FMU: Fair Machine Unlearning via Distribution Correction

Anonymous authors
Paper under double-blind review

Abstract

Machine unlearning, a technique used to remove the influence of specific data points from a trained model, is often applied in high-stakes scenarios. While most current machine unlearning methods aim to maintain the performance of the model after removing requested data traces, they may inadvertently introduce biases during the unlearning process. This raises the question: Does machine unlearning actually introduce bias? To address this question, we evaluate the fairness of model predictions before and after applying existing machine unlearning approaches. Interestingly, our findings reveal that the model after unlearning can exhibit a greater bias. To mitigate the bias induced by unlearning, we developed a novel framework, Fair Machine Unlearning (FMU), which ensures group fairness during the unlearning process. Specifically, for privacy preservation, FMU first withdraws the model updates of the batches containing the unlearning requests. For debiasing, it then deletes the model updates of sampled batches that have reversed sensitive attributes associated with the unlearning requests. To validate the effectiveness of FMU, we compare it with standard machine unlearning baselines and one existing fair machine unlearning approach. FMU demonstrates superior fairness in predictions while maintaining privacy and comparable prediction accuracy to retraining the model. Furthermore, we illustrate the advantages of FMU in scenarios involving diverse unlearning requests, encompassing various data distributions of the original dataset. Our framework is orthogonal to specific machine unlearning approaches and debiasing techniques, making it flexible for various applications. This work represents a pioneering effort, serving as a foundation for more advanced techniques in fair machine unlearning.

1 Introduction

Building responsible machine learning algorithms for protecting data privacy (Jain et al., 2016; Chen et al., 2021) is an ever-present concern that holds significant importance within both academia and industry. In recent years, significant progress has been made in the development of vital legal regulations for protecting personal data. These include the General Data Protection Regulation (GDPR), a landmark initiative by the European Union (Commission, 2018), the California Consumer Privacy Act (CCPA) (DoJ, 2018), the comprehensive Personal Information Protection and Electronic Documents Act, and the Brazilian General Data Protection Law. Of particular significance is the concept of the "right to be forgotten" (RTBF), introduced as a fundamental component of the GDPR (Rosen, 2011). The RTBF mandates organizations and institutions to promptly and diligently erase personal data upon a user's request. This provision underscores the commitment to individual privacy rights (Viorescu et al., 2017). In the context of machine learning (ML), the RTBF translates

Table 1: Utility and fairness were evaluated before (fair training) and after applying machine unlearning (Amnesiac) on the `Adult` and `ACS-I` datasets, with gender as the sensitive attribute. For unlearning, 30% of data randomly sampled from unprivileged groups (females positive on `Adult` and males positive on `ACS-I`) was removed. The last column shows a performance drop, indicating increased bias after unlearning.

	Metric	Fair Training	Amnesiac	Drop (%)
`Adult`	Acc. (↑)	80.5±0.1	78.1±1.8	2.98
	AUC (↑)	86.5±0.8	80.3±2.8	7.17
	$\triangle_{EO}$ (↓)	2.6±0.1	22.3±5.2	757.69
	$\triangle_{DP}$ (↓)	1.6±0.9	30.8±9.4	1825.00
`ACS-I`	Acc. (↑)	77.4±1.4	76.7±1.9	0.90
	AUC (↑)	85.2±1.7	79.0±1.1	7.28
	$\triangle_{EO}$ (↓)	2.3±0.6	28.8±7.9	1152.17
	$\triangle_{DP}$ (↓)	2.1±0.3	23.9±4.3	1038.09

into the obligation for model providers to effectively eliminate both the "information" pertaining to the data and its corresponding "impact" from ML models. This process, known as "machine unlearning" (MU), presents an intriguing challenge to researchers.

With the rapid evolution of machine unlearning applications, including recommender systems (Chen et al., 2022a; Li et al., 2023; Zhang et al., 2023c; Xu et al., 2023b; Liu et al., 2023), graph learning (Chen et al., 2022b; Cheng et al., 2023; Cong & Mahdavi, 2022; Klicpera et al., 2019), and federated learning (Liu et al., 2021; 2022; Yuan et al., 2023), extensive efforts have been made towards developing MU framework to achieving the RTBF compliance. While much attention is given to privacy concerns in these investigations, there is a lack of focus on *trustworthy* issues of MU, particularly the fairness (Mehrabi et al., 2021; Dwork et al., 2012). Only one literature exists that studies the fairness of MU frameworks (Zhang et al., 2023a). It concludes that when data deletion is non-uniform, certain MU methods outperform others in achieving fairness.

In light of this observation, we pioneer through a more diverse empirical investigation to reveal the relationship between machine unlearning and fairness. Our exploration has revealed a significant concern: machine unlearning poses a considerable risk of causing a decline in fairness, as exemplified in Table 1. This potential discrimination within a machine learning model can give rise to significant social and ethical issues, thus limiting the applicability of the model after unlearning in real-world applications, including healthcare (Ahmad et al., 2020; Chen et al., 2018), job recruitment (Mehrabi et al., 2021), credit scoring (Kozodoi et al., 2022), and criminal justice (Berk et al., 2021). Hence, there is a pressing need to develop a fair machine unlearning approach that remains free from biases after unlearning (Li et al., 2021; Mehrabi et al., 2021; Barocas et al., 2017). This gives rise to a key question that anchors our study:

Can machine unlearning both ensure fairness and keep prediction accuracy?

The primary challenge of this study is the input to the MU framework is a trained model. In contrast, many advanced debiasing approaches, such as in-processing methods, depend on solving an objective function that incorporates a debiasing module during the training process. Moreover, specific machine unlearning frameworks rely on distributed training, so they need to consider both local and global fairness concerns. For instance, directly applying debiasing techniques to models trained on different data subsets does not guarantee global fairness, especially when predictions from these subsets are subsequently aggregated (e.g., through voting) for decision-making purposes. An alternative approach could involve centralized fair retraining after unlearning, but it contradicts the fundamental purpose of unlearning, which is to circumvent the cost of retraining on the remaining data.

To tackle these challenges, we introduce FMU, designed to ensure fairness and privacy while maintaining the accuracy of predictions, which is crucial for downstream tasks. One key benefit of FMU is its simplicity in implementation, requiring just a few extra lines of code to perform debiasing for a machine unlearning approach. FMU works efficiently in two steps. First, it removes the model updates linked to the data batches marked for unlearning. Then, it compensates by offsetting a similar number of model updates from the opposite-sensitive data groups within the trained model. The overview framework of FMU is illustrated in Figure 1. This method allows FMU to ensure that model parameters are balanced across all sensitive groups from a global view. Our main contributions are summarized below:

- We pioneer an investigation to reveal the impact of machine unlearning on group fairness through the following observations: 1) both exact machine unlearning and approximate unlearning methods can make a fair model unfair; 2) the distribution shift of the model weights is the cause of the bias after the unlearning request has been executed by the unlearning methods.
- To the best of our knowledge, this is one of the first attempts at studying the fair machine unlearning problems. We propose a novel framework, FMU which provides an extremely simple yet effective solution that is adaptive to any trained ML model.
- Comprehensive experimental results on five real-world datasets validate FMU in achieving fairness and privacy while maintaining utility, i.e., high prediction accuracy in prediction tasks.
- FMU is straightforward in implementation and does not require any additional storage beyond a standard machine unlearning. Besides, it is agnostic to the ML models and debiasing methods, which makes it versatile in practice.

2 Fair Machine Unlearning

In this section, we first offer background knowledge, including existing techniques for machine unlearning as well as fairness in machine learning. Then, we present our approach FMU, which involves two key steps. Firstly, we detail the unlearning process for the specific data that the users request to unlearn. Secondly, we explain the debiasing step, which involves the removal of an equal amount of model updates associated with data exhibiting the same label but having an opposite sensitive attribute.

Problem Statement of Machine Learning. Suppose we have a training dataset with N data samples, denotes as $D = \{\mathbf{X}, \mathbf{Y}\} = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)\} \in \mathcal{D}$, where $\mathcal{D}$ is the input space with dimension $\mathbb{R}^{N \times d}$, $\mathbf{x}_i \in \mathbf{X}$ is a feature vector of dimension d; $\mathbf{y}_i \in \mathbf{Y} = \{0, 1\}$ is a binary label; and $\mathbf{s}_i \in \mathbf{S} = \{0, 1\}$ is a binary sensitive attribute. A *machine learning* problem can be formulated as a mapping $M : \mathcal{D} \to \mathcal{H}$ that maps a dataset D to a model $M(D)$ in a hypothesis space $\mathcal{H}$.

2.1 Machine Unlearning

The goal of machine unlearning is to deal with the requests invoking the "right to be forgotten". Given a set of samples that are requested to be unlearned $D_u \subseteq D$, machine unlearning aims to efficiently find a model that resembles the model retrained on $D \backslash D_u$ from scratch. We define an unlearning mechanism as a mapping U: $\mathcal{H} \times \mathcal{D} \times \mathcal{D} \to \mathcal{H}$ that maps a learned model $M(D)$, the original dataset D, and a subset $D_u \subseteq D$ of data points to be deleted, to a new unlearned model $U(M(D), D, D_u)$. Ideally, the unlearned model $U(M(D), D, D_u)$ is statistically indistinguishable from $M(D \backslash D_u)$. Both the learning algorithm M and the unlearning mechanism U are assumed to be randomized, i.e., their outputs produce a probability distribution over the hypothesis $\mathcal{H}$ given the input.

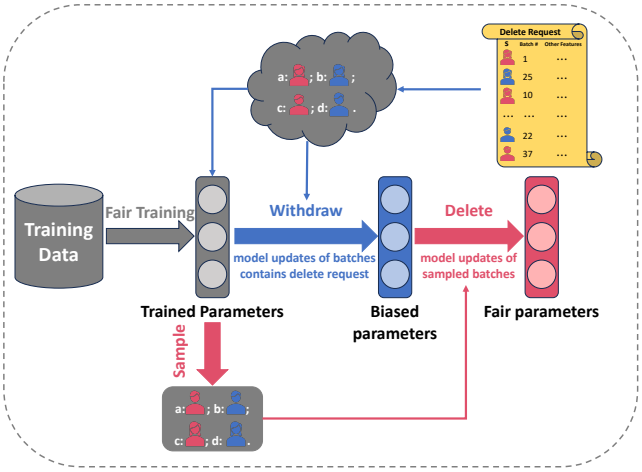


Figure 1: The framework of the proposed FMU. When the unlearning request comes, FMU withdraws the model updates of batches that contain the data that the users request to unlearn, and then deletes the model updates of batches that have the data with the reversed sensitive attribute of the unlearning request.

There are two primary approaches to machine unlearning: exact and approximate machine unlearning (Nguyen et al., 2022; Zhang et al., 2023b). The machine unlearning problem can be formulated as the comparison between two distributions of ML models. Let $\Pr(M(D))$ denote the distribution of a model trained on a dataset D by a learning algorithm $M(\cdot)$. Let $\Pr\left(U\left(D, D_u, M(D)\right)\right)$ be the distribution of a unlearned model. The rationale behind representing the output of $U(\cdot)$ as a distribution rather than a single point is the learning algorithms $M(\cdot)$ and unlearning algorithms $U(\cdot)$ are randomized, as discussed previously.

