EAB-FL: Exacerbating Algorithmic Bias through Model Poisoning Attacks in Federated Learning

Syed Irfan Ali Meerza and **Jian Liu**
University of Tennessee, Knoxville
smeerza@vols.utk.edu, jliu@utk.edu,

Abstract

Federated Learning (FL) is a technique that allows multiple parties to train a shared model collaboratively without disclosing their private data. It has become increasingly popular due to its distinct privacy advantages. However, FL models can suffer from biases against certain demographic groups (e.g., racial and gender groups) due to the heterogeneity of data and party selection. Researchers have proposed various strategies for characterizing the group fairness of FL algorithms to address this issue. However, the effectiveness of these strategies in the face of deliberate adversarial attacks has not been fully explored. Although existing studies have revealed various threats (e.g., model poisoning attacks) against FL systems caused by malicious participants, their primary aim is to decrease model accuracy, while the potential of leveraging poisonous model updates to exacerbate model unfairness remains unexplored. In this paper, we propose a new type of model poisoning attack, EAB-FL, with a focus on exacerbating group unfairness while maintaining a good level of model utility. Extensive experiments on three datasets demonstrate the effectiveness and efficiency of our attack, even with state-of-the-art fairness optimization algorithms and secure aggregation rules employed. Code is available at https://github.com/irfanMee/EAB-FL

1 Introduction

Federated Learning (FL) [Konečný *et al.*, 2016] has recently emerged as a promising solution that enables multiple clients to collaboratively learn a shared prediction model while keeping all the training data on the device. Due to its privacy-preserving nature, in recent years, FL has benefited a wide variety of privacy-sensitive application domains, such as medical research [Rauniyar *et al.*, 2023], and financial fraud [Liu *et al.*, 2023], etc. However, the distributed nature of FL makes it inherently vulnerable to poisoning attacks, in which the model can be compromised by malicious clients uploading malicious model updates. Without a central authority to validate clients' participation, these malicious clients can indirectly and consequently manipulate the parameters of the learned model and thereby reduce its overall performance [Cao and Gong, 2022].

In addition, compared to centralized learning, FL models are more susceptible to algorithmic bias against specific demographic groups (e.g., racial and gender groups) due to its inherent characteristics, such as data heterogeneity, party selection, and client dropping out [Abay *et al.*, 2020]. Compounding this issue, the involvement of malicious participants within FL environments can further exacerbate these biases.

Attacker's Motivations. Attacking FL models through poisoning attacks to exacerbate their unfairness could provide an attacker with various benefits. For instance, e-commerce websites can be targeted with fairness attacks on their recommendation algorithms to suggest certain groups of products or services for the benefit of their providers while harming others. In the case of loan applications, an attacker might attempt to manipulate the FL model to unfairly discriminate against or favor specific groups of people, resulting in unjust loan decisions. Furthermore, the attacker can also attack the models used in criminal justice areas to make them unfair to reduce the trust and credibility in the criminal justice system. Given the aforementioned potential and motivations of fairness attacks, it is crucial to have a comprehensive understanding of the attack surfaces of FL, particularly in the domain of fairness.

Existing Attacks. There are a few studies exploring poisoning attacks against FL systems, either by adding poisonous instances or adversarially changing model updates [Cao and Gong, 2022; Bagdasaryan *et al.*, 2020]. However, these attacks are proposed with the purpose of reducing the model's classification accuracy without any regard for the model's fairness. In centralized machine learning, there have been attacking efforts to exacerbate algorithmic bias, such as the gradient-based poisoning attacks [Solans *et al.*, 2021] and the anchoring and influence attacks [Mehrabi *et al.*, 2021], which aim to maximize the covariance between the sensitive attributes and the decision outcome to affect the model fairness. However, these attacks can hardly be adapted to the FL settings due to the challenges in measuring the impact of each data sample on fairness violation in the learned model. This is mainly because the data used in FL is often decentralized and not directly accessible, making it hard to evaluate the specific impact of each sample on the model. To the best of our knowledge, the potential of leveraging poisoning attacks in FL to exacerbate group unfairness remains unexplored.

Our Attack. In this paper, we design a new type of model poisoning attack, EAB-FL, where an adversary can introduce

or exacerbate algorithmic bias against certain groups of individuals or samples while maintaining a relatively good level of model utility (i.e., classification accuracy). We assume the adversary has compromised a small fraction of client devices (a.k.a., malicious clients) and can manipulate the training process on these devices. To maintain the overall model utility, EAB-FL first identifies the redundant space of the locally trained model by applying layer-wise relevance propagation (LRP) [Bach *et al.*, 2015]. This space comprises neurons that remain relatively stable during the learning process and exhibit minimal correlation with the prediction task for the privileged demographic group (i.e., the one that the adversary does not want to impact). Subsequently, EAB-FL adjusts the model parameters within the redundant space by solving an optimization problem on a subset of local datasets that adversely influences the model's performance for the targeted group. Consequently, our method can preserve high utility for the privileged group while simultaneously reducing it for the targeted group, thus inducing algorithmic bias in the model.

To make the attack less suspicious and more generalized across different FL settings, EAB-FL has the following major properties: *(1) High Model Utility:* Unlike other poisoning attacks in FL that target model classification accuracy, our attack aims to exacerbate group unfairness only, with a minimum impact on the model's overall utility, thereby rendering the attack less noticeable; *(2) Persistence and Stealth Attack:* We intent to embed adversarial features into the redundant space of the model to improve the persistence of attack while remaining the alterations of the model parameters minimal, aiming to ensure that the attack remains robust against secure aggregations; and *(3) Effective under Fairness Optimizations:* Our attack can remain effective under existing fairness optimization strategies (e.g., FEDFB [Zeng *et al.*, 2022], FAIRFED [Ezzeldin *et al.*, 2023]), as the adversary optimizes the poisoning each time it participates in the training, which allows it to nullify the effect imposed by any fairness optimization methods.

Through extensive experiments on three datasets with different fairness optimization, we demonstrate the effectiveness of EAB-FL in achieving the desired goal of exacerbating group unfairness. We also evaluate our attack under various state-of-the-art secure aggregation rules in FL, and the results demonstrate its sustained efficacy.

