criterion for hypervolume maximization. This criterion iteratively computes the expectation over the posterior distribution $(P)(\theta(U_m)|O_m)$, which represents the function values at previously evaluated points U_m, given the noisy observations $O_m = \{u_r, y_r, (\sum r)\}^m_{r=1}$. The acquisition function based on $NEHVI$ can be defined as:

$$\alpha_{NEHVI}(U) = \int \alpha_{EHVI}(U|\mathcal{P}^*_m)\mathbb{P}(\theta|O_m)d\theta \tag{18}$$

The integral in the above equation can be estimated using MC integration. Let the samples from the posterior represented as $\tilde{\theta}_r \sim \mathbb{P}(\theta|O_m) for\ r = 1, ..., M$ and the Pareto frontier over previously evaluated points be $\mathcal{P}^*_r = \left\{\tilde{\theta}_r(u)|u\,\epsilon\,U_m, \tilde{\theta}_r(u) \succ \tilde{\theta}_r(u')\ \forall\ u'\epsilon\,U_m\right\}$ then the acquisition function for a batch of q points can be computed as:

$$\alpha_{qNEHVI}(U) \approx \frac{1}{M}\sum_{r=1}^{M}\alpha_{qEHVI}(u|\mathcal{P}^*_r) \tag{19}$$

3.4 Convergence and Complexity Results

In this section, we elucidate the convergence guarantees of TrustFed for both client side and server side. Additionally, we detail the computation and communication costs associated with TrustFed.

Client side convergence: Each client updates its classifier, parameterized by ψ, in each iteration t by performing fairness-constrained optimization as per Equation 9. Assuming convexity of $J(\psi)$ and $F(\psi)$, compactness of ψ's feasible set, and a suitable learning rate $lr^{(t)}$, $\psi^{(t)}$ converges to a fair classifier ψ^* that minimizes $J(\psi)$ within the constraint $F(\psi) \leq c$.

Serve side convergence: Assume that $\mathcal{U}$ is a compact set and f (sample) has a Multi-Output Gaussian Process prior with continuously differentiable mean and covariance functions. If the base samples $\{\epsilon_i\}^M_{i=1}$ are IID, drawn from a multivariate normal distribution $\mathcal{M}(0, I_{qN})$, and if $u^* \in \underset{u\in\mathcal{U}}{\text{argmax}}\ \hat{\alpha}^M_{qNEHVI}(u)$ then:
1. $\alpha_{qNEHVI}(\hat{u}^M_*) \rightarrow \alpha^*_{qEHVI}$ a.s. $\rightarrow$ The estimated acquisition function α_{qEHVI} converges almost surely (a.s.) to the true acquisition function α^*_{qNEHVI} as $M \rightarrow \infty$.
2. dist$(\hat{u}^M_*, \mathcal{U}^*) \rightarrow 0$ a.s. $\rightarrow$ The distance between the set of maximizers of the estimated acquisition function $\hat{\mathcal{U}}^*$ and the set of true maximizers $\mathcal{U}^*$ goes to zero as $M \rightarrow \infty$. The proof of the above convergence result can be found in [13].

Computation Cost The time complexity for local dataset processing through each client's model is $T_1 = \mathcal{O}(C)$. On the server side, computing the volume of $2^q - 1$ (q represents the number of candidate points) hyper-rectangles for each of K hyperrectangles and M MC samples leads to $T_2 = \mathcal{O}(MNK(2^q - 1))$ on a single-threaded machine. Specifically for $N = 2$, $K = |\mathcal{P}| + 1$. MOBO also includes evaluation of local clients to find new solutions corresponding to the candidates lr and ζ. Assuming each client takes equal time to train and share local updates, the total time complexity is $T = T_1 + C \times T_2 \sim \mathcal{O}(CMK(2^q - 1))$.

Communication Cost MOBO converges in a fraction of the wall time of traditional optimization algorithms [13], enabling TrustFed to reach optimal performance in under 15 communication rounds, as shown in Figure 1. However, competing baselines fail to reach this even after 50 rounds.

4 Experimental Setup

Benchmark Datasets We evaluate TrustFed with *five real-world datasets*: (1) Bank [3], (2) Default [3], (3) Adult [3], (4) Law [40], and (5) ACS [14]. These are considered benchmarks in the fairness domain [2] and are widely used for evaluation. This selection also aligns with the datasets used by all reported baselines, ensuring a fair comparison in evaluating TrustFed. To further demonstrate TrustFed's scalability and real-world applicability, we employ the ACS dataset, with over 1.3 million instances from 50 US states.
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, Bank, Default, Law, and ACS are $1:3.0$, $1:15.11$, $1:7.87$, $1:3.52$, $1:3.50$, and $1:1.70$ respectively. 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; income for Law) to mirror more realistic scenarios Missing the slit of ACS. Note that the baseline methods also use some of these datasets but ignore class imbalance and report accuracy as the measure of performance. We report the evaluation metrics based on the average of 5 random shuffles of each dataset that passes through model.