2.1.1 Exact Unlearning:

In exact unlearning, requested data information is directly removed from the training set, achieved by retraining models on the remaining data subsets after deleting the requested data (Bourtoule et al., 2021; Cao & Yang, 2015b; Ginart et al., 2019; Bourtoule et al., 2020). The distributions of the unlearned model should be indistinguishable from the original model to ensure attackers cannot recover any information from the unlearned model (Xu et al., 2023a). The definition can be formulated as follows.

Definition 2.1 (Exact Unlearning). Given a learning algorithm $M(\cdot)$, we say the process $U(\cdot)$ is an exact unlearning process iff $\forall \mathcal{T} \subseteq \mathcal{H}, D \in \mathcal{D}, D_u \subset D$:

$$\Pr\left(M\left(D \backslash D_u\right) \in \mathcal{T}\right) = \Pr\left(U\left(D, D_u, M(D)\right) \in \mathcal{T}\right). \tag{1}$$

One representative method is SISA (Bourtoule et al., 2020). This is an exact machine unlearning method aiming to reduce the computational cost of the retraining process by employing a data partitioning technique.

2.1.2 Approximate Unlearning:

In contrast, approximate unlearning adjusts model parameters to remove the impact of deleted data (Guo et al., 2019; Izzo et al., 2021; Neel et al., 2021; Thudi et al., 2022; Graves et al., 2021).

2.2 Definition (ϵ-Approximate Unlearning). Given $\epsilon > 0$, an unlearning mechanism U performs ϵ unlearning for a learning algorithm M if $\forall \mathcal{T} \subseteq \mathcal{H}, D \in \mathcal{D}, \mathbf{x} \in \mathbf{X}$:

$$e^{-\epsilon} \leq \frac{\Pr(U(D, \mathbf{x}, M(D)) \in \mathcal{T})}{\Pr(M(D \backslash \mathbf{x}) \in \mathcal{T})} \leq e^{\epsilon}, \tag{2}$$

where $\mathbf{x}$ is the removed sample. In our experiment, we compare our method with a typical approximate unlearning method, Amnesiac (Graves et al., 2021). It leverages batch training characteristics and records the updated parameters of a model for each batch in storage.

2.2 Group Fairness

The literature on group fairness has put forth numerous concepts of fairness. These definitions, each target distinct statistical measures to achieve equilibrium among subgroups within the data. We consider Equalized Opportunity and Demographic Parity as the fairness metrics in this paper.

Equal Opportunity (EO). Equal opportunity (EO) (Hardt et al., 2016) requires the classifier to maintain equal true positive rates across different subgroups, aiming for a perfect classifier. The fairness measurement corresponding to EO can be expressed as follows:

$$\triangle_{EO} = |P(\hat{\mathbf{y}} = 1 | \mathbf{y} = 1, \mathbf{s} = 0) - P(\hat{\mathbf{y}} = 1 | \mathbf{y} = 1, \mathbf{s} = 1)|. \tag{3}$$

A low $\triangle_{EO}$ indicates that the probability of an instance in a positive class being assigned to a positive outcome for both subgroup members is relatively small.

Demographic Parity. Demographic Parity (DP) (Zafar et al., 2017; Feldman et al., 2015) requires the prediction $\hat{\mathbf{y}}$ to be independent with the sensitive attribute $\mathbf{s}$, i.e., $\hat{\mathbf{y}} \perp \mathbf{s}$. The majority of the literature focuses on binary classification and binary attributes, i.e., $\mathbf{y} \in \{0, 1\}$ and $\mathbf{s} \in \{0, 1\}$. Similar to equal opportunity, the fairness in terms of DP can be measured by:

$$\triangle_{DP} = |P(\hat{\mathbf{y}} = 1 | \mathbf{s} = 0) - P(\hat{\mathbf{y}} = 1 | \mathbf{s} = 1)|. \tag{4}$$

A lower $\triangle_{DP}$ indicates a more fair classifier. Both the DP and EO can be easily extended to multi-class and multi-category sensitive attributes problems by ensuring $\hat{\mathbf{y}} \perp \mathbf{s}$.

2.3 The FMU Framework

Training. Recall that in the original fairness-aware training (Zliobaite, 2015; Calders et al., 2009), most of the existing models can be summarized as a learning process:

$$\min_{\theta} \mathcal{L} = \min_{\theta} (\mathcal{L}_{utility} + \alpha \mathcal{L}_{fairness}), \tag{5}$$

where θ represents the learned parameters, $\mathcal{L}_{utility}$ denotes the loss function for the utility, $\mathcal{L}_{fairness}$ corresponds to the applied fairness regularize to achieve the group fairness, defined in Section 2.2, `DiffDP` (Chuang & Mroueh, 2020), and α controls the trade-off between utility and fairness:

$$\theta_M = \theta_{initial} + \sum_{p=1}^{P} \sum_{b=1}^{B} \Delta_{\theta_{p,b}}, \tag{6}$$

where $\Delta_{\theta_{p,b}}$ is the model update for the epoch p batch b, and P stands for the number of epochs, while B corresponds to the total count of batches within each epoch. θ_{initial} denotes the initial parameters of the ML model. This displays the accumulation of parameter changes across epochs and batches, contributing to the model training progression.

During the training process, a record detailing which specific batches including each individual sample is maintained. This record can take various forms, such as an index correlating batches with each example in the training dataset, an index associating batches with respective classes, or any other format. As a result, this record includes the batch index of the data instances that users request for unlearning, denoted as B_u. Furthermore, upholding the updates made to the model parameters from each batch containing sensitive data is important. However, a comprehensive strategy involves detailed documentation of every single batch update, and a more targeted approach can be adopted. If the user's concern is primarily on a subset of data, then only the parameter updates in the batches that correspond to this subset need to be obtained by retraining on the remaining data.

Algorithm 1: Algorithm FMU

Input: Trained model parameters θ_M, a unlearning request list B_u.
(1) Unlearning stage:
Initialize $\theta_{M'} = \theta_M$.
for $u \in B_u$ **do**
  $\theta_{M'} = \theta_{M'} - \Delta_{\theta_u}$.
(2) Debiasing stage:
Sample a list $R \in B \backslash B_u$, with $|R| = n'_{1-\mathbf{s},\mathbf{y}} = |B_u| = n_{\mathbf{s},\mathbf{y}}$
Initialize $\theta_{M'_{fair}} = \theta_{M'}$.
for $r \in R$ **do**
  $\theta_{M'_{fair}} = \theta_{M'_{fair}} - \Delta_{\theta_r}$.
Return: $\theta_{M'_{fair}}$.

Unlearning. After the training procedure is finished, the unlearned model M' can be generated by deleting the parameter updates of each batch u in list B_u from the original model θ_M (Graves et al., 2021):

$$\theta_{M'} = \theta_{initial} + \sum_{p=1}^{P} \sum_{b=1}^{B} \Delta_{\theta_{p,b}} - \sum_{u \in B_u} \Delta_{\theta_u} = \theta_M - \sum_{u \in B_u} \Delta_{\theta_u}. \tag{7}$$

Debiasing. FMU starts by sampling a batch list R from the remaining data in $B \backslash B_u$. This step guarantees the sampled batches excluding any requested user information, thus preventing potential privacy leakage. The sampling for R follows the rules below.

Suppose the unlearning batch list B_u includes $n_{\mathbf{s},\mathbf{y}}$ samples for each combination of sensitive attribute $\mathbf{s} \in \{0, 1\}$ and label $\mathbf{y} \in \{0, 1\}$. For each group defined by $\mathbf{s}$ and $\mathbf{y}$ in B_u, we include in R an equal number of samples with the opposite sensitive attribute but the same outcome as those in B_u. In other words, if B_u has n samples for a particular combination of $\mathbf{s}$ and $\mathbf{y}$, then R will also contain n samples with the opposite $\mathbf{s}$, i.e., $1 - \mathbf{s}$, but the same $\mathbf{y}$. In mathematical notation, we set $n'_{1-\mathbf{s},\mathbf{y}} = n_{\mathbf{s},\mathbf{y}}$, where n and n' represent the number of samples in B_u and R, respectively, and $\mathbf{s}$ and $\mathbf{y}$ can take values of either 0 or 1.

To ensure fairness in the prediction after unlearning, we need to balance the effects of removing samples from B_u. Hence, from the remaining data in $B \backslash B_u$, we delete the model updates of batches in the list R to offset the effect of the unlearning from B_u. The fair model after the unlearning can be obtained by:

$$\theta_{M'_{fair}} = \theta'_M - \sum_{r \in R} \Delta_{\theta_r}. \tag{8}$$

Finally, the resulting model after unlearning maintains fairness and accuracy, by balancing the model updates after the removal of requested data. The algorithm of FMU is summarized in Algorithm 1.

Discussion. The process of unlearning can cause a fair-trained model to become unfair since it changes the distribution of sensitive groups. One way to tackle this is by excluding *updates related to fairness loss* and only keeping those that improve utility for the sensitive group that was removed. This helps balance out the distribution of sensitive groups after data removal. However, this approach has an extra need for storage to keep fairness and utility updates *separately*. To tackle this problem, we suggest a different approach: deleting a similar number of updates from the opposing sensitive group. This is similar to a technique called upsampling mentioned in previous research (Visa & Ralescu, 2005; Tran et al., 2022). One advantage of this method is that it requires the same amount of storage as the standard unlearning method. It can also be extended to handle multiple sensitive attributes and cases with multiple classes.

3 Experiments

In this section, we mainly investigate the performance of FMU from three perspectives: **1) the machine unlearning performance, 2) the utility, i.e., task-specific prediction accuracy, and 3) the fairness performance, under different experimental settings.** In particular, we examine whether the final model is able to protect data privacy against attacks, and how the prediction and fairness performance changes with different unlearning ratios and sensitive groups. Besides, we visualize a comparison of the prediction distribution of unlearning baselines across sensitive groups. Finally, we evaluate how the trade-off between fairness and accuracy changes with the controlling hyperparameter α in Equation (5).

3.1 Datasets

We conducted experiments using five widely used fairness datasets to assess the fairness and utility of machine unlearning baselines. These include four tabular datasets: `Adult` (Kohavi & Becker, 1996), `COMPAS` (Larson et al., 2016), `ACS-I`, and `ACS-T` (Ding et al., 2021), as well as one image dataset, `CelebA-A` (Liu et al., 2015). Following the previous work (Biswas & Rajan, 2020; Chakraborty et al., 2020; Zhang & Harman, 2021), in our experiment, we select gender and race for the sensitive attributes. Further information on the datasets is provided in Appendix E.1, and the detailed statistics are summarized in Table 11.

