Add-Remove-or-Relabel: Practitioner-Friendly Bias Mitigation via Influential Fairness

Brianna Richardson
richardsonb@ufl.edu
University of Florida
Gainesville, Florida, USA

Prasanna Sattigeri
Dennis Wei
Karthikeyan Natesan Ramamurthy
Kush R. Varshney
Amit Dhurandhar
psattig@us.ibm.com
dwei@us.ibm.com
knatesa@us.ibm.com
krvarshn@us.ibm.com
adhuran@us.ibm.com
IBM Research
Yorktown Heights, New York, USA

Juan E. Gilbert
juan@ufl.edu
University of Florida
Gainesville, Florida, USA

ABSTRACT

Commensurate with the rise in algorithmic bias research, myriad algorithmic bias mitigation strategies have been proposed in the literature. Nonetheless, many voice concerns about the lack of transparency that accompanies mitigation methods and the paucity of mitigation methods that satisfy protocol and data limitations of practitioners. Influence functions from robust statistics provide a novel opportunity to overcome both issues. Previous work demonstrates the power of influence functions to improve fairness outcomes. This work proposes a novel family of fairness solutions, coined influential fairness (IF), that is human-understandable and also agnostic to the underlying machine learning model and choice of fairness metric. We conduct an investigation of practitioner profiles and design mitigation methods for practitioners whose limitations discourage them from utilizing existing bias mitigation methods.

KEYWORDS

machine learning, fairness, ethics, bias mitigation

ACM Reference Format:
Brianna Richardson, Prasanna Sattigeri, Dennis Wei, Karthikeyan Natesan Ramamurthy, Kush R. Varshney, Amit Dhurandhar, and Juan E. Gilbert. 2023. Add-Remove-or-Relabel: Practitioner-Friendly Bias Mitigation via Influential Fairness. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3593013.3594039

1 INTRODUCTION

Algorithmic bias is a persistent obstacle in the realm of machine learning (ML), impacting nearly every industry where it is applied. Concerns have arisen with respect to image recognition and object detection [20, 112], health assessment [40, 80, 90], advertisement systems [21, 42, 69, 109], and others [5, 19, 78]. Fairness research focuses on both the detection and the mitigation of bias in machine learning algorithms. The bulk of these contributions have been in the last decade; across the landscape of this research, there is an incredibly large collection of mitigation methods catering to diverse use cases, tutorials for utilizing mitigation methods, and toolkits that easily integrate into existing ML pipelines [89]. Furthermore, institutions are encouraging the production of fair machine learning tools now more than ever [51]. Despite the prolific number of research works, there is a severe lack of application of fairness technologies in practice [48, 88, 98, 105]. Practitioners require explanations of unfairness, transparent mitigation methods, and methods that are sensitive to their limitations as practitioners who abide by protocols and regulations established by their employer or by their own accord.

Influence functions from robust statistics [30] have been transformative for transparent and explainable ML. Using the Hessian matrix and the loss gradient, one can compute the influence that each training point has on a test outcome. This strategy has been shown to increase transparency, explain model behaviors, and identify adversarial examples [62]. Furthermore, recent work by [92] has shown that influence functions can also be used with group fairness objectives. This work aims to demonstrate the extent of that finding and employ the properties of influence functions to induce more transparency in fair machine learning pipelines.

This work introduces a new avenue of fair AI research coined influential fairness (IF). IF aims to add explainability, transparency, and contextualization to the procedure of bias mitigation via fairness influence functions. Furthermore, it aims to provide a diverse array of solutions to fit the diverse needs of practitioners across applications. Through the formulation of practitioner profiles, we

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06…$15.00
https://doi.org/10.1145/3593013.3594039

delineate four types of practitioners that emerge in fairness literature. Utilizing these profiles, we design mitigation methods that satisfy the user needs required by these practitioners. Our proposed mitigation methods take complex functions from robust statistics and transform them into simple add, remove, or transform strategies that encourage human-in-the-loop implementation. We propose four mitigation methods and provide additional implementation options to align with the limitations of our practitioner profiles. We demonstrate the effectiveness of our proposed strategies by testing their performance on benchmarked fairness datasets with several group fairness objectives. Our research contributions are as follows:

- A novel formulation of profiles that organize the diverse needs and limitations of practitioners.
- New mitigation methods utilizing group fairness influence functions, designed with transparency and practitioner limitations in mind, by allowing practitioners complete control of the type and number of modifications to their data.
- A novel evaluation of black-box and glass-box estimates of fairness influence functions to assist practitioner profiles with only black-box access to models.
- Lastly, a mapping of our curated methods to the practitioner profiles that best match their limitations.

We provide a software implementation of our methods on Github.[1]

2 BACKGROUND AND RELATED WORK

2.1 Bias Mitigation

The real harms [33] from algorithmic bias have motivated the growth of trustworthy AI research, which focuses on optimizing algorithms through accountability, transparency, explainability, fairness, privacy, etc. [103]. When investigating the important tenets of machine learning using the ethics guidelines of 84 institutions, [51] identified 11 unique tenets that often emerged in these documents. The most prominent of these tenets were fairness and transparency.

Fairness research is composed of socio-technical solutions to unwanted bias [97]; a variety of bias mitigation methods have emerged. Non-exhaustively surveying techniques by task type, domain, and intervention time point, there have been mitigation strategies for regression [3, 39], classification [2, 114], and ranking [17, 95, 117]. Some mitigation methods depend on the problem domain, such as natural language processing [13, 15, 18] and computer vision [56, 108, 113]. Also, many mitigation methods depend on when the practitioner intends to implement the fairness intervention, whether it be in the planning phase of a project, a.k.a. preprocessing [37, 53, 54, 67], during the implementation of the model, a.k.a. in-processing [38, 55, 66, 120, 121], or in the post-processing phase [47, 59, 84, 85]. Furthermore, mitigation methods have been built especially for specific notions of fairness [24], such as individual fairness [72, 84, 93], group fairness [47, 54], or counterfactual fairness [34, 67].

Despite the diverse selection of mitigation strategies that exist, there is a lack of application by practitioners [48, 88, 98, 105]. Many practitioners voice a concern with the transparency of mitigation methods that modify the model in a statistically-complex and seemingly black-box fashion [48, 88, 105]. Furthermore, many toolkits make generalizations about the needs and the freedoms of practitioners. When interviewed, practitioners preferred methods to evaluate their datasets and methods that provide recommendations for data collection, instead of non-transparent, complex methods of mitigation [88, 105]. Despite the research that encourages new data collection processes to minimize fairness disparity [28], few works have given curated feedback on data collection. Finally, many practitioners are limited in some respect: either they have institutional regulations preventing their use of certain strategies [9] or they have limited access to new data or sensitive features [48].

This work proposes four add/remove strategies and a relabeling strategy. While such methods exist in the literature [25, 32, 37, 43, 49, 50, 54, 55, 73, 99], our proposed strategies allow practitioners complete control of the number of modifications to their data. Furthermore, each method is singular in the type of modification to maximize transparency (e.g., adding and removing points are not being done at the same time, as is done in [54]).

2.2 Influence Functions

Influence functions [30, 44] have a rich history and have been used to study the effect of a single training point on the model parameters. Recent work [62] has shown that this machinery can be extended to estimate the influence of a training instance on the loss incurred by a model on new unseen test instance. Consider a supervised learning task that maps the input space $\mathcal{X}$ to the output space $\mathcal{Y}$. We are given N training points, $z_1, ..., z_N$, where $z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$. For a point z and a set of parameters θ in the space of the parameters Θ, let the loss associated with a training instance z be $L(z, \theta)$ and, therefore, let the objective be as follows:

$$\theta^* = \underset{\theta \in \Theta}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} L(z_i, \theta), \tag{1}$$

to find an optimal set of parameters θ^*.

By measuring the change in parameters, after upweighting a training instance z by some small ϵ, one can estimate the influence of that instance on the optimal model parameters. Under certain conditions [30], this influence can be estimated by taking the product of the loss gradient and Hessian matrix:

$$I(z) = \frac{d\theta^*_{z,\epsilon}}{d\epsilon}|_{\epsilon=0} = -H^{-1}_{\theta^*} \nabla_\theta L(z, \theta^*), \tag{2}$$

where H is the second derivative of loss with respect to the model parameters and $\nabla_\theta L(z, \theta^*)$ is the loss gradient with respect to the parameters for the upweighted training instance.