Baselines To demonstrate the superiority of our method, we have compared our results with **seven** most recent SoTA baselines. We employ FedAvg (along with privacy protection [1]), FF-SMOTE (debiases prediction locally using Fair-SMOTE), Agnostic-Fair, FCFL, FairTrade, FairFed, and FedFB to compare the performance of

[2] The requirement for sensitive attribute information limits the choice of datasets

Dataset	FedAvg(2016)			FF-SMOTE(2021)			Agnostic-Fair(2021)			FCFL(2021)			FedFB(2021)			FairFed(2023)			FairTrade(2024)			TrustFed		
	Acc	BA	Stp	Acc	BA	Stp	Acc	BA	Stp	Acc	BA	Stp	Acc	BA	Stp	Acc	BA	Stp	Acc	BA	Stp	Acc	BA	Stp
Bank	0.26	0.56	-0.003	0.88	0.51	-0.0006	0.87	0.53	0.064	0.89	0.56	-0.003	0.89	0.57	0.0006	0.89	0.57	0.030	0.88	0.78	0.020	0.74	0.79	-0.002
Law	0.90	0.55	0.007	0.90	0.62	0.051	0.93	0.50	0.001	0.91	0.62	-0.029	0.91	0.58	0.004	0.91	0.59	0.021	0.88	0.68	-0.008	0.85	0.76	0.0005
Default	0.47	0.61	0.043	0.80	0.63	0.043	0.79	0.52	0.006	0.80	0.57	0.089	0.80	0.59	0.009	0.79	0.58	0.026	0.77	0.69	-0.0101	0.55	0.65	0.005
Adult	0.23	0.50	0	0.82	0.68	0.070	0.77	0.58	0.067	0.80	0.62	-0.093	0.76	0.50	0.0005	0.76	0.50	0.0008	0.77	0.75	0.0012	0.612	0.72	0.022

Table 1. Acc, BA, and Stp achieved by TrustFed($\varepsilon = 3$), FedAvg($\varepsilon = 3$) and other competitors(without privacy guarantee) across all datasets with data distributed randomly among 3 clients. Most competitors yield a BA close to 0.5 (akin to a random classifier) and Stp nearing 0, TrustFed achieves the best trade-off between BA and Stp.

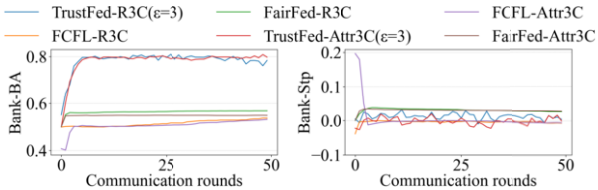

Figure 1. Comparison between BA and Stp values achieved by TrustFed ($\varepsilon = 3$), FCFL, and FairFed for Bank dataset with random and attribute based distribution of data among 3 clients over different communication rounds. Notably, TrustFed reaches high BA in under 10 rounds with low Stp, whereas competitors only achieve near-random BA ($\sim 50\%$) invalidating the significance of their low Stp.

Split	FedAvg		FF-SMOTE		FCFL		FedFB		FairFed		TrustFed	
	BA	Eqop	BA	Eqop	BA	Eqop	BA	Eqop	BA	Eqop	BA	Eqop
R3C	0.49	0.0062	0.68	-0.050	0.57	-0.102	0.50	0.666	0.50	0.666	0.74	-0.002
Attr3C	0.50	0	0.60	-0.130	0.51	0.025	0.50	0	0.49	0.00003	0.71	0.013

Table 2. Equal Opportunity ($Eqop$) and BA achieved by TrustFed for Adult dataset with R3C and Attr3C data splits.

TrustFed. For further details, see section 6. Note that we limit ourselves to comparing only with FL specific methods. While there exists several solutions for a centralized setup, we deem comparison with such methods to be inappropriate.