3.2 Baselines

For the baselines, we consider fairness-aware machine learning, standard machine unlearning, and existing fair machine unlearning approaches. The introduction of the standard machine unlearning baselines is as follows:

- **Retraining (Fair)** is conducted by integrating a fairness regularizer into the standard retraining in our experiment. The standard retraining, referred to as 'Retraining (Standard)' in this paper, is widely recognized as the benchmark for machine unlearning literature (Zhang et al., 2023b; Xu et al., 2023a; Nguyen et al., 2022). It involves retraining the model on the remaining data following the removal of the requested information from the original dataset. Its privacy-preserving performance serves as the ceiling baseline. However, retraining comes with a significant cost, motivating the exploration of alternative machine unlearning techniques aimed at circumventing this expense.
- **SISA (Bourtoule et al., 2020)** is a representative exact machine unlearning approach. Initially, the original dataset is divided into subsets called shards. Each shard is then split into several slices. A deep learning (DL) model is created for each shard, and its refinement involves progressively adding slices. The parameters of a DL model are all retained in storage. The final outputs are generated by aggregating the outputs of DL models. For addressing requests for removing data, SISA automatically identifies the relevant shards and slices and retrains the corresponding DL models from a cached stage before the deleted data slices are incorporated.
- **Amnesiac (Graves et al., 2021)** is a typical approximate machine unlearning approach. When the RTBF requests arrive, Amnesiac automatically locates the batches containing the instances that need to be deleted. Then, the parameters of the DL model are rolled back to remove the impact of the deleted data on the trained DL model. However, Amnesiacdeletes the model update directly, which will lead to a distribution shift, further amplifying the bias within the model.
- **Oesterling (Oesterling et al., 2023)** is a recently proposed fair machine unlearning method that utilizes a convex fairness regularizer with pairwise comparisons and unrolls the comparisons in an unlearnable form. However, the assumption of convex fairness regularizer limits their application for many commonly used fairness regularizations including a gap regularization, which is often non-convex (Chuang & Mroueh, 2020), including demographic parity, equalized opportunity, and equalized odds, making it impracticable for Oesterling to directly optimize them.

3.3 Experimental Setup

Data unlearning strategies. To simulate the "*right to be forgotten*" (RTBF) requests in the real world, we consider five types of data unlearning requests for each dataset, including four sensitive groups, and a mixed

Table 2: Utility and fairness performance on `Adult`, `COMPAS`, `ACS-I`, `ACS-T`, and `CelebA-A` before (original training) and after unlearning. For standard unlearning, retraining, and fair unlearning, the unlearning ratio is 30% if otherwise stated. We use Amne., Acc., Sens. Attr. to represent Amnesiac, accuracy, and sensitive attribute for short, correspondingly.

	Sens. Attr.	Metric	Original Training		Standard Unlearning				Retraining		Fair Unlearning	
			Standard	Fair	SISA (10%)	SISA	Amne. (10%)	Amne.	Standard	Fair	Oesterling	FMU
Adult	Race	Acc. (↑)	84.4±0.3	80.1±0.7	77.3±0.9	76.3±1.2	78.9±1.2	76.7±1.3	82.3±0.5	80.4±0.5	77.3±2.6	78.4±1.3
		AUC (↑)	90.1±0.5	85.3±0.5	82.3±0.7	78.4±0.4	80.3±1.7	77.1±1.7	87.2±1.9	83.2±1.3	80.4±2.3	80.9±1.6
		$\triangle_{EO}$ (↓)	9.25±3.8	1.25±1.0	13.5±2.4	19.8±5.2	15.3±3.4	29.4±7.1	10.6±5.6	1.32±0.2	5.82±2.1	1.85±0.5
		$\triangle_{DP}$ (↓)	13.4±0.8	0.95±0.6	16.4±3.1	21.4±2.6	9.45±4.5	26.2±6.3	15.2±4.8	1.54±0.4	6.37±3.2	1.74±0.8
	Gender	Acc. (↑)	84.6±0.3	80.5±0.1	78.4±1.5	76.4±1.1	78.6±1.3	76.1±1.8	82.1±0.7	79.4±1.7	78.4±0.7	78.9±0.2
		AUC (↑)	90.8±0.2	86.5±0.8	82.4±2.1	79.3±2.1	80.2±2.2	79.3±2.8	88.4±1.2	82.1±0.8	79.3±1.9	81.2±0.5
		$\triangle_{EO}$ (↓)	8.43±3.2	2.64±0.1	9.42±4.1	20.9±3.7	14.3±3.4	22.3±5.2	9.21±2.5	1.75±0.3	5.47±1.2	1.63±0.2
		$\triangle_{DP}$ (↓)	16.5±0.9	1.62±0.9	12.5±3.2	24.6±4.3	20.1±4.2	30.8±9.4	15.2±2.1	1.14±0.6	6.70±2.6	1.87±0.4
COMPAS	Race	Acc. (↑)	66.9±1.0	60.9±1.1	56.4±1.2	54.2±0.8	56.3±1.5	54.7±1.4	65.3±1.8	60.9±0.6	57.4±1.8	59.9±0.5
		AUC (↑)	72.4±0.8	69.4±1.8	63.4±1.7	61.1±1.9	64.7±1.8	61.9±2.3	72.1±2.1	67.2±1.5	64.4±1.7	66.5±1.1
		$\triangle_{EO}$ (↓)	19.4±4.6	0.91±0.3	16.8±3.6	19.5±4.2	20.4±2.2	27.6±3.3	16.3±6.2	1.02±0.7	7.35±3.8	1.24±0.3
		$\triangle_{DP}$ (↓)	17.2±4.1	1.37±1.1	26.5±7.7	34.2±10.2	23.1±6.6	29.0±8.4	18.4±5.4	1.05±0.3	9.86±2.3	1.93±0.6
	Gender	Acc. (↑)	66.8±0.7	62.1±1.5	57.3±2.0	55.4±1.9	56.7±1.2	54.9±2.6	65.4±1.6	62.6±1.5	59.3±1.4	60.1±0.8
		AUC (↑)	72.1±0.9	68.2±1.6	64.9±1.8	61.2±0.8	63.2±1.5	60.3±2.1	66.7±1.0	62.3±0.8	60.0±1.8	61.4±0.6
		$\triangle_{EO}$ (↓)	12.4±5.8	2.12±0.5	16.5±3.5	29.3±6.7	15.3±5.5	31.3±14.8	16.2±4.8	2.68±1.9	8.31±2.9	2.90±1.5
		$\triangle_{DP}$ (↓)	17.2±4.3	1.73±0.8	20.8±6.3	33.5±16.7	27.6±7.4	34.6±17.5	18.8±6.2	1.62±0.5	9.08±3.5	1.92±0.8
ACS-T	Race	Acc. (↑)	66.3±0.3	65.7±0.2	62.3±2.6	59.3±1.4	61.2±1.5	58.7±0.8	65.2±1.0	64.4±0.8	59.2±2.0	61.5±1.3
		AUC (↑)	72.6±0.2	71.2±0.2	68.2±1.4	65.3±1.9	68.0±2.3	64.2±2.4	70.3±1.6	68.2±2.4	65.4±2.1	67.8±1.1
		$\triangle_{EO}$ (↓)	6.93±1.2	0.86±0.5	10.3±7.5	17.4±9.5	11.6±6.7	15.4±8.1	7.4±2.4	1.01±0.3	5.31±2.8	1.53±0.6
		$\triangle_{DP}$ (↓)	10.2±1.6	2.83±0.7	15.7±12.5	18.2±7.4	16.7±8.8	14.3±9.9	11.4±2.4	2.32±1.1	8.12±3.3	1.82±0.5
	Gender	Acc. (↑)	66.2±0.4	65.9±0.4	60.8±2.5	57.2±2.1	61.0±2.4	58.9±2.7	64.3±0.7	63.5±0.8	60.5±1.9	62.7±0.7
		AUC (↑)	72.7±0.2	72.0±0.3	68.4±0.8	66.2±1.8	68.2±2.1	67.2±1.9	70.2±2.4	67.3±1.0	64.1±1.3	68.6±0.8
		$\triangle_{EO}$ (↓)	6.14±3.5	1.83±0.5	11.2±5.6	17.9±7.3	11.4±9.7	22.5±10.5	8.8±2.2	2.45±1.0	6.42±4.3	2.73±1.1
		$\triangle_{DP}$ (↓)	7.25±2.6	1.54±0.4	13.0±4.1	15.6±9.3	10.3±4.5	17.8±7.4	6.2±3.5	2.23±0.7	5.29±5.7	2.91±1.3
ACS-I	Race	Acc. (↑)	81.2±0.1	80.2±1.2	76.4±1.8	75.3±1.9	75.2±1.3	75.0±1.1	80.9±1.5	77.9±1.5	76.5±0.8	78.3±0.9
		AUC (↑)	90.1±0.1	87.0±1.6	83.2±2.0	82.5±2.4	83.5±2.5	81.5±1.6	89.7±2.1	86.6±1.1	84.1±1.6	85.2±1.0
		$\triangle_{EO}$ (↓)	7.42±0.6	1.96±0.5	14.6±5.2	23.5±4.4	17.9±5.1	25.3±4.0	9.07±1.4	1.23±0.7	7.64±2.2	1.36±0.2
		$\triangle_{DP}$ (↓)	10.0±1.8	1.66±0.2	18.4±1.5	26.3±7.7	14.3±4.7	23.3±6.2	11.3±3.8	1.02±0.4	3.42±1.8	0.92±0.4
	Gender	Acc. (↑)	82.0±0.2	80.9±0.6	76.3±2.0	75.4±2.2	77.3±0.8	76.7±1.9	81.1±0.6	78.3±1.2	75.3±0.8	78.0±0.8
		AUC (↑)	90.1±0.1	85.2±1.7	81.9±2.4	80.2±2.6	80.9±2.4	79.0±1.1	88.4±1.8	84.2±2.0	82.1±0.9	84.0±1.5
		$\triangle_{EO}$ (↓)	6.12±0.6	2.31±0.6	12.4±0.7	24.3±3.3	11.6±5.2	28.8±7.9	9.05±0.5	1.40±0.3	10.2±2.7	1.66±1.0
		$\triangle_{DP}$ (↓)	10.2±1.6	2.13±0.3	17.2±0.9	28.4±6.8	15.4±3.7	23.9±4.3	12.4±2.2	1.68±0.8	2.55±0.6	1.67±1.2
CelebA-A	Race	Acc. (↑)	78.1±0.6	75.0±1.9	71.3±1.7	68.3±0.8	69.2±1.2	66.7±0.4	77.2±0.9	76.4±0.4	74.2±2.0	76.5±0.8
		AUC (↑)	85.9±0.8	81.3±0.6	77.2±1.9	75.3±0.9	78.2±1.3	74.2±1.8	84.3±1.2	81.2±1.8	77.4±2.1	80.8±0.7
		$\triangle_{EO}$ (↓)	19.3±1.5	0.93±0.3	23.3±5.6	27.4±11.2	21.6±12.6	35.4±13.5	21.4±1.7	1.01±0.5	5.31±2.8	1.53±0.2
		$\triangle_{DP}$ (↓)	38.7±1.8	1.62±1.3	35.7±8.9	41.2±21.4	36.7±18.8	40.3±20.4	37.4±1.6	2.32±0.9	8.12±3.3	1.82±0.3
	Gender	Acc. (↑)	78.1±0.6	74.3±1.6	70.8±1.7	67.2±1.9	71.0±1.6	68.9±1.8	78.4±0.5	73.5±0.6	70.5±1.9	72.7±0.7
		AUC (↑)	85.9±0.8	81.7±1.2	77.4±1.2	76.2±2.1	78.2±1.9	77.2±1.4	84.2±0.8	82.3±1.2	80.1±1.3	81.6±0.8
		$\triangle_{EO}$ (↓)	35.6±2.1	1.12±0.4	33.2±15.6	37.9±21.3	31.4±8.4	42.5±20.5	37.8±26.7	2.45±0.7	12.4±4.3	2.73±0.6
		$\triangle_{DP}$ (↓)	50.6±2.6	2.21±0.5	39.0±14.1	45.6±19.3	40.3±24.3	47.8±27.4	53.2±24.3	2.23±0.6	15.2±5.7	1.81±0.5

unlearning. For example, if the sensitive attribute is gender, then the five sensitive groups are female positive, female negative, male positive, male negative, and mixed. In the context of privacy concerns, we focus on investigating the following three unlearning requests:

- *Unlearning from a Privileged Group:* It typically refers to those who historically have been more inclined to be categorized favorably in a binary classification task within machine learning. Privilege emerges from disparities in power dynamics, and it is important to note that the same groups may not universally enjoy privilege across all contexts, even within the same society (Varshney, 2019). This demographic often seeks to protect their personal data, particularly due to their extensive education, thus prioritizing the privacy of their sensitive information (Upton et al., 2001; eur, 2019).