2 Related Work

Model poisoning attacks against FL systems have received considerable attention in recent years. These attacks are designed to manipulate the global model by tampering with local training processes on a fraction of participating clients. For instance, malicious clients can significantly degrade the performance of the global model by adding random noise to the local model to mislead the global model [Hossain *et al.*, 2021; Cao and Gong, 2022]. Xingchen *et al.* [Zhou *et al.*, 2021] proposed an optimization-based model poisoning attack to inject poisonous neurons into the model's redundant space, identified using the Hessian matrix. Moreover, it has been shown that the adversary can replace the aggregated model with the malicious model through one compromised device to perform model-replacement attacks [Xie *et al.*, 2020; Bagdasaryan *et al.*, 2020]. Henger *et al.* [Li *et al.*, 2022] proposed a model poisoning attack using reinforcement learning, where malicious clients collectively learn the data distribution to launch an optimal attack. While the aforementioned attacks have shown a great chance of compromising FL models, they only target the model's utility by either decreasing its overall classification accuracy or making specific test samples classified as adversary-desired labels.

In addition, it has been shown that FL models are more likely to suffer from algorithmic bias compared to centralized learning due to their inherent characteristics, such as data heterogeneity, party selection, and client dropping out [Abay *et al.*, 2020]. To address this issue, initial studies [Li *et al.*, 2020; Mohri *et al.*, 2019; Lyu *et al.*, 2020; Li *et al.*, 2021a; Zhong *et al.*, 2022] focus on client parity and aim to promote equalized accuracies across all participating clients in FL. This is typically achieved by reducing the client's performance disparity [Li *et al.*, 2020] or maximizing the performance of the worst client [Mohri *et al.*, 2019]. More recent studies have addressed group fairness [Abay *et al.*, 2020; Zhang *et al.*, 2020; Rodríguez-Gálvez *et al.*, 2021; Chu *et al.*, 2021; Ezzeldin *et al.*, 2023; Du *et al.*, 2021]. These works propose solutions for providing fair performance across different sensitive groups. These studies mainly utilize deep multi-agent reinforcement learning [Zhang *et al.*, 2020], re-weighting mechanisms [Abay *et al.*, 2020; Ezzeldin *et al.*, 2023; Du *et al.*, 2021], optimization with fairness constraints (e.g., via alternating gradient projection [Chu *et al.*, 2021] or the modified method of differential multipliers [Rodríguez-Gálvez *et al.*, 2021]) to achieve group fairness.

In adversarial scenarios, attacks targeting fairness measures in machine learning are a relatively new concept, and only a few studies have been proposed in this area. Solans et al. [Solans *et al.*, 2021] were among the first to propose a fairness attack that uses a gradient-based poisoning attack to introduce classification disparities among different groups. Mehrabi et al. [Mehrabi *et al.*, 2021] proposed anchoring and influence attacks to introduce algorithmic bias in machine learning algorithms. More recently, Chhabra et al. [Chhabra *et al.*, 2023] proposed a black-box fairness attack on clustering algorithms. However, all these attacks have been proposed in centralized learning settings. To the best of our knowledge, the potential of adversarial attacks aiming to exacerbate model unfairness in FL remains unexplored, which is vital for us to fully understand the attack surfaces of FL and thereby help facilitate corresponding mitigations to improve its resilience.

3 Preliminaries

3.1 Federated Learning

Federated Learning (FL) is a collaborative approach in machine learning where a global model is trained across multiple distributed clients under the supervision of a central server. This method is distinct because it does not require direct access to client data, thereby enhancing privacy and data security. FL's primary aim is to optimize the global model's parameters while effectively utilizing the diverse, decentralized data held by each client. The key formula governing FL is:

$$\min_{\theta_g} \quad f(\theta_g) = \sum_{k=1}^{n} p_k \mathcal{L}_k(\theta_g), \quad \mathcal{L}_k = \frac{1}{d_k} \sum_{j_k=1}^{d_k} l_{j_k}(\theta_g), \qquad (1)$$

where, θ_g represents the global model parameters, n is the total number of clients, p_k is the probability of each client k's participation, $\mathcal{L}_k$ is the empirical loss for client k, l_{j_k} is the loss for each data sample j of client k, and d_k denotes the number of data samples of client k. The optimization in FL typically involves selecting a subset of clients in each training round, based on their participation probability, and then applying local optimizers like Stochastic Gradient Descent (SGD). A widely used model aggregation method, FedAvg [McMahan *et al.*, 2017], involves averaging the participating client models. However, this approach often results in performance inconsistencies among clients, with potential biases towards clients with larger datasets or those participating more frequently. This can inadvertently introduce biases against certain demographic groups in the dataset.

3.2 Group Fairness in FL

In a binary classification task, we deal with training samples of the form $(x_1, y_1, g_1), ..., (x_d, y_d, g_d)$ where each example consists of an instance $x_i \in X$, a label $y_i \in Y$, and a sensitive attribute $g_i \in G$. The goal is to develop a classification model, denoted as $f(\theta) : X \to Y$, which aims to minimize the cumulative loss, $\mathcal{L}_m(\theta, \mathcal{D}) = \sum_{(x,y)\in\mathcal{D}} l(f(x, \theta), y)$, over the training dataset $\mathcal{D} = (X, Y, G)$ to find the optimal parameters. In the context of Federated Learning (FL), the framework strives not only for accuracy but also for fairness in model predictions concerning the sensitive attribute g_i. Fairness is evaluated based on certain notions, such as demographic parity and equal opportunity. A model is considered fair from a group fairness perspective if it performs equally well for both the privileged group ($g_i = 1$) and the underprivileged group ($g_i = 0$). For a model yielding binary predictions $\hat{Y}$, given data samples X and their corresponding labels Y, we use the following two metrics to assess the model's fairness:

Demographic Parity [Dwork *et al.*, 2012]: If a classifier's predictions $\hat{Y}$ is statistically independent of the sensitive characteristic G, it meets demographic parity under a distribution (X, Y, G). This is equivalent to $\mathbb{E}[\hat{Y}|G = a] = \mathbb{E}[\hat{Y}]$, where $a = 0$ or 1 for a binary group. Demographic parity can be defined as:

$$Pr\{\hat{Y} = 1|G = 1\} = Pr\{\hat{Y} = 1|G = 0\}. \tag{2}$$

Equal Opportunity [Hardt *et al.*, 2016]: If a classifier's predictions $\hat{Y}$ is conditionally independent of the sensitive feature given the label, it meets equalized opportunity under a distribution (X, Y, G). This is the same as $\mathbb{E}[\hat{Y}|G = a, Y = 1] = \mathbb{E}[\hat{Y}|Y = 1]$. In this case, we want the true positive rate $Pr\{\hat{Y} = 1|Y = 1\}$ to be the same for each population with no regard for the errors when $Y = 0$. Equal Opportunity thus can be defined as:

$$Pr\{\hat{Y} = 1|Y = 1, G = 1\} = Pr\{\hat{Y} = 1|Y = 1, G = 0\}. \tag{3}$$

3.3 Influence Score

In EAB-FL, to exacerbate model bias, each malicious client needs to identify a subset of their local training samples that can detrimentally affect the performance of a specific demographic group (i.e., targeted group). To quantify the impact of each local training sample on the model's performance concerning the targeted demographic group, we use "influence score" [Wang *et al.*, 2022] to assess how a model's prediction on the data samples from the targeted group would change if a training sample (x_i, y_i, g_i) is excluded from the training dataset, particularly under fairness constraints imposed on the classifier. Specifically, the influence score of a training example (x_i, y_i, g_i) concerning a specific demographic group (e.g., $g_j = \tau$) can be represented as:

$$\begin{aligned}\text{inf}_l(\mathcal{D}, i) \approx \int_{g_j=\tau} \Theta(x_i, x_j; \theta)\Big(\frac{\partial \mathcal{L}(w, y_i)}{\partial w}\Big|_{w=f(x_i;\theta)} \\ + \frac{\partial \phi(f, g_i)}{\partial l}\Big)\Big|_{f(x_i;\theta)}\Big) \mathrm{d}Pr(x_j, y_j, g_j),\end{aligned} \tag{4}$$

where $\text{inf}_l(\mathcal{D}, i) \in \mathbb{R}$, Θ is the Neural Tangent Kernel (NTK) [Jacot *et al.*, 2018] and θ represents the parameters of the classification model. $\phi(f, g_i)$ denotes a differentiable surrogate for commonly used fairness constraints, such as Demographic Parity or Equal Opportunity, μ is a tolerance parameter for fairness deviations, and $\mathrm{d}Pr(x_j, y_j, z_j)$ refers to the differential probability measure over the data distribution $\mathcal{D}$. We solve this equation by calculating the Jacobian of the function that multiplies the model's output with the demographics attribute $(g_j = \tau)$, and then compute a kernel matrix from the gradients. A positive value suggests that including the training sample (x_i, y_i, g_i) improves the model's performance for the targeted demographic group, while a negative value implies it hinders accuracy. The magnitude of the value shows the strength of this impact.

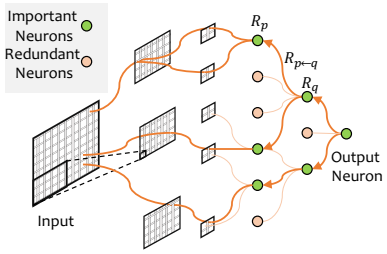


Figure 1: Illustration of the LRP procedure.

3.4 Layer-wise Relevance Propagation

Layer-wise Relevance Propagation (LRP) [Bach *et al.*, 2015] is an explanation technique, aiming to propagate the model prediction $f(x, \theta)$ backward in the model to quantify the contributions of each neuron to the model prediction. Specifically, during the backward propagation of LRP, as shown in Figure 1, each neuron redistributes the received relevance scores to the preceding layer in equal proportions. By examining the propagated relevance scores, LRP can determine the degree of influence each neuron has on the model prediction. Neurons with higher relevance scores are deemed more "important" in decision-making, while those with lower scores are considered relatively "redundant". If we consider p and q as neurons in two successive layers, the relevance scores $(R_q)_p$ at one layer are transferred to the neurons in the layer below by applying the following rule:

$$R_p = \sum_q \frac{z_{pq}}{\sum_p z_{pq}} R_q, \tag{5}$$

where z_{pq} is the product of the activation of neuron p and the weight of the connection from neuron p to neuron q, which

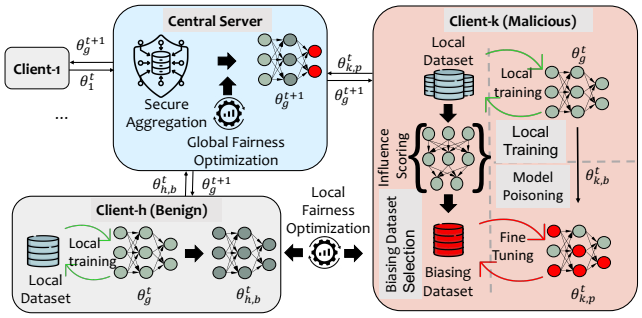


Figure 2: Illustration of the proposed EAB-FL. text before after

quantifies the contribution of neuron p to the relevance of neuron q. The denominator acts as a normalizing factor to uphold the conservation principle, ensuring that the sum of relevance scores propagated from a neuron matches the sum of scores it received. In EAB-FL, we use LRP procedure to identify the model's redundant space and confine all poisoning alterations within this space to exacerbate model bias while ensuring minimal impact on the model utility.

4 Poisoning Attacks against Fairness in FL

4.1 Threat Model and Adversarial Goals

Threat Model: In this work, we assume an adversary has compromised a small fraction of client devices (a.k.a., malicious clients), and the adversary can tamper with the local training process on the compromised client devices during the learning to exacerbate model bias. The attack does not involve manipulating the local datasets, making it easy to bypass any security measures focused on data integrity. During the FL participation process, we assume that the central server may employ security measures to validate the credibility of the submitted model updates [Pillutla *et al.*, 2022; Bhagoji *et al.*, 2019]. For instance, the server can evaluate the accuracy of the submitted model updates on a validation set, and the server can also verify model updates through secure aggregation rules (e.g., Krum [Yin *et al.*, 2018]) to mitigate or reject anomalous model updates.

Adversarial Goal: Different from traditional poisoning attacks, which only aim to decrease the model's utility, the adversary in our attack aims to rely on the compromised client devices to exacerbate the algorithmic bias in the learned global model while maintaining a relatively good level of model utility (e.g., classification accuracy).