5 Experimental Evaluation and Discussion

We evaluate TrustFed using both Independent Identically Distributed (IID) setting (i.e., randomly distributing data across clients) and non-IID data distributions (i.e., distributing data across clients based on a specific attribute), reflecting the diverse data scenarios encountered in FL. This dual approach allows us to explore the efficacy of our framework under varied data homogeneity scenarios. It is important to note, however, that within the realm of FL, non-IID data distributions present a more realistic scenario. However, to ensure a fair comparison with baseline models that predominantly use IID data, we also include it in our analysis. This approach allows for a comprehensive evaluation, demonstrating our framework's adaptability and effectiveness across different data distributions and highlighting its advantages over existing methods.

5.1 IID Data distribution

In this setting, the dataset is randomly distributed across the clients which emulates an IID setting. We compare TrustFed with seven baseline methods across five datasets. For fair comparison, we follow the experimental setup proposed in [19]. The number of clients is set at 3. We consider statistical parity (Stp) as the fairness metric. The results in Table 1 show that TrustFed reports the best trade-off between balanced accuracy (BA) and Stp along with privacy guarantee compared to the baselines. 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

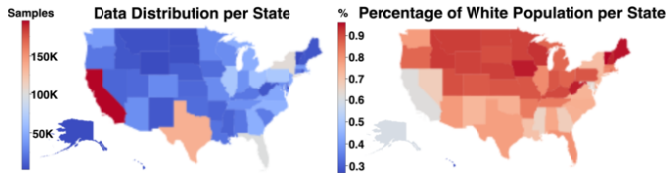

Figure 2. Data Distribution of ACS Income dataset per state

FedFB			FairFed			TrustFed		
Acc	BA	Stp	Acc	BA	Stp	Acc	BA	Stp
0.630	0.500	0.0001	0.624	0.500	0	0.590	0.668	0.0006

Table 3. Acc, BA, and Stp achieved by TrustFed($\varepsilon = 3$), FedFB, and FairFed across ACS dataset. The competitors yield a BA close to 0.5 (akin to a random classifier) and Stp nearing 0, TrustFed achieves the best trade-off between BA and Stp.

methods (without privacy guarantees), Agnostic-Fair, FCFL, FedFB, and FairFed show high Acc with the corresponding BA values ~ 0.5 and Stp scores close to 0. Such low BA values categorize these FL models as random classifiers, rendering their low Stp scores insignificant. In contrast, TrustFed achieves significantly lower Stp with remarkably higher BA for all datasets along with privacy budget of $\varepsilon = 3$. Figure 1 shows a comparison between BA and Stp values achieved by TrustFed, FCFL, and FairFed across Bank dataset over 50 communication rounds. The figure demonstrates TrustFed's ability to achieve fairness without compromising the model's BA.

Generlizability: Table 2 shows TrustFed's BA and Equal Opportunity ($Eqop$) results for Adult dataset with random data distribution among 3 clients. The table shows that TrustFed consistently maintains high BA values while achieving low $Eqop$ scores. This suggests that TrustFed is agnostic with respect to the chosen fairness metric.

5.2 Non-IID Data Distribution

To simulate the Non-IID scenario, datasets are distributed among clients based on specific attributes ('age' for Bank, Default, Adult; 'income level' for Law). Table 2 shows that TrustFed shows comparable performance to the SoTA baselines even under non-IID data scenarios (Attr3C) for Adult dataset. Figure 1 shows that even for non-IID data (Attr3C), TrustFed achieves and maintains the optimal trade-off between BA and StP for across Bank dataset. Similar trend can be observed for other datasets.

Additionally, for non-IID data scenario we utilize the American Community Survey (ACS) Dataset [14] which naturally exhibits a non-IID setup as demonstrated in Figure 2, which presents the demographic distribution of ACS data among 50 US states. Data corresponding to each state is considered a client. Table 3 presents the BA and Stp achieved by our model and competing baselines. The superior performance achieved by TrustFed underscores the effectiveness of our approach in handling real-world data complexities.

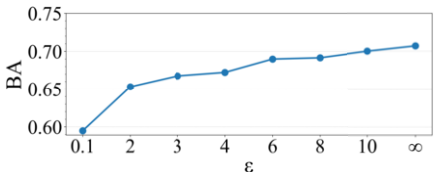

Figure 3. Privacy-Utility trade-off across ACS dataset.

Dataset	Eval. Metric	$\varepsilon = \infty$	$\varepsilon = 0.1$	$\varepsilon = 2$	$\varepsilon = 3$	$\varepsilon = 4$
Law	BA	0.771	0.598	0.763	0.767	0.769
	Eqop	-0.0327	-0.0228	0.0089	-0.0213	-0.0217
ACS	BA	0.713	0.571	0.657	0.673	0.675
	Eqop	0.00006	0.0069	-0.0013	-0.0013	-0.0005

Table 4. BA and $Eqop$ achieved by TrustFed for Law dataset with random data distribution among 3 clients and for ACS dataset (50 clients). BA increases as we increase the privacy budget (ε) with the best trade-off observed at $\varepsilon = 3$. $\varepsilon = \infty$ implies that no Differential Privacy (DP) noise is added.