- *Unlearning from an Unprivileged Group:* In contrast to the privileged group, this refers to a group less likely to receive favorable predictions. Previous studies have shown that unprivileged groups face higher privacy risks and costs for achieving fairness in machine learning models (Chang & Shokri, 2021; Strobel & Shokri, 2022). As a result, protecting their privacy and ensuring fairness becomes important.

- *Mixed Unlearning:* This involves unlearning requests from all sensitive groups and is the most common scenario. For this unlearning process, we implement random sampling from the entire demographic, ensuring that each sample has an equal chance of being selected for unlearning.

We test different types of unlearning requests by randomly selecting data at a specified ratio from the corresponding group or from the entire training dataset if it's a mixed group, simulating the unlearning request. We assess the performance of the model across unlearning ratios ranging from 10% to 50%.

Evaluation metrics. We consider two types of evaluation metrics: utility and fairness. Following the previous studies, for measuring the utility, the *accuracy* and *AUC (Area Under the ROC Curve)* are tested, and for evaluating the fairness, the $\triangle_{EO}$ and $\triangle_{DP}$ are considered, as introduced in Section 2.2.

Training setup. For implementing FMU, we use Pytorch (Paszke et al., 2017) with Adam Optimizer (Kingma & Ba, 2014). We set the initial learning rate to 0.01, and weight decay to 0. For tabular data, a 2-layer Multilayer Perceptron (MLP) with 256 hidden neurons is utilized. For image data, a 2-layer MLP with 128 hidden neurons is employed. The batch size used for all datasets is 32. The number of epochs for the training or unlearning process is 100, with a StepLR step of 0.1, and a StepLR gamma of 50. We use the binary cross entropy loss for calculating the utility loss.

3.4 Can FMU Surpass Existing MU Methods in Fairness Performance While Maintaining Strong Prediction Capabilities?

To validate the efficacy of the proposed method, we provide a detailed comparison of FMU with the aforementioned baselines on five datasets, which contain both tabular data and image data. For this experiment, we delete the unprivileged group on each dataset, e.g., female positive on `Adult`, and with 10% and 30% ratios and report both mean and standard deviation of 20 runs for all models across all datasets. Based on the results in Table 2, we can make the following observations:

Observation 1: Both exact machine unlearning, e.g., SISA, and approximate machine unlearning, e.g., Amnesiac, can introduce bias to a fairly trained model. On all five datasets in Table 2, unlearning 10% or 30% of unprivileged groups using SISA or Amnesiac will lead to more unbiased predictions compared to the standard training, in terms of Δ_{DP} and Δ_{EO}. Specifically, 1) Higher unlearning ratios, such as 30%, result in more bias in the model compared to lower ratios like 10%; 2) Removing data from an unbalanced dataset (`Adult`) induces more bias than from a balanced dataset (`ACS-I`) because unbalanced data is more sensitive to the removal of the unprivileged group; 3) Deleting data from a larger dataset leads to a smaller drop in accuracy compared to a smaller dataset, but FMU can approximate the retraining accuracy for both scenarios.

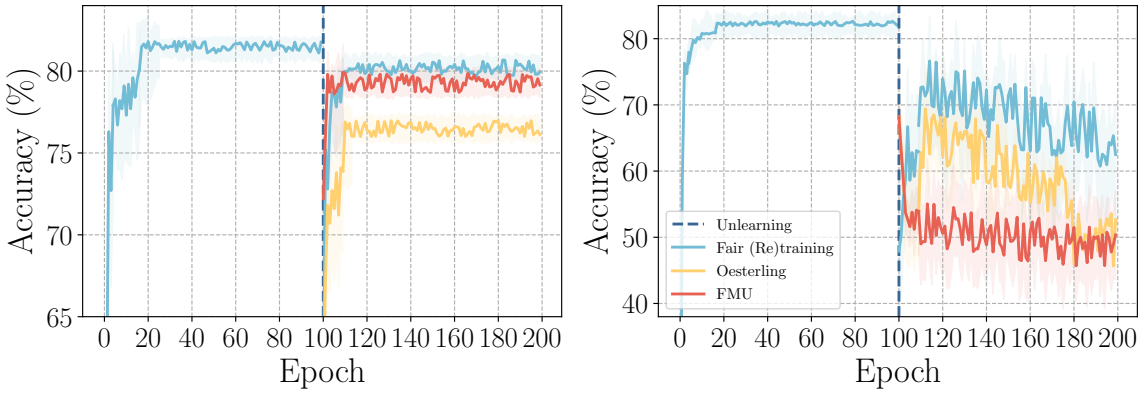
Observation 2: The proposed method demonstrates comparable accuracy to retraining while ensuring fairness. FMU shows comparable fairness and utility performance with fair retraining, and it does not require additional computational and storage costs. In contrast to a fair machine unlearning baseline Oesterling, our method exhibits superior fairness. Additionally, Oesterling imposes a convex assumption on the objective function, thereby limiting its capacity to employ advanced debiasing techniques.

3.5 The Machine Unlearning Performance of FMU

The primary goal of machine unlearning is to remove the information requested for unlearning by the user on a model. To evaluate the effectiveness of FMU, we conducted two experiments: Assessing model accuracy on both the target and non-target groups before and after unlearning, and (2) Evaluating the performance of membership inference attacks.

3.5.1 Will FMU Efficiently Unlearn the Requested Information?

To answer this question, we evaluate the performance of the model on both unlearned and remaining data before and after applying FMU or the other two unlearning baselines. Following the prior research, we select a group of data as the target group for unlearning, while the remaining data constitutes the non-target group. In Figure 2, we report the utility on `ACS-I`. Here, we specify the male positives as the target group



(a) Non-Target Data Accuracy

(b) Target Data Accuracy

Figure 2: The training accuracy of both target and non-target data on ACS-I is analyzed, considering a 30% unlearning ratio, with gender as the sensitive attribute.

for unlearning, while the non-target groups include male negatives, female positives, and female negatives. Initially, we conduct fair training. At epoch 100, we apply each unlearning approach to forget the entire target group data and obtain an unlearned model. Then, we continue training on each unlearned model and report the test accuracy for both the target and non-target groups in Figure 2. The obtained results lead to the following observations:

Table 3: Recall of Membership Inference Attacks.

Epoch	Retraining	Oesterling	FMU
0	0.97±0.17	0.97±0.17	0.97±0.17
0*	0.97±0.17	0.97±0.17	0.00±0.05
5	0.12±0.03	0.00±0.0	0.00±0.00
10	0.05±0.01	0.00±0.0	0.00±0.00
15	0.01±0.00	0.00±0.0	0.00±0.00
20	0.00±0.00	0.00±0.0	0.00±0.00

Observation 1: FMU preserves the utility on non-target data comparing to retraining. In Figure 2(a), fair retraining exhibits the highest accuracy after unlearning the target group at epoch 100, benefiting from retraining on the remaining data. While FMU initially experiences a slight dip, a brief period of training corrects this, resulting in comparable accuracy with fair retraining. Additionally, FMU outperforms Oesterling, which requires more epochs to recover the accuracy on the non-target data. This demonstrates the effectiveness of FMU in maintaining performance after unlearning for the non-target group.

Observation 2: FMU is effective at unlearning the requested data. Figure 2(b) illustrates that FMU exhibits the best unlearning performance, showing the lowest prediction accuracy on the target data, which is close to random guessing for the binary classification task. This suggests that the unlearned model obtained from FMU has largely forgotten the target data and struggles to make accurate predictions when trained solely on the remaining data. In contrast, both fair retaining and Oesterlingmodels achieve higher prediction accuracies on the target data compared to FMU, indicating that they retain some level of knowledge regarding the target data within their unlearned models. Fair retraining achieves the highest performance among the three methods, reaching its peak accuracy at approximately 110 epochs. However, there is a decline in performance afterward. This decline may result from the model remembering knowledge of the target data at the beginning, but as it is now training on the remaining data, which does not contain the target data anymore, inconsistency occurs, impairing its ability to effectively predict the target data.

3.5.2 Can A Membership Inference Attack Deduce Information From Unlearned Data?

The threat model. The adversary aims to determine whether specific records are present in the training data by employing membership inference attacks. Such leakage of membership information can significantly compromise privacy. In this scenario, the adversary has white-box access to the currently published model but lacks access to any previously released versions. The adversary can use data from a distribution similar to that used to train the target model. While the adversary's data may include duplicate records from the training set, such duplicates are not required.

Membership Inference Attack. We conducted a membership inference attack (MIA) (Yeom et al., 2018) on the trained model both prior to and after applying machine unlearning techniques. MIAs employ 20