Using the derivation in [62], the resulting influence of a point z on the loss of a test point z_{test} is approximated as:

$$I(z, z_{test}) = -\nabla_\theta L(z_{test}, \theta^*)^\top H^{-1}_{\theta^*} \nabla_\theta L(z, \theta^*). \tag{3}$$

This assumes, however, that the loss function is strictly convex in the model parameters θ and the Hessian H_θ is a positive-definite matrix. In addition to the derivation, [62] demonstrated how influence functions could be employed to increase model transparency, explain model behavior, and identify adversarial samples.

Influence functions have been used across a number of works, many focusing on adding transparency to a diverse set of ML

[1] https://github.com/bricha2/InfluentialFairness

tasks such as causal inference models [4], NLP models [18, 45], multi-stage models [27], latent factor models [29], information diffusion networks [35], recommendation systems [119], and unsupervised tasks [100]. Outside of transparency, influence functions have been shown to prevent data poisoning [63, 101], increase model performance via relabeling [65] or data imputation [77], prevent data memorization [16], generate robustness scores for predictions [74, 96], and help identify harmful data [62, 92, 118]. Many works have built upon [62] optimizing influence function computations to include complex models, larger datasets, or generally faster computations [10, 12, 41, 61, 94, 122]. Furthermore, many works have provided alternative estimators of influence [6, 18, 46, 58, 60, 87, 107].

2.3 Influence Functions for Bias Mitigation

Influence functions are promising in algorithmic fairness, as they afford an opportunity to provide example-based explanations and build transparent mitigation methods. Nonetheless, to date, only a few works have studied the use of group fairness metrics as the objective in the influence function computation. Yu et al. [119] used influence functions to mitigate data distribution-based bias in recommendation systems. While their work used influence functions to mitigate data bias from feedback loops that stem from reinforcement learning, it did not focus on group-based metrics. Brunet et al. [18] proposed a novel method for estimating fairness influences in natural language processing (NLP) tasks using differential bias of word associations as an objective function. Again, their objective functions differs from the proposed work. Furthermore, the work herein goes a step further to investigate the use of the derived influence functions as a mitigation strategy.

Sattigeri et al. [92] proposed a post-hoc mitigation strategy that uses an infinitesimal jackknife-based approach to selectively sample points to remove and modify the model without refitting. Specifically, they consider several incarnations of group fairness metrics, denoted as $M(D_{val}, \theta^*)$ and define the fairness influence score as,

$$I_M(z, D_{val}) = -\nabla_\theta M(D_{val}, \theta^*)^\top H_{\theta^*}^{-1} \nabla_\theta L(z, \theta^*). \tag{4}$$

where, D_{val} is a held-out set and z is a training instance. The proposed work investigates these fairness influence functions in a broader scope, identifying the best practices for employing influential fairness and proposing several alternative mitigation methods besides selective removal. While Sattigeri et al. [92] utilized group fairness influence functions to mitigate bias, removing samples holds assumptions that practitioners have no protocols prohibiting such action and that the original model had enough data to balance the removal.

Another restriction of practitioners is the lack of access to training data or the model internals. We investigate if the fairness influence definition in [107] is a useful proxy for estimating fairness influence on the model parameters in such a black-box setting. Specifically, the definition of Wang et al. [107] aims to measure the change in fairness metric as measured on a new, unseen-by-the-model set when a point in this same new set is slightly upweighted. We call this approach a black-box estimator of fairness influence. Unlike the training instance fairness influence estimator in (4), these black-box estimators admit simpler analytical expressions. For example, the influence of a validation instance z_{val} on the statistical parity (SP) fairness metric value computed on the whole validation set can be computed as follows,

$$I_{M=SP}(z_{val}, D_{val}) = -h(z_{val}; \theta^*) + \hat{\mu}_0 \tag{5}$$

where h represents the trained model and $\hat{\mu}_0$ is the positive outcome rate of the unprivileged group. See Proposition 4 in [107] for black-box influence function expressions for other metrics. Wang et al. [107] then proceed to create a counterfactual distribution that renders the current model fair. This is then used to learn an optimal transport based pre-processor that transforms the features of a test instance to match the fair counterfactual distribution. In contrast, in this work, we restrict to a setting where we cannot alter feature values. We do investigate however whether the black-box fairness influence estimates of [107] can reduce the disparity of the original model. In summary, we aim to investigate mitigation methods based on several studied limitations of practitioners.

3 PRACTITIONER PROFILES

The first major contribution of this work is a novel mapping and delineation of practitioner profiles in the context of fairness. User-centered design has stressed for several decades the importance of persona or profile creations to assist in the design of inclusive technologies [86, 106]. For the scope of this work, we define *practitioner* as an individual actively engaged in the evaluation and mitigation of fairness in ML. We define four distinct practitioner roles.

The Free-Range Practitioner. This practitioner has glass-box access to their model, the training data used to train the model, and external data that can be incorporated into their model. Furthermore, they have few regulatory protocols limiting their ability to modify the data and the models (remove bad data, update model, control over data collection, etc.). There are no assumptions needed with respect to the free-range practitioner. Most existing solutions assume that most practitioners are free-range. Such a practitioner can do any type of modification because they are limited neither by protocols nor data.

The Protocol Practitioner. This practitioner also has glass-box access to their model, the training data, and external data. However, they also follow strict protocols that prohibit certain types of modifications to the model. These protocols can be based on personal philosophy or they could be limitations put in place by the practitioner's corporate or government regulations [81, 88, 98]. Transparency is especially important for this practitioner. They have stakeholders that they must explain modifications to, so it is critical they understand the functionality of their mitigation method [57, 105]. Furthermore, they have regulations that disallow for the relabeling of points, the deletion of data, or often even the use of sensitive attributes [81]. This practitioner is most likely to utilize methods that allow them to collect better data instead of making modifications to existing data [88, 105].

The Data-Limited Practitioner. This practitioner also has glass-box access to their model and the training data. They have freedoms with respect to manipulating the data and the model, but they have no access to new data, perhaps due to the cost or impracticality of data collection [8, 28]. A special category of data-limited practitioners are those without access to sensitive attributes [70]. Several works propose viable solutions for this type of limitation [31, 104, 116, 123]. The data-limited practitioner requires bias mitigation methods that utilize the data that the practitioner does have.

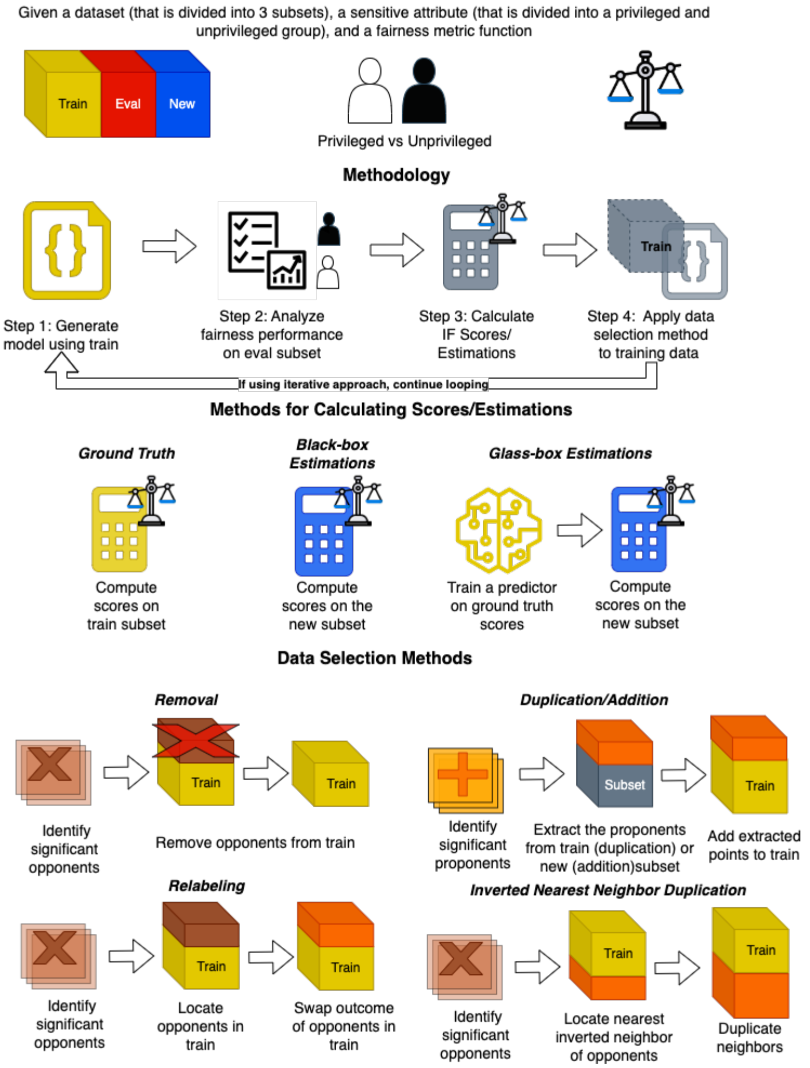
Given a dataset (that is divided into 3 subsets), a sensitive attribute (that is divided into a privileged and unprivileged group), and a fairness metric function
Train
Eval
New
Privileged vs Unprivileged
Methodology
Step 1: Generate model using train
Step 2: Analyze fairness performance on eval subset
Step 3: Calculate IF Scores/ Estimations
Train
Step 4: Apply data selection method to training data
If using iterative approach, continue looping
Methods for Calculating Scores/Estimations
Ground Truth
Compute scores on train subset
Black-box Estimations
Compute scores on the new subset
Glass-box Estimations
Train a predictor on ground truth scores
Compute scores on the new subset
Data Selection Methods
Removal
Identify significant opponents
Train
Train
Remove opponents from train
Duplication/Addition
Identify significant proponents
Subset
Extract the proponents from train (duplication) or new (addition)subset
Train
Add extracted points to train
Relabeling
Identify significant opponents
Train
Locate opponents in train
Train
Swap outcome of opponents in train
Inverted Nearest Neighbor Duplication
Identify significant opponents
Train
Locate nearest inverted neighbor of opponents
Train
Duplicate neighbors