5.3 Ablation Study

Sensitivity: To gain a deeper understanding of our proposed method, we also perform a set of sensitivity experiments: *varying the number of clients* and *varying the number of communication rounds*. The performance of TrustFed remains consistent with varying number of clients, which proves its adaptability regardless of the client count. Figure 1 presents the BA and Stp values achieved by TrustFed over 50 communication rounds across Bank dataset. We can observe that TrustFed achieves an optimal trade-off between BA and Stp within the initial 10 rounds and maintains it, which highlights TrustFed's efficiency in achieving and sustaining optimal trade-offs between BA and Stp while protecting the privacy rights of clients.

Effect of Varying ε: ε represents the degree of privacy guaranteed by the DP mechanism. A smaller value of ε indicates stronger privacy, as it reveals less information about individuals. Table 4 presents the values of BA and $Eqop$ achieved by TrustFed for all datasets for various values of the privacy budget (ε). It can be observed that the BA values decrease with a reduction in the privacy budget and increase when the budget is raised. The optimal trade-off is noted at $\varepsilon = 3$; hence, this value is selected for presenting the remainder of the results. We further investigate this effect by increasing ε beyond 4. Figure 3 presents the trade-off between privacy and utility (BA) for ACS dataset, where $\varepsilon = \infty$ indicates no noise. As ε increases (moving right along the x-axis), there is a general trend of increasing BA (upward movement on the y-axis). This suggests that reducing the privacy constraints (i.e., increasing ε) leads to improvements in the model's utility. The best trade-off can be observed around $\varepsilon = 3$.

6 Related Work

Fairness-aware Learning: In the field of machine learning, significant interest has developed in methods for identifying and addressing bias. A comprehensive overview of these techniques is provided in [31]. Broadly, these approaches are categorized into pre-processing, in-processing, and post-processing. *Pre-processing methods* aim to modify the training dataset to remove bias prior to model training. This approach includes methods developed by [9] and [24]. Additionally, Fair-SMOTE [10], represents a SMOTE based SoTA technique that addresses discrimination. *In-processing methods* involve modifications to the classification model itself to achieve fair outcomes. These modifications might include changes to the optimization objectives, as seen in [32], or adaptive reweighting strategies as proposed by [4]. *Post-processing methods* focus on adjusting the output of classifiers to mitigate bias. Notable methodologies in this category have been proposed by [27] and [22].

Note that all the above methods are primarily designed for centralized machine learning systems and may not directly extend to distributed settings like FL.

Fairness-aware Federated Learning: Recent efforts in FL have focused on developing methods to reduce bias. One such method, Agnostic-Fair [15], is a fairness-aware FL framework that removes discrimination through the reweighting of training data. Similarly, [20] and [42] have proposed methods to locally solve optimization problem with fairness constraints. However, these methods are not agnostic in terms of the employed fairness notion. Also they require sharing of local sensitive features with the global server. FedFB [41] offers a framework where clients independently correct bias in their predictions by utilizing Fair-Batch [35]. In another approach, FairFed [19], the authors introduce a weight aggregation technique in FL that considers discrimination. This method lets each client's degree of fairness determine its contribution to the combined global parameters, promoting greater overall fairness. Additionally, FCFL, introduced by [12], is a gradient-based method that aims to evenly distribute Pareto utility, which encompasses both accuracy and fairness, among all clients. FairTrade [6] is another fairness aware federated framework. However, it is unable to handle MIA. FAC-Fed is another fairness-aware FL framework in which the authors present an adaptation of the Synthetic Oversampling Technique (SMOTE) to mitigate discrimination in streaming environments [5].

These approaches neglect either one or both of the fundamental issues in fairness-aware FL: *privacy leakage and class imbalance*.

Differential Privacy and Federated Learning: In the realm of secure machine learning, FL and Differential Privacy (DP) have emerged as crucial areas of research and application. A Bayesian DP based method has been proposed by [36] to cater for the privacy rights of individuals in an FL system. [39] proposed an ε-DP based privacy preserving federated framework and proved that a trade-off exists between privacy budget and the rate of convergence of the learning framework. Another privacy enhancing FL framework based on local DP has been presented by [37]. For further insights into DP based privacy preserving methods in FL systems please consult [17].