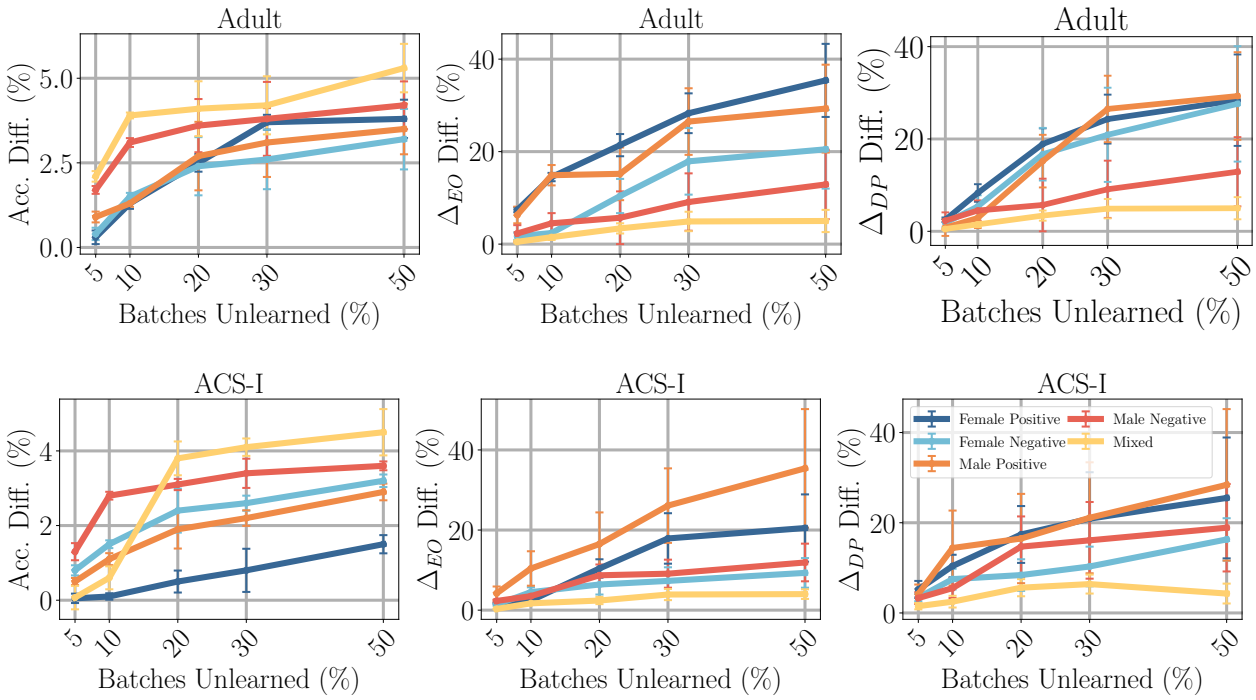


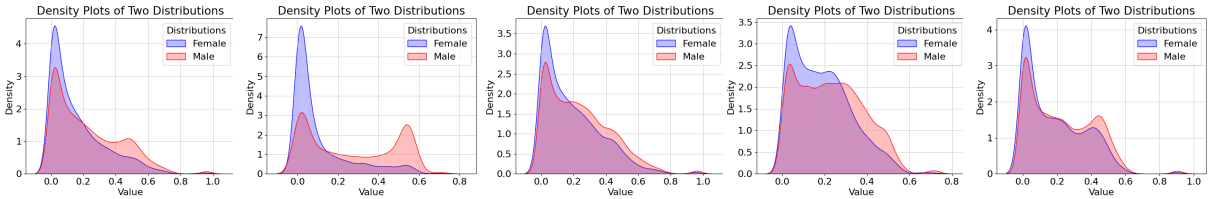
Figure 3: Accuracy difference (Acc. Diff.), $\triangle_{EO}$ difference ($\triangle_{EO}$ Diff.), and $\triangle_{DP}$ difference ($\triangle_{DP}$ Diff.) on `Adult` and `ACS-I` at different unlearning ratios $\{5\%, 10\%, 20\%, 50\%\}$, measured by the performance distance between applying Amnesiac and FMU. The sensitive attribute is gender. `Adult`'s unprivileged group is *Female Positive*, while `ACS-I` has more balanced classes.

shadow models, and the target model has trained for 100 epochs. Our evaluation of the effectiveness of these MIAs is based on the recall metric. These MIAs were systematically executed with a focus on specific individual examples. Following that, data removal techniques were employed in an effort to eliminate any learned information associated with this subset of individual examples. Notably, the unlearning process was initiated at epoch 0, and the outcomes of the MIA on the model, after applying the data removal technique but before any retraining, are observable at epoch 0^*. The recall value of the attack is reported.

Observation. As shown in Table 3, the results demonstrate that fair retraining offers minimal protection against the leakage of private class information for a span of more than 20 complete retraining epochs. By comparison, both Oesterling and FMU protect against MIA with less than 5 epochs of retraining, with FMU even showcasing the efficiency without any retraining.

3.6 How Do Different Unlearning Ratios And Distributions Impact FMU?

Due to the complexity of unlearning requests in the real world, we delve into a comprehensive analysis that considers various scenarios, each characterized by distinct ratios within different groups. Our assessment



(a) Fair Training (fair) (b) Amnesiac (unfair) (c) Fair Retraining (fair) (d) Oesterling (less fair) (e) FMU (more fair)

Figure 4: The distribution comparison for different baselines. The distribution is based on the predicted probability of male and female groups on the `Adult` dataset. The value indicates $\hat{\mathbf{y}}$.

focuses on the sensitive attribute *gender*, and we compare our method against the model after applying Amnesiac. This study involves a relatively balanced dataset, `Adult`, and an imbalanced dataset, `ACS-I`.

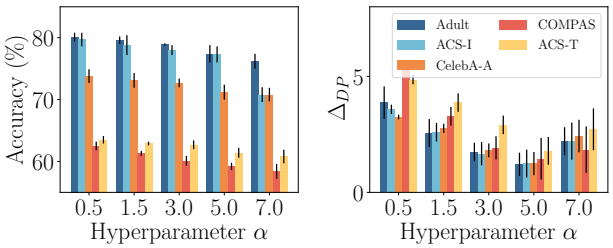
We report the performance improved, measured by the distance between Amnesiac and FMU. Specifically, for the utility metric, the difference is determined by the increment in value from Amnesiac to FMU, i.e., Acc. Diff.=Acc.(FMU)-Acc.(Amnesiac). While for the fairness metrics, the difference is determined by the value decrease from Amnesiac to FMU. In other words, i.e., $\triangle_{DP}$ Diff.=$\triangle_{DP}$(Amnesiac)-$\triangle_{DP}$(FMU), and $\triangle_{EO}$ Diff.=$\triangle_{EO}$(Amnesiac)-$\triangle_{EO}$(FMU). From the results synthesized in Figure 3, the following observations come to light:

Observation 1: The accuracy and fairness performance is correlated with original data distribution. Specifically, when deleting the unprivileged group of a dataset, the fairness difference is large since the standard unlearning can generate a highly biased model. This phenomenon becomes especially apparent when dealing with imbalanced datasets like `Adult`, in such cases, eliminating data from the unprivileged group tends to result in more significant fairness issues compared to data removal from the privileged group. However, when testing with balanced datasets like `ACS-I`, the fairness discrepancy stemming from data removal among different groups is generally smaller.

Observation 2: The mixed unlearning request will have less impact on the fairness by the unlearning since the trained model is already fair, and it will cause fewer distribution shifts, thus leaving less space for FMU to improve. While it can cause performance degeneration to a large degree since the unlearning is based on the whole dataset, the same percentage will cause more samples to be removed.

3.7 How Does FMU Achieve Better Fairness?

To understand why FMU is effective at reducing bias, we compared its prediction distribution with baseline methods. We tested it by removing 30% of the unprivileged group (female positive) from the `Adult` dataset, where gender is the sensitive attribute. The results, shown in Figure 4, confirm our previous observations: **using standard machine unlearning techniques can make an initially fair trained model (Figure 4(a)) become unfair (Figure 4(b))**. However, FMU achieves fairness levels similar to those obtained through fair retraining. FMU works by mitigating discrepancies in the representation of different groups. Remarkably, FMU outperforms Oesterling in fairness, which further demonstrates its effectiveness.

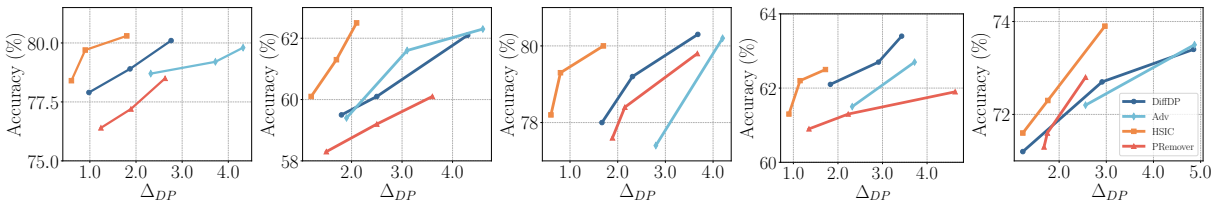


(a) Accuracy (b) $\triangle_{DP}$

Figure 5: Effect of α on accuracy and fairness.

3.8 Can We Balance the Accuracy and Fairness Trade-Off by Adjusting α?

One important hyperparameter in FMU is α, controlling the influence of fairness regularizer on the model prediction. To explore how sensitive the model is to α and to strike a balance between high accuracy and low discrimination, we train FMU on various datasets using different α values. Specifically, we test α values of $0.5, 1.5, 3.0, 5.0, 7.0$. Analyzing the results in Figure 5, we observe that, in general, higher α values initially lead to better fairness in FMU, but as α increases further, the debiasing ability degenerates.



(a) `Adult` (b) `COMPAS` (c) `ACS-I` (d) `ACS-T` (e) `CelebA-A`

Figure 6: Accuracy and $\triangle_{DP}$ trade-off on five datasets. Results located in the upper-left corner are preferable.

Specifically, from Figure 5(a), we observe that when $\alpha \leq 1.5$, the classification performance is almost unaffected. Once α is too large, the accuracy will decay sharply. The impacts of α on the discrimination score are presented in Figure 5(b). **When increasing the value of α, FMU will first have decreased discrimination scores, while when the value is too large, the fairness drops** because it would be difficult to optimize the model to the global minimum. Therefore, to achieve a balance of accuracy and fairness, in all other experiments, we choose $\alpha = 3.0$ to perform FMU.

3.9 How Well Do Various Fairness Regularizers Work?

We further consider the accuracy-fairness trade-offs on different debiasing methods, including `HSIC` (Li et al., 2019), PrejudiceRemover (`PRemover`) (Kamishima et al., 2012), adversarial debiasing (`Adv`) (Zhang et al., 2018b), as well as `DiffDP` (Chuang & Mroueh, 2020). Figure 6 displays the Pareto front curves (Yao et al., 2023; Ling et al., 2022) produced by a grid search of hyperparameters for each method. The top left data point represents the ideal performance, with the highest accuracy and lowest bias.

For the fairness loss design in FMU, the Demographic Parity gap regularization `DiffDP` (Chuang & Mroueh, 2020) was employed, defined by: $\mathcal{L}_{fairness} = \frac{1}{N}\sum_{i=1}^{N}(|P(\hat{\mathbf{y}}_i = 1|\mathbf{s}_i = 0) - P(\hat{\mathbf{y}}_i = 1|\mathbf{s}_i = 1)|)$. While `DiffDP` is chosen as the primary regularizer in our work, other fairness regularizers could also be utilized in FMU. To evaluate the impact of various fairness regularizers, we conducted an ablation study, and the results are exhibited in Figure 6. This study aims to understand how different choices of fairness regularizers affect the trade-off between fairness and utility.

From Figure 6, we observe that: 1) at equal accuracy, `HSIC` exhibits lower $\triangle_{DP}$ compared with baselines, 2) while `Adv` shows its power in utility but it lacks stable debiasing ability, 3) `PRemover` exhibits its advantage in debiasing while performing less on the prediction task, 4) **`DiffDP` shows a balanced trade-off for both discrimination scores and accuracy**. Finally, we concluded that `DiffDP` achieves the best balance between fairness and utility, making it our final choice for implementing FMU in all other experiments.

4 Conclusion

While current methods for machine unlearning focus on protecting the privacy of requested data, they often overlook the crucial issue of introducing bias during the unlearning process. To tackle this problem, we develop a new framework called FMU that aims to promote fairness among different groups during unlearning. Our approach involves removing the model updates corresponding to unlearning requests, while also incorporating updates from sampled batches where sensitive attributes are aligned inversely with the unlearning requests. We demonstrate that FMU is effective in achieving better fairness while still maintaining privacy and utility. FMU is adaptable, working with various unlearning techniques and model designs for different web applications. Additionally, we show that FMU is versatile in handling different unlearning requests, making it flexible across a wide range of unlearning scenarios.

References

Charter of fundamental rights and general data protection regulation, 2019. URL `https://europa.eu/eurobarometer/surveys/detail/2222`.

Muhammad Aurangzeb Ahmad, Arpit Patel, Carly Eckert, Vikas Kumar, and Ankur Teredesai. Fairness in machine learning for healthcare. In *KDD*, pp. 3529–3530, 2020.