Adversary's Knowledge: In the threat model, we assume that the adversary has full knowledge of the structure and parameters of the global model θ_g^t received at each communication round. The adversary is considered to be active, meaning that the adversary can tamper with the local training process and locally deploy an optimization-based approach on each compromised client device. However, the adversary does not have any knowledge of the aggregation rule used on the server or the fairness optimization methods applied in FL. We assume a non-colluding malicious client, where the adversary has no partner to exchange any information nor does it have any knowledge of the benign clients' local training.

4.2 Attack Design Overview

In this work, we propose an optimization-based model poisoning attack in FL, EAB-FL, that targets the group fairness measure of the learned global model. As shown in Figure 2, the FL system consists of a central server and n clients, a small fraction of which have been compromised by an adversary. During each communication round, benign (non-compromised) clients (e.g., the client h) train the global model (θ_g^t) provided by the central server on their local datasets for a certain number of epochs and send the updated model ($\theta_{h,b}^t$) back to the server. Additionally, the FL system may employ either secure aggregation rules (e.g., Krum [Yin *et al.*, 2018] and FLDetector [Zhang *et al.*, 2022]) and/or fairness optimization strategies (e.g., FEDFB [Zeng *et al.*, 2022] and FAIRFED [Ezzeldin *et al.*, 2023]) to defend against potential poisoning attacks and ensure model fairness.

On those malicious client devices, the adversary can launch the attack through three stages: (1) each client (e.g., the client k) follows the normal procedure by training the global model on their local datasets for a certain round of epochs, generating a benign model $\theta_{k,b}^t$; (2) the adversary uses the global model (θ_g^t) to calculate the influence score of each local data sample from the privileged demographic group over the targeted group using Equation 4 and creates a local biasing dataset ($\mathcal{D}_k^{bias}$) that can reduce the model's classification accuracy for the targeted group; and (3) the adversary first identifies the redundant space (i.e., the neurons that remain relatively invariant during training) using LPR in the benign model update $\theta_{k,b}^t$, and then manipulates the neurons within this space using the created biasing dataset, leading to the poisoned model $\theta_{k,p}^t$. This poisoned model is sent to the server, where it is aggregated with updates from other clients to update the global model (θ_g^{t+1}) for the next communication round. With this multi-step approach, the adversary can induce overfitting of the model to the privileged demographic group while decreasing accuracy for the targeted group, thereby exacerbating model bias, while the impact on its overall utility remains minimal.

4.3 Design of EAB-FL

The crux of EAB-FL lies in inducing bias without compromising the attack's persistence or stealthiness. Unlike conventional poisoning strategies that only attack the global model's utility, our attack employs an optimization-based approach that is more sophisticated and tailored for inducing bias in FL. The attack pipeline on each malicious client unfolds as:

(1) **Regular Local Training**: The adversary intends to introduce or exacerbate model bias by maintaining a high model utility for the privileged demographic group while decreasing the model accuracy for the targeted demographic group. To ensure high utility for the privileged demographic, during the communication round t, each malicious client (e.g., client k) follows a similar procedure as benign clients to conduct training on their local dataset $\mathcal{D}_k$. The objective function can be represented as:

$$\min_{\theta_{k,b}^t} \quad \frac{1}{|\mathcal{D}_k|} \sum_{i=1}^{|\mathcal{D}_k|} \mathcal{L}_c(f(x_i; \theta_{k,b}^t), y_i) \quad s.t. \phi(f, g_i) \leq \mu, \quad (6)$$

where $\phi(f, g_i)$ serves as a differentiable proxy for the fairness constraints (i.e., Demographic Parity used in EAB-FL), with

μ representing a fairness tolerance parameter, and $|\mathcal{D}_k|$ is the number of local training samples.

(2) **Biasing Dataset Selection**: The adversary then uses the current global model (θ_g^t) to calculate the influence scores of local data samples from the privileged demographic group on the targeted demographic group, τ, using Equation 4. Upon calculating these influence scores, the adversary ranks the samples in ascending order and earmarks a predetermined fraction—denoted by κ of the size of the privileged demographic group to create $\mathcal{D}_k^{bias}$. Choosing κ is a strategic decision, that significantly impacts the attack's effectiveness.

(3) **Model Poisoning**: In FL, the adversary needs to ensure their model manipulations persist and remain effective amidst model updates aggregated from other clients on the server. While a straightforward strategy would be to introduce a significantly large local update, this method could jeopardize the attack's stealthiness and might be easily neutralized by secure aggregation rules. To tackle this, inspired by existing studies (e.g., [Li *et al.*, 2019b]) which show that certain neurons in a model are relatively redundant for model predictions and tend to remain invariant during training, EAB-FL embeds adversarial influences into this "redundant" space, aiming to establish a persistent and robust adversarial path within the model that can withstand the aggregation process in FL.

To effectively poison the model, the adversary initially identifies the model's redundant space in $\theta_{k,b}^t$. This is accomplished by utilizing Layer-Wise Relevance Propagation (described in Section 3.4), which assigns relevance scores to each neuron, thereby highlighting their respective contributions to the model's primary classification task for the privileged group. Subsequently, the adversary proceeds to inject adversarial influences into this space by solving the following optimization problem with the selected biasing dataset $\mathcal{D}_k^{bias}$:

$$\min_{\theta_{k,p}^t} \frac{1}{|\mathcal{D}_k^{bias}|} \sum_{i=1}^{|\mathcal{D}_k^{bias}|} \mathcal{L}(l(x_i;\theta_{k,p}^t), y_i) + \gamma \sum_{\theta^* \in \theta_{k,p}^t} h(\theta_{k,b}^t)(\Delta\theta^*)^2 + \rho||\theta_{k,p}^t - \theta_{k,b}^t||_2, \tag{7}$$

where γ and ρ are weight factors for the two regularization terms, which are employed to ensure model poisoning occurs within the redundant space and to penalize the poisoned model update ($\theta_{k,p}^t$), which deviates much from the benign model update ($\theta_{k,b}^t$) produced in the initial step; $h(\theta_{k,b}^t)$ represents the LRP score matrix of the model $\theta_{k,b}^t$, indicating the importance of neurons; and $\Delta\theta^*$ reflects the adjustments made during the model poisoning.