Figure 1: Visual depiction of methodology and experimentation done in the proposed work. Ground Truth scores computed via methodology proposed in [92] and estimations of Black-box scores done as in [107].

Their reliance on data might discourage them from removing data. Existing bias mitigation methods that this user is the most likely to engage with include data augmentation or data synthesis.

The Auditing Practitioner. This practitioner has only black-box access to the model and access to external data. They can run data through the model and get outcomes, but they cannot see the internal mechanisms of the model. They also do not have access to the data used to train the model. Auditing practitioners can be further delineated into cooperative and independent. Cooperative practitioners are those who work in collaboration with the practitioners with access to the model internals and training data. In contrast, independent practitioners do not work in collaboration with practitioners with access. Auditing practitioners are rarely considered in fairness research despite the fact that many fairness practitioners are auditing practitioners. Fairness auditing interrogates proprietary software [1, 5, 14, 68, 75, 82, 102]. These practitioners aim to improve upon existing black-box models, despite their lack of access to the internal components. They are likely to use data pre-processing strategies or post-hoc strategies that do not require access to the original model.

In addition to aligning with these roles, the assumptions of this work align with the general requests made by practitioners in previous studies [28, 48, 88, 98, 105]: we aim to minimize modifications to the model, have a good fairness-accuracy trade-off, and optimize fairness outcomes. Furthermore, in concert with human-in-the-loop fairness [115], the proposed strategies are transparent and allow human intervention and easy manipulation.

4 IF MITIGATION METHODS

In this section, we introduce the family of proposed IF bias mitigation methods. A visual depiction of the different variations is given in Figure 1. The details are provided throughout the remainder of this section. At the end of the section, we discuss which proposed variations are amenable to which practitioner profiles.

4.1 High-Level Approach

In IF, we propose the following basic approach represented as pseudocode in Algorithm 1. Given a trained model, we evaluate its disparity according to a given group fairness metric M (line 4 in Algorithm 1) and then either exactly compute or estimate the influence function for M (line 5). Training data points with a large negative influence value are termed *proponents* and those with a large positive influence value are termed *opponents*; collectively we term training data points with large absolute value of influence as *significant points* (denoted ★ in Algorithm 1). Identifying thresholds for significance is further discussed in Section 4.3.2. To mitigate bias measured by M, we employ a strategy (line 7), which may involve adding, removing, and/or relabeling the significant points, either in a single pass or iteratively. If iteratively, the underlying machine learning model is retrained on every iteration.

Algorithm 1: General approach for iterative IF. Single-step IF is the same algorithm without the do-while loop, only its internal steps.

```
Data: train, eval
Input: disp(M, subset) disparity function that measures disparity for a
       given metric M on a subset subset; calcInf(subset) influential
       fairness score calculating function for a given subset; strat(★)
       mitigation function that takes in significant points ★ and applies to
       train; thresh threshold for allowable disparity; mod_limit integer
       threshold for permitted number of modifications;
Result: Bias-mitigated model.
1 Initialization ;
2 do
3     Train model with train subset ;
4     d = disp(M, eval) ;
5     calcInf() ;
6     Identify significant points ★ (either proponents or opponents) using
        knee approach;
7     mod, train = strat(★) ;
8 while abs(d)>thresh or len(★)> 0 or mod < mod_limit;
```

In one of the prior influence function works [62], the authors describe negative points that contribute to test loss as 'harmful' and positive points that reduce test loss as 'helpful'. Ref. [87] changed the language from 'helpful' and 'harmful' to 'proponents' and 'opponents', respectively. We adopt the language proposed by [87], except since our objective is to minimize disparity (in contrast to maximizing accuracy), our negative points are our 'proponents' for reducing disparity and our positive points are our 'opponents' which contribute to disparity.

To satisfy the diverse use cases of bias mitigation methods, there must be a high level of flexibility available to practitioners. The basic approach in Algorithm 1 allows for several layers of flexibility as described next.

4.2 Specific Strategies

The first choice is that of different modification strategies (line 7 in Algorithm 1). For this we propose (1) *removal & duplication*, (2) *nearest inverted neighbor duplication*, (3) *relabeling opponents*, and (4) *addition.*

Removal & Duplication. The first strategy is to remove opponents from the training data since these increase disparity. For proponents, the equivalent strategy is to duplicate them (i.e., have two copies in the training data). The removal of biased data via group-based fairness influence functions is also demonstrated in [92], but their implementation does not provide practitioners the freedom to customize the extent of their mitigation, as all the proposed methods do.

Nearest Inverted Neighbor (NIN) Duplication. Since the removal of opponents may not be an option for data-limited or protocol-limited practitioners, we propose a second kind of duplication as an alternative to removal. We call this proposed strategy *nearest inverted neighbor duplication*, depicted in the bottom-right corner of Figure 1. It duplicates the training point closest to an opponent but with the opposite label to counteract the impact of the opponent. The duplicated training point is not necessarily a proponent or opponent itself. To identify nearest inverted neighbors efficiently, we build two k-d trees, one for all samples in *train* with the favorable outcome (label) and the second for all samples with the unfavorable outcome. For each opponent, we use the k-d tree for the opposite outcome class to find the sample that is closest. In the context of unseen data, nearest neighbors can also be added from the *new* subset.

Relabeling Opponents. Influence functions also provide an opportunity to identify points that are mislabeled [62, 87]. Previous works have demonstrated the strength of relabeling to remove bias [22, 32, 54, 73]. By identifying points that contribute to bias and changing their label, these methods improve disparity outcomes. Under the assumption that opponents may be mislabeled samples, in this method, the labels of the most significant opponents are flipped.

Addition. The final proposed mitigation method is a selective sampling strategy that requires external labeled data. Furthermore, it requires an influence estimator (see 4.3.3) to estimate the influence of these new, external points that have not been seen by the model. Using an influence estimator, proponents from the new data are identified and then added to the training data before the model is retrained. This step most resembles duplication, but it does not require the proponents to already exist in the training data. Since samples are coming from an external source and haven't been used to train the model, ground truth scores cannot be formulated for this data. Our objective, therefore, is to identify the estimators that most improve outcomes.

While the simple procedures of reweighing and relabeling have been seen in bias mitigation literature, the collaboration of these methods with group fairness-based influence functions is entirely novel. Additionally, previous methods approximated samples to reweigh or relabel, based on proximity to the decision boundary or class distribution [50, 54]. Furthermore, the proposed methods provide practitioners the flexibility of choosing their depth of perturbation, which, to the extent of our knowledge, is entirely novel in fairness literature.