Notably, these methods overlook the crucial issue of fairness and the challenge of class imbalance, which are vital for ensuring equitable and unbiased outcomes in FL models.

7 Conclusion

We proposed a novel Federated learning (FL) framework, TrustFed, aimed at achieving Pareto-optimal trade-offs between balanced accuracy and fairness while maintaining privacy rights of individuals in FL applications. Our methodology, employing Multi-Objective Optimization (MOO), presents a major leap forward from traditional SoTA FL frameworks that primarily focus on accuracy without effectively guaranteeing privacy protection. The efficacy of TrsutFed is further demonstrated through experiments across several benchmark datasets and fair FL methods, with TrustFed consistently achieving better fairness-balanced accuracy trade-off along with effective privacy guarantees. It is agnostic to the fairness metric and effectively generalizes to diverse client data distributions and varying numbers of clients. It also generalizes efficiently to real world complexities.

Acknowledgements

This work is partially supported by the research project SoBigData++ funded by the European Commission under the Horizon 2020 program with grant agreement number 871042.

Ethical Statement

Eliminating gender specific discrimination and eradicating exploitation of marginalised groups of society plays a decisive role in forging a path to a sustainable world. Our approach aims to make fair and high-quality predictions in accordance with the levels of moral equivalence; therefore, this methodology has the potential to address the issues of "gender equality" and "reducing inequalities".

We conducted our experiments with publicly available datasets, we are however committed to avoiding disclosure of personal data. We have tried to remain transparent about the capabilities and limitations of our approach by avoiding false claims and fabricating false results.

References

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

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

[3] K. Bache and M. Lichman. Uci ml repository. 2013.

[4] M. Badar and M. Fisichella. Fair-cmnb: Advancing fairness-aware stream learning with naïve bayes and multi-objective optimization. *Big Data and Cognitive Computing*, 8(2):16, 2024.

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

[6] M. Badar, S. Sikdar, W. Nejdl, and M. Fisichella. Fairtrade: Achieving pareto-optimal trade-offs between balanced accuracy and fairness in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10962–10970, 2024.

[7] M. Balandat, B. Karrer, D. Jiang, S. Daulton, B. Letham, A. G. Wilson, and E. Bakshy. Botorch: A framework for efficient monte-carlo bayesian optimization. *Advances in neural information processing systems*, 33:21524–21538, 2020.

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

[9] T. Calders and et al. Building classifiers with independency constraints. In *IEEE International Conference on Data Mining Workshops*. IEEE, 2009.

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

[11] H. Chen, T. Zhu, T. Zhang, W. Zhou, and P. S. Yu. Privacy and fairness in federated learning: on the perspective of tradeoff. *ACM Computing Surveys*, 56(2):1–37, 2023.

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

[13] S. Daulton, M. Balandat, and E. Bakshy. Parallel bayesian optimization of multiple noisy objectives with expected hypervolume improvement. *Advances in Neural Information Processing Systems*, 34:2187–2200, 2021.

[14] F. Ding, M. Hardt, J. Miller, and L. Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

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

[16] C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9 (3–4):211–407, 2014.

[17] A. El Ouadrhiri and A. Abdelhadi. Differential privacy for deep and federated learning: A survey. *IEEE access*, 10:22359–22380, 2022.

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

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

[20] B. R. Gálvez, F. Granqvist, R. van Dalen, and M. Seigel. Enforcing fairness in private federated learning via the modified method of differential multipliers. In *NeurIPS 2021 Workshop Privacy in Machine Learning*, 2021.

[21] I. Gradshteyn and I. Ryzhik. Table of integrals, series, and products 6th edn (new york: Academic). 2000.

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

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

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

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

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

[27] F. Kamiran, T. Calders, and M. Pechenizkiy. Discrimination aware decision tree learning. In *ICDM*, pages 869–874. IEEE, 2010.

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

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

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

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

[32] M. Padala and S. Gujar. Fnnc: achieving fairness through neural networks. In *IJCAI*, 2020.

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

[34] C. P. Robert, G. Casella, C. P. Robert, and G. Casella. Monte carlo integration. *Monte Carlo statistical methods*, pages 71–138, 1999.

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

[36] A. Triastcyn and B. Faltings. Federated learning with bayesian differential privacy. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 2587–2596. IEEE, 2019.

[37] S. Truex, L. Liu, K.-H. Chow, M. E. Gursoy, and W. Wei. Ldp-fed: Federated learning with local differential privacy. In *Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking*, pages 61–66, 2020.

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

[39] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020.

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

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

[42] D. Y. Zhang, Z. Kou, and D. Wang. Fairfl: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1051–1060. IEEE, 2020.