Solving this optimization problem enables the adversary to induce overfitting of the model to the privileged demographic group while decreasing accuracy for the targeted group, thus exacerbating model bias while maintaining overall model utility. The two regularization terms are designed to keep the poisonous updates closely aligned with benign updates and to confine modifications primarily to the redundant space, which tends to remain relatively stable during training. Such a strategy renders the attack more covert and helps it endure the aggregation process on the server in FL.

5 Evaluation

5.1 Federated Datasets

We evaluate the proposed EAB-FL using the following three datasets in non-IID settings:

(1) *CelebA* [Liu *et al.*, 2018]: A collection of $200k$ celebrity face images from the Internet that have been manually annotated. The dataset has up to 40 labels, each of which is binary-valued. For CelebA, each subject's gender (male or female) is a sensitive attribute. (2) *Adult Income* [Dua and Graff, 2017]: A tabular dataset that is widely investigated in machine learning fairness literature. It contains $48,842$ samples with 14 attributes. In this dataset, race (white or non-white) is used as the sensitive attribute. (3) *UTK Faces* [Zhang *et al.*, 2017]: A large-scale face dataset with more than 20,000 face images with annotations of age, gender, and ethnicity. Race (white or non-white) is used as the sensitive attribute in this dataset.

To show the real-world implications, we also apply EAB-FL to the *MovieLens 1M* dataset [Harper and Konstan, 2015] (a movie recommendation system).

5.2 Evaluation Metrics

(1) *Equal Opportunity Difference (EOD)*: We use the EOD of each sensitive group to measure group fairness. Specifically, EOD $= |Pr\{\hat{Y} = 1|G = 0, Y = 1\} - Pr\{\hat{Y} = 1|G = 1, Y = 1\}|$.

(2) *Demographic Parity Difference (DPD)*: DPD is another metric used for measuring group fairness, which is calculated as DPD $= \Big|Pr\{\hat{Y} = 1|G = a\} - Pr\{\hat{Y} = 1\}\Big|$.

(3) *Utility*: In our experiments, we use the model's prediction accuracy to quantify global model utility.

5.3 Fairness Attack Baselines

Since no fairness attack specifically designed for FL currently exists, we adapted two fairness attacks originally designed for centralized learning (i.e., gradient-based [Solans *et al.*, 2021] and anchoring-based attack [Mehrabi *et al.*, 2021]) to FL settings to demonstrate the superiority of EAB-FL. Specifically, to introduce or exacerbate model bias, the gradient-based attack employs a bi-level optimization process to inject a small fraction of poisoning points into the training data, while the anchoring-based attack strategically introduces poisoned data points near the targeted demographic group, sharing the same demographic characteristics but with opposite labels. In FL settings, we adapted these attacks by enabling malicious clients to employ them locally during their training processes. Unlike centralized learning, since we cannot measure the fairness impact of each local data sample on the global model, these attacks adapted for FL primarily target degrading the fairness level within their respective local datasets, rather than the global model as a whole.

5.4 Fairness Optimization Strategies

To evaluate the effectiveness of our attack under certain fairness optimization strategies applied in FL, besides FEDAVG [McMahan *et al.*, 2017], we also, adopt the following state-of-the-art FL fairness optimization methods to evaluate our attack's effectiveness: *(i)* Q-FFL [Li *et al.*, 2020]: Q-FFL is one of the client-fairness-based methods, aiming to equalize the accuracies of all the clients by dynamically reweighting the

Fairness Optimization	Attack	CelebA ($\epsilon = 0.1$)			Adult Income ($\epsilon = 0.2$)			UTK Faces ($\epsilon = 0.2$)		
		EOD (↓)	DPD(↓)	Utility	EOD (↓)	DPD(↓)	Utility	EOD (↓)	DPD(↓)	Utility
		Gender	Gender	(↑)	Race	Race	(↑)	Race	Race	(↑)
FEDAVG [McMahan *et al.*, 2017]	No Attack	0.23	0.21	**91**%	0.25	0.27	**83**%	0.24	0.22	**87**%
	Gradient-based	0.25	0.23	85%	0.27	0.27	79%	0.29	0.31	82%
	Anchoring-based	0.24	0.22	81%	0.27	0.29	77%	0.27	0.28	81%
	EAB-FL	**0.41**	**0.43**	88%	**0.41**	**0.44**	80%	**0.38**	**0.34**	84%
Q-FFL [Li *et al.*, 2020]	No Attack	0.19	0.20	**89**%	0.22	0.24	**82**%	0.18	0.21	**84**%
	Gradient-based	0.21	0.21	84%	0.26	0.28	79%	0.22	0.25	80%
	Anchoring-based	0.21	0.23	83%	0.24	0.24	76%	0.21	0.21	79%
	EAB-FL	**0.36**	**0.39**	85%	**0.39**	**0.38**	79%	**0.33**	**0.30**	81%
GIFAIR-FL [Yue *et al.*, 2023]	No Attack	0.19	0.19	**88**%	0.21	0.23	**82**%	0.17	0.20	**83**%
	Gradient-based	0.21	0.20	84%	0.25	0.28	78%	0.20	0.21	76%
	Anchoring-based	0.21	0.22	81%	0.22	0.24	76%	0.20	0.19	73%
	EAB-FL	**0.33**	**0.37**	84%	**0.38**	**0.36**	79%	**0.31**	**0.28**	78%
FAIRFED [Ezzeldin *et al.*, 2023]	No Attack	0.16	0.16	**87**%	0.18	0.19	**80**%	0.17	0.15	**82**%
	Gradient-based	0.19	0.20	83%	0.23	0.25	75%	0.23	0.21	72%
	Anchoring-based	0.21	0.19	79%	0.19	0.21	72%	0.20	0.19	70%
	EAB-FL	**0.31**	**0.26**	84%	**0.29**	**0.30**	78%	**0.28**	**0.28**	76%
FEDFB [Zeng *et al.*, 2022]	No Attack	0.15	0.16	**84**%	0.19	0.19	**79**%	0.16	0.17	**82**%
	Gradient-based	0.21	0.23	79%	0.25	0.25	72%	0.20	0.22	75%
	Anchoring-based	0.19	0.18	76%	0.24	0.25	70%	0.19	0.20	72%
	EAB-FL	**0.31**	**0.29**	81%	**0.32**	**0.33**	75%	**0.30**	**0.30**	77%

Table 1: Group fairness comparison of different fairness optimization methods under different attack scenarios.

aggregation, favoring the clients with poor performance. *(ii)* GIFAIR-FL [Yue *et al.*, 2023]: GIFAIR-FL aims to achieve client fairness using regularization terms to penalize the spread in the loss. *(iii)* FEDFB [Zeng *et al.*, 2022]: FEDFB is a group fairness method, where they have fitted the concept of the fair batch from centralized learning into FL by leveraging the shared group-specific positive prediction rate for each client. *(iv)* FAIRFED [Ezzeldin *et al.*, 2023]: FAIRFED is a group fairness method that can improve both local and group fairness. It employs FEDAVG and a fairness- based re-weighting mechanism to account for the mismatch between global fairness measure and local fairness measure.