Wael Alghamdi, Hsiang Hsu, Haewon Jeong, Hao Wang, Peter Michalak, Shahab Asoodeh, and Flavio Calmon. Beyond adult and compas: Fair multi-class prediction via information projection. *NeurIPS*, 35: 38747–38760, 2022.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *NeurIPS tutorial*, 1:2017, 2017.

Thomas Baumhauer, Pascal Schöttle, and Matthias Zeppelzauer. Machine unlearning: Linear filtration for logit-based classifiers. *Machine Learning*, 111(9):3203–3226, 2022.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

Sumon Biswas and Hridesh Rajan. Do the machine learning models on a crowd sourced platform exhibit bias? an empirical study on model fairness. In *ESEC*, pp. 642–653, 2020.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine Unlearning, December 2020. URL `http://arxiv.org/abs/1912.03817`. arXiv:1912.03817 [cs].

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159. IEEE, 2021.

Alexander E Burakov, Evgeny V Galunin, Irina V Burakova, Anastassia E Kucherova, Shilpi Agarwal, Alexey G Tkachev, and Vinod K Gupta. Adsorption of heavy metals on conventional and nanostructured materials for wastewater treatment purposes: A review. *Ecotoxicology and environmental safety*, 148: 702–712, 2018.

Maarten Buyl and Tijl De Bie. Optimal transport of classifiers to fairness. In *NeurIPS*, 2022.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *ICDM Workshop*, pp. 13–18. IEEE, 2009.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *NeurIPS*, volume 30. Curran Associates, Inc., 2017a. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf`.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. *NeurIPS*, 30, 2017b.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480. IEEE, 2015a.

Yinzhi Cao and Junfeng Yang. Towards Making Systems Forget with Machine Unlearning. In *2015 IEEE Symposium on Security and Privacy*, pp. 463–480, San Jose, CA, May 2015b. IEEE. ISBN 978-1-4673-6949-7. doi: 10.1109/SP.2015.35.

Joymallya Chakraborty, Suvodeep Majumder, Zhe Yu, and Tim Menzies. Fairway: a way to build fair ml software. In *ESEC*, pp. 654–665, 2020.

Hongyan Chang and Reza Shokri. On the privacy risks of algorithmic fairness. In *EuroS&P*, pp. 292–303. IEEE, 2021.

Chong Chen, Fei Sun, Min Zhang, and Bolin Ding. Recommendation unlearning. In *WWW*, pp. 2768–2777, 2022a.

Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *NeurIPS*, 31, 2018.

Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. When machine unlearning jeopardizes privacy. In *WAHC*, pp. 896–911, 2021.

Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. Graph unlearning. In *CCS*, pp. 499–513, 2022b.

Jiali Cheng, George Dasoulas, Huan He, Chirag Agarwal, and Marinka Zitnik. Gnndelete: A general strategy for unlearning in graph neural networks. *arXiv preprint arXiv:2302.13406*, 2023.

Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. In *ICLR*, 2020.

European Commission. 2018 reform of eu data protection rules, 2018. URL https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf.

Weilin Cong and Mehrdad Mahdavi. Grapheditor: An efficient graph representation learning and unlearning approach. 2022.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *KDD*, KDD '17, pp. 797–806, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. doi: 10.1145/3097983.3098095. URL https://doi.org/10.1145/3097983.3098095.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *NeurIPS*, 34:6478–6490, 2021.

CA DoJ. California consumer privacy act (ccpa), 2018. URL https://oag.ca.gov/privacy/ccpa.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *ITCS*, pp. 214–226, 2012.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, pp. 259–268, 2015.

Antonio Ginart, Melody Y. Guan, Gregory Valiant, and James Zou. Making AI Forget You: Data Deletion in Machine Learning, November 2019.

Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *AAAI*, volume 35, pp. 11516–11524, 2021.

Neil D Gross, David M Miller, Nikhil I Khushalani, Vasu Divi, Emily S Ruiz, Evan J Lipson, Friedegund Meier, Yungpo B Su, Paul L Swiecicki, Jennifer Atlas, et al. Neoadjuvant cemiplimab for stage ii to iv cutaneous squamous-cell carcinoma. *New England Journal of Medicine*, 387(17):1557–1568, 2022.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *NeurIPS*, 29, 2016.

Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In *AISTATS*, pp. 2008–2016. PMLR, 2021.

Priyank Jain, Manasi Gyanchandani, and Nilay Khare. Big data privacy: a technological perspective and review. *Journal of Big Data*, 3:1–25, 2016.

Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. In *UAI*, pp. 862–872. PMLR, 2020.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *KAIS*, 33(1):1–33, 2012.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *ECML PKDD*. Springer, 2012.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Johannes Klicpera, Aleksandar Bojchevski, and Stephan Günnemann. Combining neural networks with personalized pagerank for classification on graphs. In *ICLR*, 2019.

Ronny Kohavi and Barry Becker. Uci adult data set. *UCI Meachine Learning Repository*, 5, 1996.

Nikita Kozodoi, Johannes Jacob, and Stefan Lessmann. Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research*, 297(3):1083–1094, 2022.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. Propublica compas analysis—data and analysis for 'machine bias'. `https://github.com/propublica/compas-analysis`, 2016. Accessed: 2023-03-13.

Yunqi Li, Yingqiang Ge, and Yongfeng Zhang. Tutorial on fairness of machine learning in recommender systems. In *SIGIR*, pp. 2654–2657, 2021.

Yuyuan Li, Chaochao Chen, Xiaolin Zheng, Yizhao Zhang, Biao Gong, Jun Wang, and Linxun Chen. Selective and collaborative influence function for efficient recommendation unlearning. *Expert Systems with Applications*, 234:121025, 2023.

Zhu Li, Adrian Perez-Suay, Gustau Camps-Valls, and Dino Sejdinovic. Kernel dependence regularizers and gaussian processes with applications to algorithmic fairness. *arXiv preprint arXiv:1911.04322*, 2019.

Hongyi Ling, Zhimeng Jiang, Youzhi Luo, Shuiwang Ji, and Na Zou. Learning fair graph representations via automated data augmentations. In *The Eleventh International Conference on Learning Representations*, 2022.

Gaoyang Liu, Xiaoqiang Ma, Yang Yang, Chen Wang, and Jiangchuan Liu. Federaser: Enabling efficient client-level data removal from federated learning models. In *IWQOS*, pp. 1–10. IEEE, 2021.

Jiahao Liu, Dongsheng Li, Hansu Gu, Tun Lu, Jiongran Wu, Peng Zhang, Li Shang, and Ning Gu. Recommendation unlearning via matrix correction. *arXiv preprint arXiv:2307.15960*, 2023.

Yi Liu, Lei Xu, Xingliang Yuan, Cong Wang, and Bo Li. The right to be forgotten in federated learning: An efficient realization with rapid retraining. In *INFOCOM*, pp. 1749–1758. IEEE, 2022.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, December 2015.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *ICML*, 2018.

Gemma Martinez-Nadal, Pedro Puerta-Alcalde, Carlota Gudiol, Celia Cardozo, Adaia Albasanz-Puig, Francesc Marco, Júlia Laporte-Amargós, Estela Moreno-García, Eva Domingo-Doménech, Mariana Chumbita, et al. Inappropriate empirical antibiotic treatment in high-risk neutropenic patients with bacteremia in the era of multidrug resistance. *Clinical Infectious Diseases*, 70(6):1068–1074, 2020.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *CSUR*, 54(6):1–35, 2021.

Anay Mehrotra and Nisheeth Vishnoi. Fair ranking with noisy protected attributes. *NeurIPS*, 35:31711–31725, 2022.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory*, pp. 931–962. PMLR, 2021.

Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.

Alex Oesterling, Jiaqi Ma, Flavio P Calmon, and Hima Lakkaraju. Fair machine unlearning: Data removal while mitigating disparities. *arXiv preprint arXiv:2307.14754*, 2023.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Jeffrey Rosen. The right to be forgotten. *Stan. L. Rev. Online*, 64:88, 2011.

Neelam Rout, Debahuti Mishra, and Manas Kumar Mallick. Handling imbalanced data: a survey. In *International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications: ASISA 2016*, pp. 431–443. Springer, 2018.

Saeed Sharifi-Malvajerdi, Michael Kearns, and Aaron Roth. Average individual fairness: Algorithms, generalization and experiments. *NeurIPS*, 32, 2019.

Lawrence W Sherman and Richard A Berk. The specific deterrent effects of arrest for domestic assault. In *Quantitative Methods in Criminology*, pp. 3–14. Routledge, 2017.

Changjian Shui, Gezheng Xu, Qi Chen, Jiaqi Li, Charles X Ling, Tal Arbel, Boyu Wang, and Christian Gagné. On learning fairness and accuracy on multiple subgroups. *NeurIPS*, 35:34121–34135, 2022.

David Marco Sommer, Liwei Song, Sameer Wagh, and Prateek Mittal. Towards probabilistic verification of machine unlearning. *arXiv preprint arXiv:2003.04247*, 2020.

Vimalraj S Spelmen and R Porkodi. A review on handling imbalanced data. In *2018 international conference on current trends towards converging technologies (ICCTCT)*, pp. 1–11. IEEE, 2018.

Martin Strobel and Reza Shokri. Data privacy and trustworthy machine learning. *IEEE Security & Privacy*, 20(5):44–49, 2022.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling sgd: Understanding factors influencing machine unlearning. In *EuroS&P*, pp. 303–319. IEEE, 2022.

Tuan Tran, Uyen Le, and Yihui Shi. An effective up-sampling approach for breast cancer prediction with imbalanced data: A machine learning model-based comparative analysis. *Plos one*, 17(5):e0269135, 2022.

Ioannis Tsaousis and Mohammed H Alghamdi. Examining academic performance across gender differently: Measurement invariance and latent mean differences using bias-corrected bootstrap confidence intervals. *Frontiers in Psychology*, 13:896638, 2022.

Nancy Upton, Elisabeth J Teal, and Joe T Felan. Strategic and business planning practices of fast growth family firms. *Journal of small business management*, 39(1):60–72, 2001.

Kush R Varshney. Trustworthy machine learning and artificial intelligence. *XRDS: Crossroads, The ACM Magazine for Students*, 25(3):26–29, 2019.

Razvan Viorescu et al. 2018 reform of eu data protection rules. *European Journal of Law and Public Administration*, 4(2):27–39, 2017.

Sofia Visa and Anca Ralescu. Issues in mining imbalanced data sets-a review paper. In *MAICS*, volume 2005, pp. 67–73. sn, 2005.

Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. *arXiv preprint arXiv:2108.11577*, 2021.

Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. Machine unlearning: A survey, 2023a.

Mimee Xu, Jiankai Sun, Xin Yang, Kevin Yao, and Chong Wang. Netflix and forget: Efficient and exact machine unlearning from bi-linear recommendations. *arXiv preprint arXiv:2302.06676*, 2023b.

Yao Yao, Qihang Lin, and Tianbao Yang. Stochastic methods for auc optimization subject to auc-based fairness constraints. In *International Conference on Artificial Intelligence and Statistics*, pp. 10324–10342. PMLR, 2023.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *CSF*, pp. 268–282. IEEE, 2018.