4.3 Additional Considerations

4.3.1 *Single-Step vs. Iterative Modifications.*

A second decision is whether to implement just one round of mitigation or an iterative application of mitigation. The choice of whether practitioners prefer a single step or iterative mitigation has an impact not only on the training costs (which scale with the number of iterations as seen in Algorithm 1), but also on the performance of the final model (evaluated in Section 6) and the transparency of the method. The last is because single-step mitigation methods allow practitioners to more easily keep track of modified points.

4.3.2 ***Modification Subset Size.*** For some practitioners, it may be desirable to limit the number of training points modified. In addition, since there exists a small-modification assumption in first-order approximations [12, 92], the influence of a modification of a large group of points will likely incur a large approximation error. Thus, a subset of the most influential points ($\star$ in Algorithm 1) should be chosen to maximize the impact of the mitigation methods while minimizing the number of points modified, which also controls the approximation error. This consideration applies regardless of whether the practitioner is using a single-step or iterative modification. The subset can be chosen through either its size (the "top k") or a significance threshold. Sattigeri et al. [92] chose a k that optimized fairness outcomes on a validation set. In this work, we identify the optimal k using the *knee* of the mean influence curve, specifically using the Kneedle approach proposed by [91].

4.3.3 ***Ground Truth Scores vs. Influential Estimators.*** The computation of ground-truth influential fairness scores requires access to the training data. This can limit the application of influence based bias mitigation strategies. Therefore, in this work, we also investigate two methods for estimating these influence scores that partly or wholly eliminate the need to have access to training data: glass-box predictors trained on ground truth scores, and a black-box fairness influence estimator, defined by [107].

Glass-Box Fairness Influence Predictors State-of-the-art predictors, such as gradient boosting and k-nearest neighbors, have been used across a variety of works [26, 52, 79] for their predictive power. We use these models to predict influential fairness scores. An influence score predictor well-trained on ground truth scores can then be utilized on new, unseen data without further access to training data.

Black-Box Fairness Influence Estimators Black-box estimations of influential points are calculated without access to any training data at all and using black-box access to the model. The use of black-box estimations of influential points can greatly reduce compute resources especially with complex models built with large data sets and it can provide external parties an opportunity to audit existing models. A method to effectively estimate influential points in a black-box manner could be promising to the influential fairness research space.

Towards this goal, the approach introduced by Wang et al. [107] comes closest to being completely model agnostic and hence viable as a black box influence estimator. As described in Section 2.3, Wang et al. [107]'s influence score is applicable only to an instance that is part of the test set used to measure the fairness metric and reflects this instance's contribution to the fairness metric value. In Section 6, we empirically investigate if these test distribution specific scores can be used to train the training data influence scores.

Table 1: Mitigation methods and the profiles they are associated with.

	Free-Range	Protocol	Data-Limited	Auditor
Removal	✓			
Duplication	✓	✓	✓	
Relabeling	✓		✓	
NIN Duplication	✓	✓	✓	
Addition	✓	✓		✓

4.4 Practitioner-Mitigation Associations

Table 1 maps the proposed mitigation methods to the practitioner profiles most likely to use them. Given that protocol and data limitations may differ, this mapping may not work for all practitioners. For the free-range practitioner, all of our mitigation methods are viable solutions. This practitioner may choose to utilize the solution with the best overall outcomes. (In the empirical results of Section 6, we find this to be the removal of opponents.) For the protocol practitioner, the glass-box or ground-truth implementations of the addition or duplication strategies are all viable solutions. For the data-limited practitioner, the ground-truth addition, duplication, or relabeling strategies would allow them to augment or transform their datasets without losing data. Lastly, the auditing practitioner would require a mitigation method that assumes no access to the original model. For a cooperative auditor, a glass-box influence predictor could be used in combination with the addition method. Otherwise, the independent auditor would need a black-box estimator.

5 EXPERIMENTAL SETUP

5.1 Datasets

In order to test the bounds of our methodology, we will test our procedures on three benchmarked fairness datasets [8, 71, 76, 83]: the Law School Admissions Council (`Law`), the Correctional Offender Management Profiling for Alternative Sanctions (`COMPAS`), and the `Adult` dataset. Each dataset also has a binary classification task. For each dataset, we chose a sensitive attribute and binarized it into privileged and non-privileged populations. Datasets received similar pre-processing steps: samples with missing values were removed, protected attributes were removed prior to training, and binary and categorical data were discretized without redundancy. `Law` was first used by [111] and contains performance records for over 20,000 law students. For this dataset, students are split into groups based on whether or not they classify as an Under-represented Minority (URM). `COMPAS` was first studied in [5] and contains the recidivism decision made by the software for 6,900 offenders in Broward County, Florida. The data is divided by race, with samples being classified as White or Non-White. The `Adult` dataset [64], otherwise referenced as the Census Income dataset, consists of demographic data from 1994. It is publicly available in the UCI machine learning repository [36].[2] With 45,222 samples across 10 attributes, this dataset is intended to be used to build predictors for whether an individual makes over 50,000 U.S. dollars in income.

[2] https://archive.ics.uci.edu/ml/datasets/adult

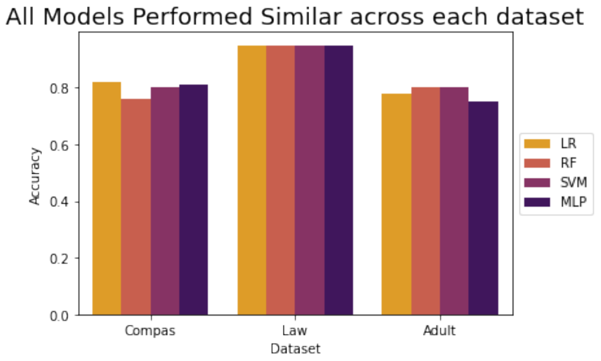

Figure 2: Comparison of four models generated for each dataset: Logistic Regression (LR), Random Forest (RF), Single Vector Machine (SVM), and Multilayer Perceptron (MLP).

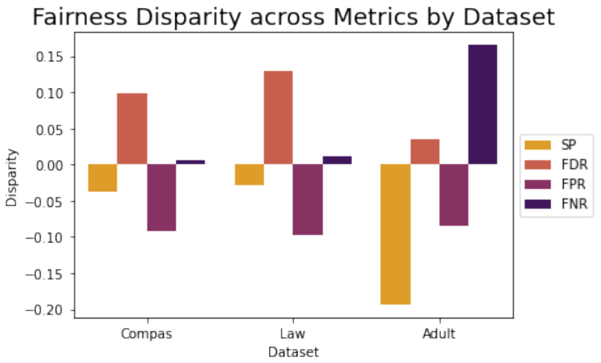

Figure 3: Disparities across the fairness objectives for LR models trained for each dataset. Highlighted metrics of interest for each dataset are: FPR for Compas, FDR for Law, and SP for Adult.

It is often used in algorithmic bias research to investigate gender, race, and/or age disparities.

5.2 Logistic Regression Models

For each dataset, we will produce a Logistic Regression (LR) model from the training subset using an Adam optimizer and an L2 regularization term. Previous work has demonstrated the strength of linear models when computing influence due to their small size in relation to alternative, more complex models [11, 61, 62]. By using an LR model, we can calculate exact Hessians instead of needing to estimate them due to the computational costs of larger models. Therefore, in this work, our ground truth scores represents exact influence scores and not approximations, removing any influence estimation error. Furthermore, we will be using a ridge regression because it preserves the convexity of the loss function. In Figure 2, we also demonstrate that LR is competitive with respect to accuracy compared to other potential models where IF can be computationally expensive or intractable to estimate accurately.

Table 2: Disparity metric function definitions for all metrics used in the proposed work. Y represents labels, $\hat{Y}$ is predicted labels, and G is the sensitive attribute.

Metric	Function
`Statistical Parity`	$M_{SP} = P(\hat{Y} \mid G = 0) - P(\hat{Y} \mid G = 1)$
`FDR Difference`	$M_{FDR} = P(Y = 0 \mid G = 0, \hat{Y} = 1) - P(Y = 0 \mid G = 1, \hat{Y} = 1)$
`FNR Difference`	$M_{FNR} = P(\hat{Y} = 0 \mid G = 0, Y = 1) - P(\hat{Y} = 0 \mid G = 1, Y = 1)$
`FPR Difference`	$M_{FPR} = P(\hat{Y} = 1 \mid G = 0, Y = 0) - P(\hat{Y} = 1 \mid G = 1, Y = 0)$

5.3 Fairness Objective Functions

There exists a multitude of fairness metrics in literature. For this work, we will focus on four metrics of fairness: Statistical Parity (SP) difference, False Discovery Rate (FDR) difference, False Positive Rate (FPR) difference, and False Negative Rate (FNR) difference. For the equations in Table 2, $\mathcal{Y} = \{0, 1\}$ is the space for both the labels Y and the predicted labels $\hat{Y}$. Additionally, the sensitive attribute $G \in \{0, 1\}$, where 1 is the privileged group.