5.5 Attack Performance

Attack Performance under FEDAVG. As shown in Table 1, if no attack is present under FEDAVG, the learned global model tends to be biased against certain demographic groups already (e.g., EOD and DPD on the CelebA dataset are 0.23 and 0.21, respectively), while the model utility is at a relatively good level (e.g., 91% on the CelebA dataset). Introducing an attack amplifies this inherent unfairness, as demonstrated in our evaluations where each participating client has an ϵ probability of being malicious. The results indicate that compared to the two fairness attacks, EAB-FL is capable of introducing significantly greater model bias while maintaining a minimal impact on its utility. In the FEDAVG setting, without employing any fairness optimizations, these baseline attacks only slightly increase model bias. This is likely due to the adversarial updates being overshadowed by benign updates, leading to a catastrophic forgetting of the adversarial update (for instance, EOD values for the gradient-based and anchoring-based attacks on the CelebA dataset are just 0.25 and 0.24, respectively). However, EAB-FL significantly impacts the global model fairness (e.g., EOD and DPD values on the CelebA dataset are 0.41 and 0.43, respectively), meanwhile keeping its utility high.

Attack Performance under Fairness Optimization. To validate whether our attack can remain effective under existing fair optimization strategies in FL, we evaluate our attack under four state-of-the-art fair FL methods, including Q-FFL [Li *et al.*, 2020], GIFAIR-FL [Yue *et al.*, 2023], FAIRFED [Ezzeldin *et al.*, 2023] and FEDFB [Zeng *et al.*, 2022]. The results, as shown in Table 1, indicate that while baseline attacks such as gradient-based and anchoring-based attacks yield a slight degradation in performance, EAB-FL significantly undermines fairness, even in the face of these advanced optimization strategies. For instance, when applying Q-FFL to the UTK Faces dataset, our attack results in an EOD of 0.33, signifying that the privileged racial group (white) is 33% more likely to receive correct classifications. Similarly, a DPD of 0.30 indicates a 30% higher likelihood for the privileged group to be positively classified. We also observe that EAB-FL is relatively more effective against Q-FFL and GIFAIR-FL than against FAIRFED and FEDFB. This is likely because FAIRFED and FEDFB implement fairness constraints at the local model level, enhancing each local model's fairness. In contrast, Q-FFL and GIFAIR-FL apply fairness during server-side aggregation, making them more susceptible to exploitation by EAB-FL.

Impact of Malicious Participant Availability (ϵ). As shown in Figure 3, we can see that the fairness measures (i.e., EOD and DPD) significantly increase from 0.23 and 0.21 to 0.41 and 0.43 for the CelebA data with the increasing probability of participating clients being malicious (ϵ, from 0 to 0.3). However, further increasing ϵ does not have a proportional impact on the model's fairness. This is due to the constraints on the model weights that make the malicious updates appear less suspicious to the server and prevent the client's local model from being trivial. We can observe a similar trend for other datasets as well, like for the Adult Income data dataset, the saturation comes at 0.4, and for UTK faces it comes at 0.3.

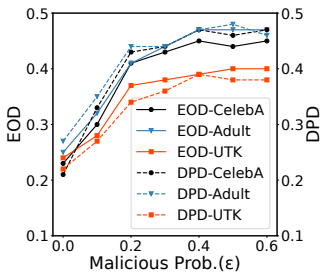


Figure 3: Impact of malicious participant probability.

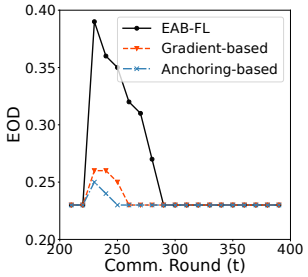
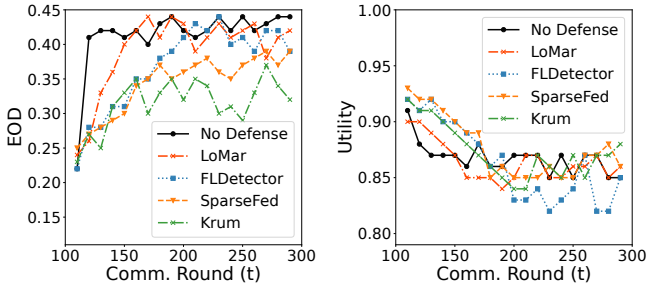


Figure 4: Attack persistence evaluation.

Attack Persistence. Attack persistence allows us to measure the durability of the attack's impact after the removal of malicious clients from the training process. To evaluate this, we conduct a single-shot attack scenario on the CelebA dataset, where the attack is initiated only once at round $t = 230$. For this round, we assume that all participating clients are malicious. Figure 4 presents a comparative analysis of EAB-FL and other baseline fairness attacks under this scenario, which indicates that EAB-FL maintains a high level of attack success (high EOD), over an extended period of model aggregation. The key to the proposed attack's sustained effectiveness lies in its strategic placement of poisoning neurons within the neural network's redundant space. By embedding the adversarial influence in these less dynamic areas of the network, we significantly reduce the likelihood of these neurons being altered during the training of the main task, thus ensuring the longevity of the attack's impact.