Wei Yuan, Hongzhi Yin, Fangzhao Wu, Shijie Zhang, Tieke He, and Hao Wang. Federated unlearning for on-device recommendation. In *WSDM*, pp. 393–401, 2023.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *AISTATS*, 2017.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *AIES*, 2018a.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *AIES*, pp. 335–340, 2018b.

Dawen Zhang, Shidong Pan, Thong Hoang, Zhenchang Xing, Mark Staples, Xiwei Xu, Lina Yao, Qinghua Lu, and Liming Zhu. To be forgotten or to be fair: Unveiling fairness implications of machine unlearning methods. *arXiv preprint arXiv:2302.03350*, 2023a.

Guanhua Zhang, Yihua Zhang, Yang Zhang, Wenqi Fan, Qing Li, Sijia Liu, and Shiyu Chang. Fairness reprogramming. *arXiv preprint arXiv:2209.10222*, 2022.

Haibo Zhang, Toru Nakamura, Takamasa Isohara, and Kouichi Sakurai. A review on machine unlearning. *SN Computer Science*, 4(4):337, 2023b.

Jie M Zhang and Mark Harman. " ignorance and prejudice" in software fairness. In *ICSE*, pp. 1436–1447. IEEE, 2021.

Yang Zhang, Zhiyu Hu, Yimeng Bai, Fuli Feng, Jiancan Wu, Qifan Wang, and Xiangnan He. Recommendation unlearning via influence function. *arXiv preprint arXiv:2307.02147*, 2023c.

Indre Zliobaite. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*, 2015.

A Related Work

A.1 Fairness on Machine Learning

In recent years, the concept of fairness in machine learning has gained substantial attention. The primary objective of fairness in this context is to ensure that machine learning models remain impartial and unbiased, regardless of the individual or group involved. This pursuit of fairness can be broadly categorized into two main domains: group fairness and individual fairness. Group fairness (Dwork et al., 2012; Hardt et al., 2016; Corbett-Davies et al., 2017; Zafar et al., 2017; Madras et al., 2018) aims to guarantee equitable treatment across diverse groups. On the other hand, individual fairness (Dwork et al., 2012; Sharifi-Malvajerdi et al., 2019) centers on the principle that similar individuals should receive similar treatment. The literature on group fairness has put forth several definitions of fairness. Among these, the three most prevalent definitions are Demographic Parity (Zafar et al., 2017; Feldman et al., 2015; Zliobaite, 2015; Calders et al., 2009), Equalized Odds (Hardt et al., 2016), and Equality of Opportunity (Hardt et al., 2016).

Various bias mitigation techniques have been developed to effectively address issues related to fairness and bias in machine learning models. These techniques can be broadly classified into three categories: pre-processing (Kamiran & Calders, 2012; Calmon et al., 2017a), in-processing (Kamishima et al., 2012; Zhang et al., 2018a; Madras et al., 2018; Zhang et al., 2022; Buyl & De Bie, 2022; Alghamdi et al., 2022; Shui et al., 2022; Mehrotra & Vishnoi, 2022), and post-processing (Hardt et al., 2016; Jiang et al., 2020). Pre-processing algorithms strive to rectify the inherent biases within the dataset to ensure that the resulting trained model attains fairness (Calmon et al., 2017b). In-processing algorithms, on the other hand, operate during the training phase by adapting conventional empirical risk minimization objectives to incorporate fairness constraints (Gross et al., 2022; Sherman & Berk, 2017; Burakov et al., 2018; Martinez-Nadal et al., 2020). Post-processing algorithms adjust the predictions of a classifier to ensure equitable treatment across groups, leveraging insights gleaned from the training data (Tsaousis & Alghamdi, 2022; Hardt et al., 2016).

A.2 Machine Unlearning

The concept of machine unlearning, introduced in (Cao & Yang, 2015a), extends the "right to be forgotten" to the realm of machine learning. This involves removing specific data points from a trained model's dataset. The conventional method for implementing machine unlearning is to eliminate the designated instances from the original training dataset and then retrain the machine learning model from scratch. However, this approach becomes impractical due to the substantial computational burden it places, particularly when dealing with large datasets and frequent requests for data removal. Consequently, recent research in the field of machine unlearning, as seen in studies such as (Baumhauer et al., 2022; Bourtoule et al., 2021; Cao & Yang, 2015a; Izzo et al., 2021), has been predominantly focused on devising strategies to mitigate the computational overhead associated with the unlearning process.

For instance, (Cao & Yang, 2015a) introduced an innovative approach aimed at transforming learning algorithms into a summation form aligned with statistical query learning principles. This transformation effectively disentangles the intricate dependencies within the training data (Cao & Yang, 2015a). In practical terms, the model owner can seamlessly eliminate a specific data sample by simply excluding its corresponding transformations from the summations reliant on that particular sample. However, it's worth noting that this method finds limitations in its applicability to certain learning algorithms, particularly those, like neural networks, that resist transformation into summation forms.

An alternative solution comes from a previous study (Bourtoule et al., 2020), which proposed a more encompassing technique known as the Split-Integrated Sub-Model Approach (SISA). SISA revolves around a fundamental concept: the segmentation of training data into distinct shards, with each shard dedicated to training a specific sub-model. Consequently, the elimination of a particular sample mandates the retraining of solely the sub-model encompassing that sample. To expedite the unlearning process further, the authors suggest partitioning each shard into multiple slices and preserving intermediate model parameters during the model's refinement by each slice. This strategy contributes to the acceleration of the unlearning process while preserving overall model integrity.

An additional avenue of research in the field of machine unlearning focuses on the validation of compliance with data deletion requests by model owners. In this context, (Sommer et al., 2020) introduced a novel approach centered around the utilization of a backdoor-based method. The fundamental concept involves enabling data owners to strategically insert a backdoor within their data prior to the training phase of a Machine Learning as a Service (MLaaS) model. Subsequently, when these data owners seek to have their data removed, they possess the means to verify the thoroughness of the data deletion process by assessing the success rate of the implanted backdoor. This innovative method not only enhances accountability but also empowers data owners to actively monitor the status of their data even after integration into the model.

B Additional Experiments on the intuitive method

We want to study why we do not directly employ the intuitive approach mentioned in Section 2.3. The main issue is it needs to figure out which data relates to specific sensitive groups and store them separately during training, so it requires double storage to store the total update for each batch in every iteration during training. And FMU just needs a list of batches with sensitive group categories, so the separate storage is saved. Basically, the intuitive approach is impractical due to high storage requirements, while our method is more feasible and efficient.

We performed additional experiments using the intuitive method, where (R) denotes the sensitive attribute as race and (G) represents gender. correspondingly. We have highlighted results superior to FMU by presenting them in bold.

Table 4: The utility and fairness performances of the intuitive method across five datasets.

	Adult (R)	Adult (G)	COMPAS (R)	COMPAS (G)	ACS-T (R)	ACS-T (G)	ACS-I (R)	ACS-I (G)	CelebA-A (R)	CelebA-A (G)
Acc.	78.2 ± 0.3	78.1 ± 0.4	58.5 ± 0.2	59.3 ± 0.5	60.1 ± 0.5	61.8 ± 0.6	77.9 ± 0.4	77.7 ± 0.6	76.1 ± 0.5	72.5 ± 0.4
AUC	80.2 ± 0.7	80.9 ± 0.6	65.7 ± 0.4	60.8 ± 0.7	67.0 ± 1.0	67.9 ± 0.7	84.7 ± 1.4	85.1 ± 0.8	80.4 ± 0.5	81.1 ± 0.9
$\triangle_{EO}$	2.31 ± 1.1	2.91 ± 0.5	2.75 ± 0.6	4.19 ± 1.0	1.74 ± 0.5	3.14 ± 0.7	$\mathbf{1.20 \pm 0.9}$	2.52 ± 1.2	2.54 ± 0.6	3.43 ± 1.8
$\triangle_{DP}$	2.26 ± 1.7	3.68 ± 2.1	3.62 ± 1.5	3.15 ± 1.3	2.84 ± 0.6	4.07 ± 2.3	1.50 ± 1.2	$\mathbf{1.26 \pm 1.4}$	2.19 ± 2.2	4.52 ± 2.9

The results in the Table 4 reveal that the intuitive method demonstrates a prediction performance that is marginally inferior but still deemed acceptable while exhibiting a fairness performance comparable to FMU. **This discrepancy in prediction performance can be attributed to imbalanced utility updates, which adversely impact predictive accuracy (Rout et al., 2018).** Additionally, a conspicuous drawback of this method is its requirement for double the storage space compared to FMU.

C Additional Experiments on Various Deletion Distributions

We conducted additional experiments involving two distinct deletion distributions: a uniform deletion distribution with mixed sensitive groups, and a deletion specifically targeting the privileged group. For both scenarios, we examined varying deletion ratios, encompassing both small and large values.

C.1 Mixed Deletion on the Overall Data

C.1.1 5% Deletion on the Overall Data

Table 5: The utility and fairness performance of FMU on ACS-I with the sensitive attribute gender after unlearning. We assess the performance of a 5% data unlearning on the overall data. 'Std.' denotes the standard process without debiasing components.

	SISA	Amnesiac	Std. Retraining	Fair Retraining	Oesterling	FMU
Acc.	79.2 ± 0.8	78.4 ± 0.4	81.6 ± 0.9	80.7 ± 0.9	78.7 ± 0.5	80.5 ± 0.3
AUC	83.4 ± 2.2	82.4 ± 1.3	89.8 ± 0.4	84.8 ± 2.5	84.8 ± 1.1	86.8 ± 1.7
$\triangle_{EO}$	5.8 ± 2.7	6.1 ± 1.9	6.6 ± 0.8	2.8 ± 0.2	5.4 ± 1.0	4.3 ± 1.2
$\triangle_{DP}$	4.1 ± 2.6	5.3 ± 1.1	9.6 ± 3.4	2.9 ± 0.3	3.2 ± 1.4	3.7 ± 0.5

Based on the findings, it is reasonable to derive the following conclusions:

- Deleting only 5% of overall data does not significantly impact the **utility performance**, primarily due to the substantial sample base in ACS-I.
- The bias introduced by this deletion is comparatively lower than that resulting from deleting the **unprivileged group**, compared to Table 2. This supports our earlier inference that bias is introduced by the **distribution shift in model weights**.
- Even when the bias is small after the unlearning process in this experiment, as evident from the observations of SISA and Amnesiac, FMU continues to exhibit effective debiasing.

C.1.2 20% Deletion on the Overall Data

Table 6: The utility and fairness performance of FMU on ACS-I with sensitive attribute gender after unlearning. We assess the performance of a 20% data unlearning on the overall data. 'Std.' denotes the standard process without debiasing components.

	SISA	Amnesiac	Std. Retraining	Fair Retraining	Oesterling	FMU
Acc.	77.4 ± 1.4	76.2 ± 0.7	78.6 ± 1.2	77.7 ± 0.5	74.1 ± 0.7	75.2 ± 0.6
AUC	81.4 ± 2.1	82.4 ± 1.3	86.8 ± 1.1	83.8 ± 1.7	81.8 ± 0.9	82.8 ± 1.0
$\triangle_{EO}$	7.3 ± 4.2	9.6 ± 5.1	5.4 ± 3.7	1.5 ± 0.6	3.4 ± 0.7	1.8 ± 0.9
$\triangle_{DP}$	6.4 ± 3.6	7.5 ± 3.9	7.9 ± 4.5	1.9 ± 1.6	3.0 ± 1.6	2.6 ± 1.8

Based on the results, several conclusions can be inferred:

- A **larger** deletion ratio across the entire dataset has a **more pronounced impact** on both **utility and fairness** performance decline. This is attributed to a greater number of model weights being affected during the unlearning process, leading to increased imbalance.
- The performance of debiasing becomes **more significant** when the remaining dataset is **smaller**, especially when compared to a 5% overall deletion.