Fairness disparities for LR models trained for each dataset utilized for fairness influence function computation were computed using the *eval* subset shown in Figure 1. Each disparity metric resulted in its own list of influential scores. Scores for unmitigated models for each dataset can be seen in Figure 3.

6 EXPERIMENTAL RESULTS

6.1 Validating Fairness Influence Functions

Prior to building mitigation methods, we validate our fairness influence functions. The standard method is to examine the change in outcomes as opponents (with values above 0) and proponents (with values below 0), ranked in order of decreasing magnitude, are removed [12, 18, 46, 92]. We conducted this study for both proponents and opponents, iteratively removing samples in decreasing magnitude until all proponents or opponents are removed, respectively. As expected, there was a rise in disparity as proponents were removed and a decrease in disparity as opponents were removed. Furthermore, there was a slow drop in accuracy as the first few points were removed that quickly accelerated due to lack of data. The results of this experiment are depicted in Figure 4.

6.2 Single-Step vs. Iterative Modifications

We first discuss the results of choosing different modification subset size for both single-step and iterative modifications. In the first row of Figure 5, we demonstrate the impact on disparity and accuracy when the top *i*% of points are removed before the model is refit and influence is recomputed. Across datasets and metrics, the optimal *i* differed substantially, making the selection of *i* another optimization problem. Interestingly, we observe that the *knee* strategy is able to achieve a nearly optimal outcome with fewer samples and this trend was common across the metrics and datasets. The middle row of Figure 5 depicts a search approach to a single-step implementation, where (similar to the removal validation) samples are perturbed in small increments using a ground truth mitigation method until disparity threshold is reached or the modification limit is reached. The last row of Figure 5 depicts the difference in performance of models trained with a single vs iterative approach. As can be seen, the choice between single and iterative application

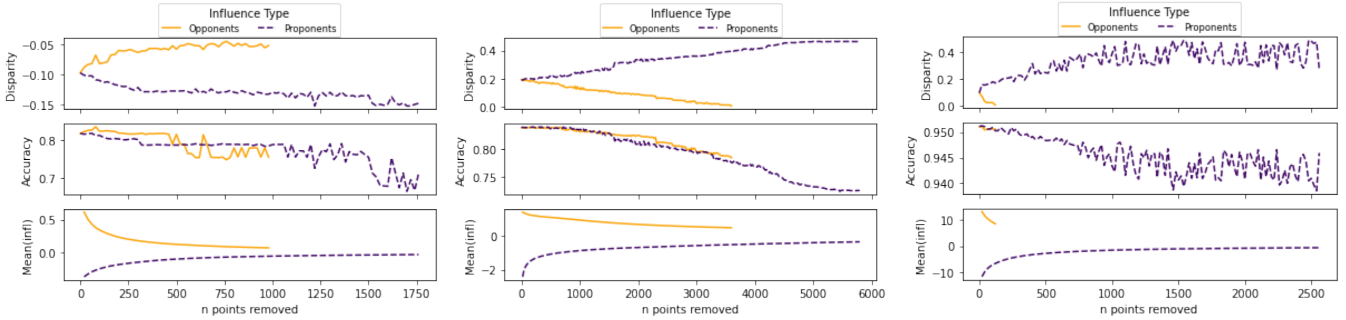

Figure 4: Removal of proponents vs removal of opponents, FDR for `COMPAS` (left), SP for `Adult` (middle), and FPR for `Law` (right). The bottom row shows the mean influence scores of points removed thus far.

is highly use-case dependent. Oftentimes, the single-step approach requires fewer perturbations and higher accuracy, but the final disparity is less consistently close to zero. Nonetheless, it may require a number of experiments to identify a good threshold for significance for influential points. Too few points modified and the model performance will remain unchanged, too many points modified and performance will change drastically.

6.3 Ground Truth Scores vs. Influential Estimators

6.3.1 ***Glass-Box Fairness Influence Predictors.*** Firstly, we compared three regressors for fairness influence prediction: gradient boosting, linear, and k-nearest neighbors. We trained the regressors on a portion of the ground truth points, and evaluated their performance using the remaining points. Evaluation was done using a rank-based criterion, Rank-Biased Overlap (RBO) [110]. Our previous results pointed to the importance of subset size, suggesting that selecting the top k or bottom k points is more important than precisely estimating the values of the points. Therefore, rank should be prioritized over regression error. Table 3 in the supplementary material shows that gradient boosting substantially outperformed the other methods for predicting influence. Therefore, this model type was used for subsequent analysis.

Next, to determine if fairness influence predictors are effective tools to mitigate bias, we applied our proposed mitigation methods using predicted influence scores. We predicted influence scores on the *new* subset and identified proponents and opponents there. We then applied each mitigation method to the original training data, finding matches in *train* as needed for the proponents and opponents in *new*. Details for each method are in the Supplementary Material. The rightmost column of Figure 7 demonstrate the performance of gradient boosting glass-box predictors. While gradient boosting did not perform as consistently well as ground truth, it was able to effectively reduce disparity and maintain accuracy across the datasets and mitigation methods.

6.3.2 ***Black-Box Fairness Influence Estimators.*** To investigate the impact of this estimator on the original model, we computed influence scores on the *new* subset and then applied the proposed mitigation methods to these estimations, in a similar manner as with gradient boosting predictions (again details are in the Supplement). Plots in the middle column of Figure 7 demonstrate the impact of applying the proposed strategies to the original model using the influential fairness black-box estimations. These experiments were conducted using the same procedure presented in Algorithm 1. None of the proposed strategies seemed to have a continuously positive impact on disparity. While this shows the seeming inability of these black-box estimations to influence the original model, we recall that the original intent in [107] was different, namely creating counterfactual models that are more fair than the original.

6.4 Comparing with Other Baselines

Figure 6 shows how the different strategies with access to ground truth influence scores perform when compared to a few baselines, chosen because they also involve reweighing or relabeling. Such existing alternatives include [23, 54, 66]. For this experimental comparison specifically, we conducted bias mitigation utilizing the reweighing strategies proposed by [50, 54]. [54] proposed a preprocessing reweighing strategy that assigns weights based on the class and sensitive attribute pairing. The duplication and removal strategies proposed here represent more transparent implementations of reweighing (weights of 2 or 0) and generally have better outcomes with respect to fairness and accuracy, as seen in Fig. 6. [50] proposed an in-processing reweighing strategy that assigns weights to samples in an iterative manner. Unlike [54], it provides solutions that are fairness metric-specific. We also conducted the relabeling strategy that was proposed in [54] and implemented in [7]. This relabeling strategy, also called data massaging, is a preprocessing strategy that relabels the training data based on the confidence of the model's prediction, the sensitive attribute, and the class label. Like the reweighing strategy proposed by [54], this strategy is not metric-agnostic.

While the ranking of the proposed influential mitigation methods was highly dependent on the dataset, as demonstrated in the second and third row of Figure 5, Figure 6 demonstrates how well the collection of these methods performs compared to existing strategies. Generally speaking, the purple markers representing the proposed methods tend to lie to the left of the others (lower disparity). In some cases, they dominate existing methods in terms of having higher accuracy and lower disparity.