Effectiveness against Secure Aggregations. We evaluate the effectiveness of EAB-FL against four robust aggregation rules: SparseFed [Panda *et al.*, 2022], Krum [Yin *et al.*, 2018], LoMar [Li *et al.*, 2021b], and FLDetector [Zhang *et al.*, 2022] on the CelebA dataset. As shown in Figure 5(a), our proposed EAB-FL can bypass the norm-bounded defense method (SparseFed) and achieves significantly high unfairness (e.g., EOD is over 0.35) in the global model. Although Krum, effective at detecting aberrant models, sometimes rejects our attack's poisonous updates, the adversary still maintains a much higher EOD (e.g., over 0.3 for most cases) compared to scenarios without attacks. Model-similarity-based methods, such as LoMar and FLDetector, fail to detect EAB-FL, as the poisonous updates closely resemble benign updates, constrained by the L2 norm. Figure 5(b) shows that EAB-FL retains high



(a) (b)

Figure 5: Attack effectiveness against secure aggregations.

Attack	EOD (↓)	DPD(↓)	Utility
	Gender	Gender	(↑)
DeSMP [Hossain *et al.*, 2021]	0.24	0.21	52%
MPAF [Cao and Gong, 2022]	0.28	0.30	49%
Sign Flipping [Li *et al.*, 2019a]	0.25	0.23	55%
DYN-OPT [Shejwalkar and Houmansadr, 2021]	0.29	0.31	53%
BDB [Shejwalkar *et al.*, 2022]	0.32	0.29	51%
EAB-FL	**0.45**	**0.50**	**83%**

Table 2: Comparison of EAB-FL with other poisoning attacks.

Attack	No Attack	Gradient-based	Anchoring-based	EAB-FL
Time (s)	1.92	19.13	4.7	**3.36**

Table 3: Time Consumption Analysis.

model utility, irrespective of the use of secure aggregations.

Comparison with Other Poisoning Attacks. To show the attack's unique impact on the model's fairness and utility, we compare EAB-FL with various state-of-the-art poisoning attacks (i.e., DeSMP [Hossain *et al.*, 2021], MPAF [Cao and Gong, 2022], Sign Flipping [Li *et al.*, 2019a], DYN-OPT [Shejwalkar and Houmansadr, 2021], and BDB [Shejwalkar *et al.*, 2022]) in FL on the CelebA dataset. As shown in Table 2, we observe that our attack achieves the highest overall impact on the model's fairness (EDO and DPD are as high as 0.45 and 0.50, respectively), and the model's utility is still at a relatively high level (i.e., over 83%). This indicates our attack's effectiveness in exacerbating group unfairness while preserving the utility of the global model. Conversely, the other poisoning attacks all have a negligible impact on fairness, i.e., less than 0.32 and 0.31 in EOD and DPD, while degrading utility to as low as 49%, mainly due to the non-uniform changes in local data or model.

Time Complexity Analysis. Table 3 shows the average time required to successfully attack the global model per communication round on the CelebA dataset using an Nvidia Quadro A100 GPU. The results show that EAB-FL needs a slightly higher computation time compared to the no-attack (FEDAVG) scenario. To demonstrate EAB-FL's feasibility on lower computational power devices like smartphones and laptops, we estimated the required time by comparing their GPUs' FLOPS. The A100 GPU, chips used in recent Apple smartphones (e.g., Apple A17 Bionic), and commonly used chips in laptops (e.g., Intel Core i7 13700) have FLOPS ratings of 9.7 TFLOPS [Nvidia, 2022], 2.15 TFLOPS [Cpu-Monkey, 2024a], and 0.82 TFLOPS [Cpu-Monkey, 2024b], respectively. The estimated consumed time of EAB-FL on smartphones is 15.16 seconds and 39.74 seconds on laptops, which confirms the feasibility of launching EAB-FL on these edge devices.

6 Conclusion

In this work, we propose a new type of model poisoning attack, EAB-FL, in FL settings, with a focus on exacerbating group unfairness while maintaining a good level of model utility. The effectiveness and efficiency of the proposed attack are demonstrated through extensive experiments on three datasets in various FL settings. The results of this study highlight the importance of fully understanding the attack surfaces of current FL systems and the need for corresponding mitigations to improve their resilience against such attacks.

References

[Abay *et al.*, 2020] Annie Abay, Yi Zhou, Nathalie Baracaldo, Shashank Rajamoni, Ebube Chuba, and Heiko Ludwig. Mitigating bias in federated learning. *arXiv preprint arXiv:2012.02447*, 2020.

[Bach *et al.*, 2015] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[Bagdasaryan *et al.*, 2020] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International conference on artificial intelligence and statistics*, pages 2938–2948. PMLR, 2020.

[Bhagoji *et al.*, 2019] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. Analyzing federated learning through an adversarial lens. In *International Conference on Machine Learning*, pages 634–643. PMLR, 2019.

[Cao and Gong, 2022] Xiaoyu Cao and Neil Zhenqiang Gong. Mpaf: Model poisoning attacks to federated learning based on fake clients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3396–3404, 2022.

[Chhabra *et al.*, 2023] Anshuman Chhabra, Peizhao Li, Prasant Mohapatra, and Hongfu Liu. Robust fair clustering: A novel fairness attack and defense framework. *International Conference on Learning Representations (ICLR)*, 2023.

[Chu *et al.*, 2021] Lingyang Chu, Lanjun Wang, Yanjie Dong, Jian Pei, Zirui Zhou, and Yong Zhang. Fedfair: Training fair models in cross-silo federated learning. *arXiv preprint arXiv:2109.05662*, 2021.

[Cpu-Monkey, 2024a] Cpu-Monkey. Apple A17 Pro (6 GPU Cores) Benchmark, Test and specs — cpu-monkey.com. https://www.cpu-monkey.com/en/igpu-apple_a17_pro_6_gpu_cores, 2024. [Accessed 17-01-2024].

[Cpu-Monkey, 2024b] Cpu-Monkey. Intel Core i7-13700 Benchmark, Test and specs — cpu-monkey.com. https://www.cpu-monkey.com/en/cpu-intel_core_i7_13700, 2024. [Accessed 17-01-2024].

[Du *et al.*, 2021] Wei Du, Depeng Xu, Xintao Wu, and Hanghang Tong. Fairness-aware agnostic federated learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 181–189. SIAM, 2021.

[Dua and Graff, 2017] Dheeru Dua and Casey Graff. Uci machine learning repository. https://archive.ics.uci.edu/ml/datasets/adult, 2017. [Online; accessed 17-December-2021].