C.2 Deletion on Privileged Group

C.2.1 10% deletion on the privileged group

Table 7: The utility and fairness performances of FMU on ACS-I with sensitive attribute gender after unlearning. We assess the performance of a 10% data unlearning on the privileged group. 'Std.' denotes the standard process without debiasing components.

	SISA	Amnesiac	Std. Retraining	Fair Retraining	Oesterling	FMU
Acc.	78.5 ± 0.5	77.8 ± 0.7	81.4 ± 1.1	80.2 ± 0.8	78.8 ± 0.4	79.9 ± 0.6
AUC	85.2 ± 0.6	84.3 ± 0.9	89.9 ± 0.8	87.4 ± 1.3	86.0 ± 0.3	86.8 ± 0.6
$\triangle_{EO}$	10.5 ± 4.7	9.6 ± 5.6	5.3 ± 2.4	2.7 ± 1.1	4.1 ± 1.6	2.5 ± 1.8
$\triangle_{DP}$	9.7 ± 3.7	10.3 ± 5.1	8.3 ± 3.6	2.4 ± 1.9	3.6 ± 1.7	2.7 ± 0.9

C.2.2 30% deletion on the privileged group

Upon analyzing the results of deletion on the privileged group from the two tables above, several observations have surfaced:

- Deleting the privileged group at **the same deletion ratio** appears to have **a lesser impact** on the overall **prediction performance**, in contrast to the deletion of the **unprivileged group**, comparing to results in Table 2 in the paper.

Table 8: The utility and fairness performances of FMU on ACS-I with sensitive attribute gender after unlearning. We assess the performance of a 30% data unlearning on the privileged group. 'Std.' denotes the standard process without debiasing components.

	SISA	Amnesiac	Std. Retraining	Fair Retraining	Oesterling	FMU
Acc.	77.6 ± 1.6	78.1 ± 0.6	81.7 ± 1.1	79.1 ± 1.5	78.2 ± 1.7	79.2 ± 0.9
AUC	84.6 ± 1.0	85.5 ± 1.5	89.7 ± 0.8	86.7 ± 1.9	85.5 ± 0.7	86.1 ± 1.3
$\triangle_{EO}$	12.8 ± 6.0	11.1 ± 7.1	8.6 ± 0.8	1.6 ± 0.2	7.7 ± 2.3	2.1 ± 0.7
$\triangle_{DP}$	15.5 ± 10.9	10.3 ± 8.8	9.6 ± 3.4	1.8 ± 0.3	2.3 ± 1.2	1.9 ± 0.6

- The **bias** introduced from deleting the privileged group at the **same deletion ratio** is found to be **lower** than that resulting from the deletion of the **unprivileged group**.
- FMU consistently showcases its **efficacy** in mitigating bias.

D Additional Experiments on Runtime and Storage

D.1 Runtime Comparison

We conducted a runtime comparison between FMU and the established baselines. The experiments were executed using NVIDIA RTX A4000 GPUs with 16GB GDDR6 Memory, with a fixed number of requests set at 100. For both tabular and image datasets, we configured the retraining epoch to 100. It's important to note that in the case of SISA, $S = 20$ corresponds to the number of shards (Bourtoule et al., 2021).

Table 9: Runtime comparison (in seconds) on different unlearning methods.

	Retraining	SISA	Amnesiac	Oesterling	FMU
Runtime (Adult)	80.3 s	25.5 s	1.4 s	4.2 s	2.1 s
Speedup (Adult)	-	3.1×	57.4×	19.1×	38.2×
Runtime (CelebA-A)	1657.1 s	137.9 s	3.6 s	15.2 s	5.2 s
Speedup (CelebA-A)	-	12.0×	460.3×	109.1×	318.7×

Based on the findings, the following observations can be made:

- The computational time for FMU is consistently less than 2.1 seconds and 5.2 seconds on Adult and seconds on CelebA-A. This represents a significant improvement, being 38.2× and 318.7× faster than the retraining approach, respectively.
- Both Amenisac and FMU demonstrate superior time efficiency when compared to retraining and SISA. Notably, SISA necessitates the retraining of sub-models, whereas Amenisac and FMU operate without such a requirement.
- FMU outperforms Oesterling in terms of time efficiency, owing to its simplified computational design.

D.2 Storage Efficiency Comparison

We have recorded the counts of parameters stored for applying FMU and baselines for 2000 unlearned samples on both Adult and CelebA-A datasets. Note that the number of shards on SISA is set at 20. We follow the evaluation (Warnecke et al., 2021).

Table 10: Number of parameter counts.

	Fair Retraining	SISA	Amnesiac	Oesterling	FMU
#Paras (Adult)	1.6×10^6	3.6×10^4	2.2×10^3	3.3×10^4	2.5×10^3
#Paras (CelebA-A)	5.2×10^7	2.7×10^5	1.1×10^4	2.5×10^5	1.2×10^4

From the results in Table 10, we have the following observations:

- Retaining necessitates the largest storage capacity, followed by SISA, and then Oesterling, FMU, and Amnesiac.
- In comparison to Amnesiac, FMU additionally stores parameters related to the fair loss of the model. Consequently, opting for a complex and advanced fair loss choice may not be as storage-friendly, making `DiffDP` a preferable option due to its simplicity.
- It is noteworthy that SISA optimizes storage by partitioning data into shards and slices. Each shard contains a single model, and the final output is an aggregation of multiple models across these shards. Only the shards containing sensitive data are retrained. However, it requires saving a model checkpoint during training to enable retraining from an intermediate state. As the number of unlearned samples increases, more shards are needed for retraining, leading to a significant growth in storage demand.
- Oesterling's storage primarily is caused by the computation and inversion of the Hessian matrix of fair loss over the remaining dataset, incurring substantial storage costs.

In summary, FMU emerges as the most favorable choice for the trade-off between performance and storage.

E More Experimental Settings

E.1 Statistics of Datasets

- `Adult`[1] (Kohavi & Becker, 1996). The `Adult` dataset contains the 1994 U.S. census data, its primary objective involves predicting whether an individual's annual earnings surpass $\$50,000$. This prediction is based on demographic and financial attributes. Age and gender are sensitive attributes.
- `COMPAS`[2] (Larson et al., 2016). The `COMPAS` dataset contains information about criminal defendants, utilized for forecasting the likelihood of a defendant's recidivism within a span of two years. This dataset comprises attributes of the defendant, including their criminal history. It contains sensitive attributes of gender and race.
- **ACS**[3] (Ding et al., 2021). The ACS dataset offers a range of prediction tasks, including determining if an individual's income exceeds $\$50,000$ and whether they are currently employed. Derived from the American Community Survey (ACS) Public Use Microdata Sample (PUMS), it includes race, gender, and other pertinent task-related attributes for all tasks.
- `CelebA-A`[4] (Liu et al., 2015) The CelebFaces Attributes dataset is a collection of $20,000$ facial images featuring $10,000$ different celebrities. For each image, there are annotations of 40 binary labels that meticulously detail distinct facial attributes, including but not limited to gender, hair color, and age.

Table 11: The statistics of the datasets. #Feat. is the number of features after pre-processing. The $N:P$ is the ratio of binary target label, i.e., $\mathbf{y}=0:\mathbf{y}=1$ and $\mathbf{s}=0:1$ is the ratio of the sensitive attributes. Sens. is sensitive attributes.

Dataset	Target	Sens.	#Data	#Feat.	$N:P$	$\mathbf{s}=0:1$	$\mathbf{s}=0:1$
`Adult`	income	gender, race	45,222	101	1 : 0.33	1 : 2.08	1 : 9.20
`COMPAS`	credit	gender, race	6,172	405	1 : 0.83	1 : 4.25	—
`ACS-I`	income	gender, race	195,665	908	1 : 0.70	1 : 0.89	1 : 1.62
`ACS-T`	travel time	gender, race	172,508	1567	1 : 0.94	1 : 0.89	1 : 1.61
`CelebA-A`	attractive	gender, age	202,599	48×48	1 : 0.95	1 : 1.40	1 : 0.29

[1] `https://archive.ics.uci.edu/ml/datasets/adult`
[2] `https://github.com/propublica/compas-analysis`
[3] `https://github.com/zykls/folktables`
[4] `https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html`

E.2 Hyperparameter Selection Range

For implementing FMU, we tune the hyperparameter α, and the specific range for hyperparameter selection is outlined below:

Table 12: The selections of fairness control hyperparameter α.

Regularizer	Fairness Control Hyperparameter α
`DiffDP`	$0.5, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0, 2.5, 3.0, 3.5, 4$
`PRemover`	$0.05, 0.2, 0.3, 0.40, 0.50, 0.7, 0.9, 1.0$
`HSIC`	$50, 100, 200, 300, 400, 500, 600, 1000$
`Adv`	$0.5, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0, 2.5, 3.0, 3.5$

F Discussion on the Limitation

F.1 Information Loss of the Proposed Method

To evaluate the potential information loss resulting from the deletion of updates related to the opposing sensitive group, it is essential to evaluate the model's overall performance rather than solely focusing on predictions within the privileged group. When dealing with imbalanced or biased data, the presence of redundant information in the privileged group exacerbates the data imbalance issue, hindering the model's utility instead of enhancing its performance (Rout et al., 2018). This phenomenon is similar to downsampling methods employed in previous studies to address unbalanced data, demonstrating empirically that reducing privileged group data can improve overall model performance (Rout et al., 2018; Spelmen & Porkodi, 2018). The proposed approach seeks to boost overall performance. In addition to task performance, the consideration of fairness performance is crucial. A trade-off between fairness and utility exists, indicating that compromising the utility of privileged groups may be necessary to achieve superior overall prediction and fairness performance.

F.2 Analysis of the Trade-offs between Fairness and Computational Efficiency

Upon examining the runtime results presented in Table 9, FMU demonstrates the most favorable trade-offs between fairness and computational efficiency overall. The analysis is detailed as follows:

Standard Machine Unlearning Methods - SISA V.S. Amnesiac: SISA guarantees precise unlearning but demands substantial memory usage and requires implementation during training. On the other hand, Amnesiac necessitates storage space for a set of parameter update values from each batch. However, the overhead of these methods is generally lower than retraining a complete model from scratch. Notably, SISA incurs higher runtime compared to Amnesiac as it involves retraining sub-models.

Approximate Machine Unlearning Methods - Amnesiac V.S. FMU: Amnesiac exhibits slightly faster performance attributed to the sampling and debiasing processes in FMU. However, FMU excels in achieving better fairness, albeit at the cost of increased runtime.

Fair Machine Unlearning Methods - Oesterling V.S. FMU: FMU proves more time-efficient than Oesterling, primarily due to its simplified computation design. Oesterling's runtime is largely influenced by the computation and inversion of the Hessian matrix of fair loss over the remaining dataset. While Oesterling demonstrates commendable debiasing performance, FMU emerges as a more economical choice in the fairness-computation trade-off.