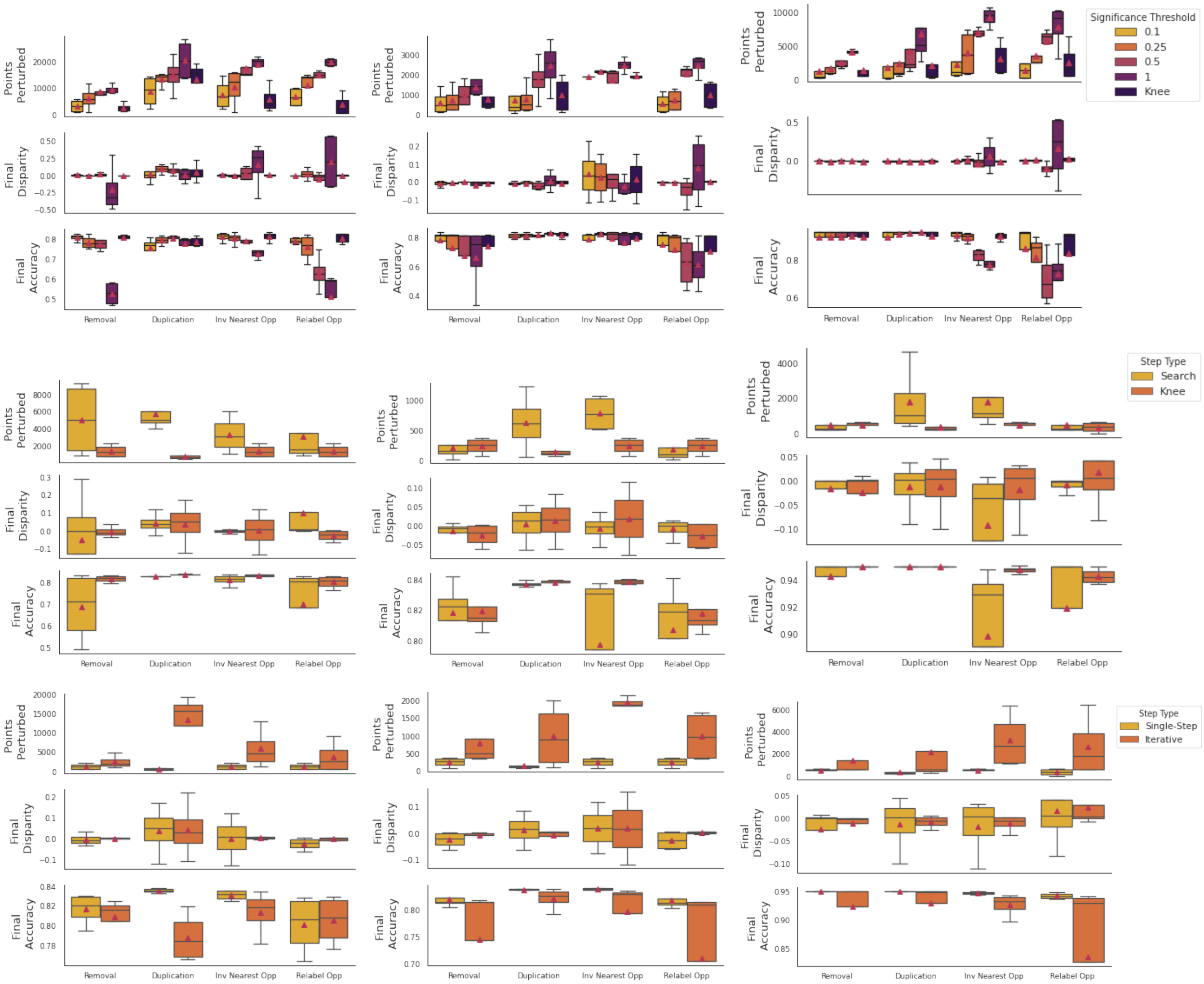

Figure 5: Comparing implementation strategies for ground truth mitigation methods across datasets, fairness objectives, and mitigation strategies. First row: Final performance of models modified using varying significance thresholds and an iterative implementation. Second row: Final performance of models modified using either a search-based or a knee-based single-step modification. Search-based implementation iteratively modifies original model with more significant points until disparity threshold met. Third row: Final performance of models utilizing either a single-step or iterative knee-based implementation. Columns align with datasets, from left to right: `Adult`, `COMPAS`, `Law`.

7 DISCUSSION

Practitioners routinely request more transparent mitigation methods that minimize disparity between groups, minimize modifications to the model, and also have a favorable fairness-accuracy trade-off. Furthermore, they desire mitigation methods that are simplistic and can be easily explained to stakeholders. The methods proposed in this paper take complex concepts from robust statistics and distill them into simple mitigation methods that either add data, remove data, or relabel data. While each of these methods has been demonstrated to reduce disparity, the magnitude of impact of their application is highly context dependent.

When comparing a single-step versus an iterative approach to modifying the original model, the results show that a single-step implementation could be just as impactful as an iterative approach, which is a novel finding to influential fairness. A single-step approach, as a post-hoc mitigation method, is substantially more transparent than an in-processing technique that requires iterative rebuilds of the model. Tracking the changes to the model becomes

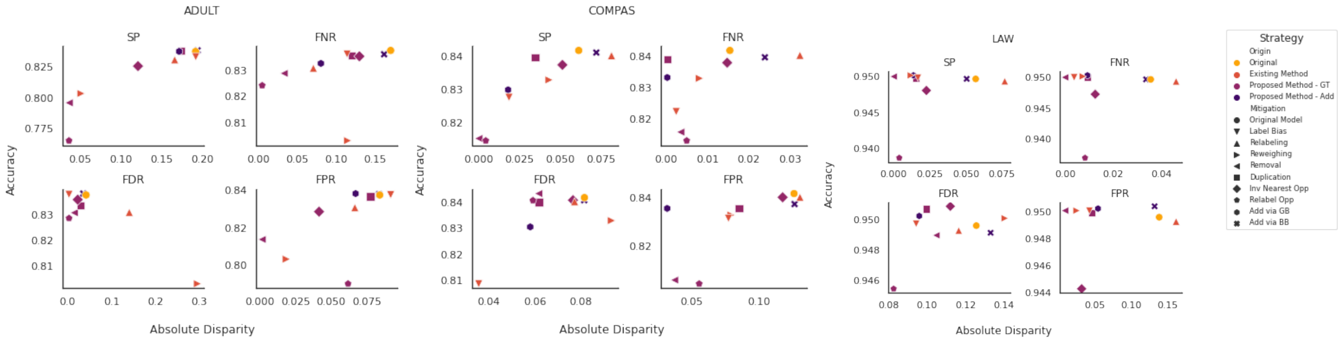

Figure 6: Performance between models modified using proposed mitigation methods (a single-step implementation with ground truth scores or, for addition only, gradient boosting predictions) compared to methods from pre-existing literature ([50, 54]).

substantially easier for practitioners. Furthermore, a single-step implementation of the proposed strategies is substantially more time efficient than existing mitigation methods that requires iterative retraining. This advantage can be amplified with existing influence function-based strategies that avoid retraining altogether [92]. When investigating the strength of influential fairness estimators to allow for the inclusion of external data and external practitioners, the findings were promising with respect to new data. Gradient boosting showed itself to be a great tool for learning influence and applying lessons to new data. Across the mitigation strategies, this glass-box predictor was able to improve disparate outcomes. On the other hand, the black-box estimator struggled to find relevant data for the original model. Nonetheless, [107] demonstrated how such a method could be used to build an alternate, fair model.

A major contribution of the proposed work is the creation of practitioner profiles. Four unique personas that emerged when investigating the literature on practitioner feedback of fairness research. Furthermore, we catered our proposed bias mitigation methods to satisfy the needs of these practitioners. While each of our mitigation methods were impactful when utilizing ground truth scores, the implementation utilizing estimations of black-box influence was not consistently impactful. Future work will investigate computing accurate black-box estimations of influential fairness scores.

Limitations. While this work introduces influential fairness through three fairness benchmarked datasets, there is still much work to be done in the demonstration of these mitigation methods. Firstly, this work intentionally focuses on small models with easily computed Hessians in order to establish ground truth scores. Many works have demonstrated effective methods to approximate influence for large, complex models [10, 12, 41, 61, 94, 122], so future work on utilizing these mitigation methods for more complex models is promising. This also holds true for applications of influential fairness in diverse use cases. Previous work has also demonstrated the strength of influence functions with multi-class labels [61, 62] and non-tabular data [18, 45, 61, 62, 119]. In this investigation of fairness influence functions, there must also be the evaluation of fairness for multiple protected attributes. Future work will also analyze the performance of the proposed strategies for such diverse use cases.

REFERENCES

[1] Philip Adler, Casey Falk, Sorelle A. Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2018. Auditing Black-Box Models for Indirect Influence. *Knowledge and Information Systems* 54, 1 (Jan. 2018), 95–122. https://doi.org/10.1007/S10115-017-1116-3/TABLES/3 arXiv:1602.07043

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *Proceedings of the International Conference on Machine Learning*. 102–119. arXiv:1803.02453

[3] Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. 2019. Fair Regression: Quantitative Definitions and Reduction-Based Algorithms. In *Proceedings of the International Conference on Machine Learning*. 120–129.