[Dwork *et al.*, 2012] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[Ezzeldin *et al.*, 2023] Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A Salman Avestimehr. Fairfed: Enabling group fairness in federated learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7494–7502, 2023.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[Harper and Konstan, 2015] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

[Hossain *et al.*, 2021] Md Tamjid Hossain, Shafkat Islam, Shahriar Badsha, and Haoting Shen. Desmp: Differential privacy-exploited stealthy model poisoning attacks in federated learning. In *2021 17th International Conference on Mobility, Sensing and Networking (MSN)*, pages 167–174. IEEE, 2021.

[Jacot *et al.*, 2018] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[Konečný *et al.*, 2016] Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtarik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. In *NIPS Workshop on Private Multi-Party Machine Learning*, 2016.

[Li *et al.*, 2019a] Liping Li, Wei Xu, Tianyi Chen, Georgios B Giannakis, and Qing Ling. Rsa: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1544–1551, 2019.

[Li *et al.*, 2019b] Xilai Li, Yingbo Zhou, Tianfu Wu, Richard Socher, and Caiming Xiong. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In *International Conference on Machine Learning*, pages 3925–3934. PMLR, 2019.

[Li *et al.*, 2020] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. *International Conference on Learning Representations (ICLR)*, 2020.

[Li *et al.*, 2021a] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pages 6357–6368. PMLR, 2021.

[Li *et al.*, 2021b] Xingyu Li, Zhe Qu, Shangqing Zhao, Bo Tang, Zhuo Lu, and Yao Liu. Lomar: A local defense against poisoning attack on federated learning. *IEEE Transactions on Dependable and Secure Computing*, 2021.

[Li *et al.*, 2022] Henger Li, Xiaolin Sun, and Zizhan Zheng. Learning to attack federated learning: A model-based reinforcement learning attack framework. *Advances in Neural Information Processing Systems*, 35:35007–35020, 2022.

[Liu *et al.*, 2018] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

[Liu *et al.*, 2023] Tao Liu, Zhi Wang, Hui He, Wei Shi, Liangliang Lin, Ran An, and Chenhao Li. Efficient and secure federated learning for financial applications. *Applied Sciences*, 13(10):5877, 2023.

[Lyu *et al.*, 2020] Lingjuan Lyu, Xinyi Xu, Qian Wang, and Han Yu. Collaborative fairness in federated learning. In *Federated Learning*, pages 189–204. Springer, 2020.

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[Mehrabi *et al.*, 2021] Ninareh Mehrabi, Muhammad Naveed, Fred Morstatter, and Aram Galstyan. Exacerbating algorithmic bias through fairness attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8930–8938, 2021.

[Mohri *et al.*, 2019] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR, 2019.

[Nvidia, 2022] Nvidia. NVIDIA A100 GPUs Power the Modern Data Center — nvidia.com. https://www.nvidia.com/en-us/data-center/a100/, 2022. [Accessed 17-01-2024].

[Panda *et al.*, 2022] Ashwinee Panda, Saeed Mahloujifar, Arjun Nitin Bhagoji, Supriyo Chakraborty, and Prateek Mittal. Sparsefed: Mitigating model poisoning attacks in federated learning with sparsification. In *International Conference on Artificial Intelligence and Statistics*, pages 7587–7624. PMLR, 2022.

[Pillutla *et al.*, 2022] Krishna Pillutla, Sham M Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022.

[Rauniyar *et al.*, 2023] Ashish Rauniyar, Desta Haileselassie Hagos, Debesh Jha, Jan Erik Håkegård, Ulas Bagci, Danda B Rawat, and Vladimir Vlassov. Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions. *IEEE Internet of Things Journal*, 2023.

[Rodríguez-Gálvez *et al.*, 2021] Borja Rodríguez-Gálvez, Filip Granqvist, Rogier van Dalen, and Matt Seigel. Enforcing fairness in private federated learning via the modified method of differential multipliers. In *NeurIPS Workshop*, 2021.

[Shejwalkar and Houmansadr, 2021] Virat Shejwalkar and Amir Houmansadr. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. In *NDSS*, 2021.

[Shejwalkar *et al.*, 2022] Virat Shejwalkar, Amir Houmansadr, Peter Kairouz, and Daniel Ramage. Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1354–1371. IEEE, 2022.

[Solans *et al.*, 2021] David Solans, Battista Biggio, and Carlos Castillo. Poisoning attacks on algorithmic fairness. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 162–177. Springer, 2021.

[Wang *et al.*, 2022] Jialu Wang, Xin Eric Wang, and Yang Liu. Understanding instance-level impact of fairness constraints. In *International Conference on Machine Learning*, pages 23114–23130. PMLR, 2022.

[Xie *et al.*, 2020] Chulin Xie, Keli Huang, Pin Yu Chen, and Bo Li. Dba: Distributed backdoor attacks against federated learning. In *8th International Conference on Learning Representations, ICLR 2020*, 2020.

[Yin *et al.*, 2018] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pages 5650–5659. PMLR, 2018.

[Yue *et al.*, 2023] Xubo Yue, Maher Nouiehed, and Raed Al Kontar. Gifair-fl: A framework for group and individual fairness in federated learning. *INFORMS Journal on Data Science*, 2(1):10–23, 2023.

[Zeng *et al.*, 2022] Yuchen Zeng, Hongxu Chen, and Kangwook Lee. Improving fairness via federated learning. *arXiv preprint arXiv:2110.15545*, 2022.

[Zhang *et al.*, 2017] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[Zhang *et al.*, 2020] Daniel Yue Zhang, Ziyi Kou, and Dong 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.

[Zhang *et al.*, 2022] Zaixi Zhang, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Fldetector: Defending federated learning against model poisoning attacks via detecting malicious clients. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2545–2555, 2022.

[Zhong *et al.*, 2022] Zhengyi Zhong, Weidong Bao, Ji Wang, Xiaomin Zhu, and Xiongtao Zhang. Flee: A hierarchical federated learning framework for distributed deep neural network over cloud, edge, and end device. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(5):1–24, 2022.

[Zhou *et al.*, 2021] Xingchen Zhou, Ming Xu, Yiming Wu, and Ning Zheng. Deep model poisoning attack on federated learning. *Future Internet*, 13(3):73, 2021.