[4] Ahmed M Alaa and Mihaela Van Der Schaar. 2019. Validating Causal Inference Models via Influence Functions. In *36th International Conference on Machine Learning*. PMLR, 191–201. https://proceedings.mlr.press/v97/alaa19a.html

[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. *Machine Bias*. Technical Report. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[6] Juhan Bae, Nathan Ng, Alston Lo, Marzyeh Ghassemi, and Roger Grosse. 2022. If Influence Functions are the Answer, Then What is the Question? (sep 2022). https://doi.org/10.48550/arxiv.2209.05364 arXiv:2209.05364

[7] Niels Bantilan. 2017. Themis-ml: A Fairness-aware Machine Learning Interface for End-to-end Discrimination Discovery and Mitigation. In *Bloomberg Data for Good Exchange Conference*. arXiv:1710.06921v1

[8] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and machine learning*. fairmlbook.org. https://fairmlbook.org/index.html

[9] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *SSRN Electronic Journal* 104 (mar 2016), 671–732. https://doi.org/10.2139/ssrn.2477899

[10] Elnaz Barshan, Marc-Etienne Brunet, and Gintare Karolina Dziugaite. 2020. RelatIF: Identifying Explanatory Training Examples via Relative Influence. *Proceedings of Machine Learning Research* 108 (mar 2020), 26–28. https://doi.org/10.48550/arxiv.2003.11630 arXiv:2003.11630

[11] Samyadeep Basu, Phillip Pope, and Soheil Feizi. 2020. Influence Functions in Deep Learning Are Fragile. (jun 2020). https://doi.org/10.48550/arxiv.2006.14651 arXiv:2006.14651

[12] Samyadeep Basu, Xuchen You, and Soheil Feizi. 2020. On Second-Order Group Influence Functions for Black-Box Predictions. In *ICML'20: Proceedings of the 37th International Conference on Machine Learning*. 715–724. https://doi.org/10.5555/3524938

[13] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604. https://doi.org/10.1162/tacl_a_00041

[14] J. Bennett and S. Lanning. 2007. The Netflix Prize. In *Proceedings of Kdd Cup and Workshop*.

[15] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., 4356–4364.

[16] Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2019. Machine Unlearning. *Proceedings - IEEE Symposium on Security and Privacy* 2021-May (dec 2019), 141–159. https://doi.org/10.48550/arxiv.1912.03817 arXiv:1912.03817

[17] Amanda Bower, Hamid Eftekhari, Mikhail Yurochkin, and Yuekai Sun. 2021. Individually Fair Rankings. In *International Conference on Learning Representations*.

[18] Marc Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2018. Understanding the Origins of Bias in Word Embeddings. *36th International Conference on Machine Learning, ICML 2019* 2019-June (oct 2018), 1275–1294. https://doi.org/10.48550/arxiv.1810.03611 arXiv:1810.03611

[19] David Buil-Gil, Angelo Moretti, and Samuel H. Langton. 2021. The accuracy of crime statistics: assessing the impact of police data bias on geographic crime analysis. *Journal of Experimental Criminology* (mar 2021), 1–27. https://doi.org/10.1007/S11292-021-09457-Y/TABLES/9

[20] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification *. In *Proceedings of Machine Learning Research*, Vol. 81. 1–15.

[21] Ian Burn, Daniel Firoozi, Daniel Ladd, and David Neumark. 2021. Machine Learning and Perceived Age Stereotypes in Job Ads: Evidence from an Experiment. (jan 2021). https://doi.org/10.3386/W28328

[22] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Min Knowl Disc* 21 (2010), 277–292. https://doi.org/10.1007/s10618-010-0190-x

[23] Toon Calders and Indrė Žliobaitė. 2013. Why unbiased computational processes can lead to discriminative decision procedures. In *Studies in Applied Philosophy, Epistemology and Rational Ethics*. Vol. 3. Springer International Publishing, 43–57. https://doi.org/10.1007/978-3-642-30487-3_3

[24] Simon Caton and Christian Haas. 2020. Fairness in Machine Learning: A Survey. (oct 2020). https://doi.org/10.48550/arxiv.2010.04053 arXiv:2010.04053

[25] L. Elisa Celis, Amit Deshpande, Tarun Kathuria, and Nisheeth K. Vishnoi. 2016. How to be Fair and Diverse? (oct 2016). https://doi.org/10.48550/arxiv.1610.07183 arXiv:1610.07183

[26] Navoneel Chakrabarty, Tuhin Kundu, Sudipta Dandapat, Apurba Sarkar, and Dipak Kumar Kole. 2019. Flight arrival delay prediction using gradient boosting classifier. *Advances in Intelligent Systems and Computing* 813 (2019), 651–659. https://doi.org/10.1007/978-981-13-1498-8_57/COVER

[27] Hongge Chen, Si Si, Yang Li, Ciprian Chelba, Sanjiv Kumar, Duane Boning, and Cho-Jui Hsieh. 2020. Multi-Stage Influence Function. In *34th International Conference on Neural Information Processing Systems*. 12732–12742. https://doi.org/10.5555/3495724.3496792

[28] Irene Y. Chen, Fredrik D. Johansson, and David Sontag. 2018. Why Is My Classifier Discriminatory? *Advances in Neural Information Processing Systems* 2018-December (may 2018), 3539–3550. https://doi.org/10.48550/arxiv.1805.12002 arXiv:1805.12002

[29] Weiyu Cheng, Linpeng Huang, Yanyan Shen, and Yanmin Zhu. 2019. Incorporating interpretability into latent factor models via fast influence analysis. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (jul 2019), 885–893. https://doi.org/10.1145/3292500.3330857

[30] R. Dennis Cook and Sanford Weisberg. 1980. Characterizations of an Empirical Influence Function for Detecting Influential Cases in Regression. *Technometrics* 22, 4 (nov 1980), 495. https://doi.org/10.2307/1268187

[31] Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R. Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. 2019. Fair transfer learning with missing protected attributes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 91–98.

[32] Bo Cowgill and Catherine Tucker. 2017. Algorithmic Bias : A Counterfactual Perspective. In *NSF Trustworthy Algorithms*.

[33] Kate Crawford. 2017. The Trouble with Bias. https://www.youtube.com/watch?v=fMym_BKWQzk

[34] Pietro G. Di Stefano, James M. Hickey, and Vlasios Vasileiou. 2020. Counterfactual fairness: removing direct effects through regularization. (2020). arXiv:2002.10774 http://arxiv.org/abs/2002.10774

[35] Nan Du, Yingyu Liang, Maria-Florina Balcan, and Le Song. 2014. Influence function learning in information diffusion networks | Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32. In *31st International Conference on Machine Learning*. 2016–2024. https://dl.acm.org/doi/abs/10.5555/3044805.3045117

[36] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[37] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2011. *Fairness Through Awareness*. Technical Report. arXiv:1104.3913v2

[38] Michael Feldman, Haverford College, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. *Certifying and removing disparate impact **. Technical Report. arXiv:1412.3756v3

[39] Jack Fitzsimons, Abdul Rahman Al Ali, Michael Osborne, and Stephen Roberts. 2019. A General Framework for Fair Regression. *Entropy 2019, Vol. 21, Page 741* 21, 8 (jul 2019), 741. https://doi.org/10.3390/E21080741 arXiv:1810.05041

[40] Milena A. Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. 2018. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA internal medicine* 178, 11 (nov 2018), 1544. https://doi.org/10.1001/JAMAINTERNMED.2018.3763

[41] Han Guo, Nazneen Fatema Rajani, Peter Hase, Mohit Bansal, and Caiming Xiong. 2021. FASTIF: Scalable Influence Functions for Efficient Model Interpretation and Debugging. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings* (2021), 10333–10350. https://doi.org/10.18653/V1/2021.EMNLP-MAIN.808 arXiv:2012.15781

[42] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Vol. 13-17-August-2016. Association for Computing Machinery, New York, NY, USA, 2125–2126. https://doi.org/10.1145/2939672.2945386

[43] Sara Hajian and Josep Domingo-Ferrer. 2013. A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering* 25, 7 (2013), 1445–1459. https://doi.org/10.1109/TKDE.2012.72

[44] Frank R. Hampel. 1974. The Influence Curve and Its Role in Robust Estimation. *J. Amer. Statist. Assoc.* 69, 346 (jun 1974), 383. https://doi.org/10.2307/2285666

[45] Xiaochuang Han, Byron C. Wallace, and Yulia Tsvetkov. 2020. Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,*. Association for Computational Linguistics (ACL), 5553–5563. https://doi.org/10.48550/arxiv.2005.06676 arXiv:2005.06676

[46] Satoshi Hara, Atsushi Nitanda, and Takanori Maehara. 2019. Data Cleansing for Models Trained with SGD. *Advances in Neural Information Processing Systems* 32 (jun 2019). https://doi.org/10.48550/arxiv.1906.08473 arXiv:1906.08473

[47] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. *Equality of Opportunity in Supervised Learning*. Technical Report. arXiv:1610.02413v1

[48] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need. In *CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3290605.3300830 arXiv:1812.05239v2

[49] Vasileios Iosifidis, Besnik Fetahu, and Eirini Ntoutsi. 2020. FAE: A Fairness-Aware Ensemble Framework. *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019* (feb 2020), 1375–1380. https://doi.org/10.48550/arxiv.2002.00695 arXiv:2002.00695

[50] Heinrich Jiang and Ofir Nachum. 2019. Identifying and Correcting Label Bias in Machine Learning. (2019). arXiv:1901.04966 http://arxiv.org/abs/1901.04966

[51] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (sep 2019), 389–399. https://doi.org/10.1038/s42256-019-0088-2

[52] Jayakumar Kaliappan, Kathiravan Srinivasan, Saeed Mian Qaisar, Karpagam Sundararajan, Chuan Yu Chang, and C. Suganthan. 2021. Performance Evaluation of Regression Models for the Prediction of the COVID-19 Reproduction Rate. *Frontiers in Public Health* 9, September (2021), 1–12. https://doi.org/10.3389/fpubh.2021.729795

[53] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication, IC4 2009*. https://doi.org/10.1109/IC4.2009.4909197

[54] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision Theory for Discrimination-aware Classification. In *IEEE 12th International Conference on Data Mining*. 924–929. https://doi.org/10.1109/ICDM.2012.45

[55] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 7524 LNAI. Springer, Berlin, Heidelberg, 35–50. https://doi.org/10.1007/978-3-642-33486-3_3

[56] Kimmo Karkkainen and Jungseock Joo. 2021. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. *Proceedings - 2021 IEEE Winter Conference on Applications of Computer Vision, WACV 2021* (2021), 1547–1557. https://doi.org/10.1109/WACV48630.2021.00159

[57] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3313831.3376219

[58] Rajiv Khanna, Been Kim, Joydeep Ghosh, and Oluwasanmi Koyejo. 2019. Interpreting Black Box Predictions using Fisher Kernels. In *Twenty-Second International Conference on Artificial Intelligence and Statistics*. PMLR, 3382–3390. https://proceedings.mlr.press/v89/khanna19a.html

[59] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding Discrimination through Causal Reasoning. In *Proceedings of the 2017 Advances in Neural Information Processing Systems*, Vol. 30.

[60] Sosuke Kobayashi, Sho Yokoi, Jun Suzuki, and Kentaro Inui. 2020. Efficient Estimation of Influence of a Training Instance. *Journal of Natural Language Processing* 28, 2 (dec 2020), 573–597. https://doi.org/10.48550/arxiv.2012.04207 arXiv:2012.04207

[61] Pang Wei Koh, Kai-Siang Ang, Hubert H K Teo, and Percy Liang. 2019. On the Accuracy of Influence Functions for Measuring Group Effects. In *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 5254–5264. https://doi.org/10.5555/3454287.3454759

[62] Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning*. PMLR, 1885–1894. https://proceedings.mlr.press/v70/koh17a.html

[63] Pang Wei Koh, Jacob Steinhardt, and Percy Liang. 2022. Stronger data poisoning attacks break data sanitization defenses. *Machine Learning* 111, 1 (jan 2022), 1–47. https://doi.org/10.1007/S10994-021-06119-Y/FIGURES/13 arXiv:1811.00741

[64] Ronny Kohavi and Barry Becker. 1996. Census Income Data Set. http://archive.ics.uci.edu/ml/datasets/Adult

[65] Shuming Kong, Yanyan Shen, and Linpeng Huang. 2022. Resolving Training Biases via Influence-based Data Relabeling. In *ICLR*.

[66] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018* (apr 2018), 853–862. https://doi.org/10.1145/3178876.3186133

[67] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. *Advances in Neural Information Processing Systems* 2017-Decem (mar 2017), 4067–4077. arXiv:1703.06856 http://arxiv.org/abs/1703.06856

[68] Kyriakos Kyriakou, Pınar Barlas, Styliani Kleanthous, and Jahna Otterbacher. 2019. Fairness in Proprietary Image Tagging Algorithms: A Cross-Platform Audit on People Images. *Proceedings of the International AAAI Conference on Web and Social Media* 13 (jul 2019), 313–322. https://doi.org/10.1609/ICWSM.V13I01.3232

[69] Anja Lambrecht and Catherine Tucker. 2019. Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Management Science* 65, 7 (apr 2019), 2966–2981. https://doi.org/10.1287/MNSC.2018.3093

[70] Po-Ming Law, Sana Malik, Fan Du, and Moumita Sinha. 2020. Designing Tools for Semi-Automated Detection of Machine Learning Biases: An Interview Study. In *Proceedings of the CHI 2020 Workshop on Detection and Design for Cognitive Biases in People and Computing Systems*.

[71] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. 2022. A survey on datasets for fairness-aware machine learning. *WIREs Data Mining Knowledge Discovery* 12, 3 (2022). https://doi.org/10.1002/widm.1452

[72] Pranay K. Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R. Varshney, and Ruchir Puri. 2018. Bias Mitigation Post-processing for Individual and Group Fairness. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 2019-May (dec 2018), 2847–2851. arXiv:1812.06135 http://arxiv.org/abs/1812.06135

[73] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 2011. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press, New York, New York, USA, 502–510. https://doi.org/10.1145/2020408.2020488

[74] David Madras, James Atwood, and A. D'Amour. 2019. Detecting Extrapolation with Influence Functions. In *ICML Workshop*.

[75] Vidushi Marda and Shivangi Narayan. 2020. Data in New Delhi's predictive policing system. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (jan 2020), 317–324. https://doi.org/10.1145/3351095.3372865

[76] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. (2019). arXiv:1908.09635v2 https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[77] Xiaoye Miao, Yangyang Wu, Lu Chen, Yunjun Gao, Jun Wang, and Jianwei Yin. 2021. Efficient and effective data imputation with influence functions. *Proceedings of the VLDB Endowment* 15, 3 (nov 2021), 624–632. https://doi.org/10.14778/3494124.3494143

[78] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism* (first ed.). NYU Press. https://doi.org/10.2307/j.ctt1pwt9w5

[79] Simon Nusinovici, Yih Chung Tham, Marco Yu Chak Yan, Daniel Shu Wei Ting, Jialiang Li, Charumathi Sabanayagam, Tien Yin Wong, and Ching Yu Cheng. 2020. Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology* 122 (jun 2020), 56–69. https://doi.org/10.1016/J.JCLINEPI.2020.03.002

[80] Ziad Obermeyer and Sendhil Mullainathan. 2019. Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million People. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*. Association for Computing Machinery (ACM), New York, New York, USA, 89–89. https://doi.org/10.1145/3287560.3287593

[81] Will Orr and Jenny L Davis. 2020. Attributions of ethical responsibility by Artificial Intelligence practitioners. (2020). https://doi.org/10.1080/1369118X.2020.1713842

[82] Cecilia Panigutti, Alan Perotti, André Panisson, Paolo Bajardi, and Dino Pedreschi. 2021. FairLens: Auditing black-box clinical decision support systems. *Information Processing & Management* 58, 5 (sep 2021), 102657. https://doi.org/10.1016/J.IPM.2021.102657 arXiv:2011.04049

[83] Dana Pessach and Erez Shmueli. 2022. A Review on Fairness in Machine Learning. *ACM Computing Surveys (CSUR)* 55, 3 (feb 2022), 1–44. https://doi.org/10.1145/3494672

[84] Felix Petersen, Debarghya Mukherjee, Yuekai Sun, and Mikhail Yurochkin. 2021. Post-processing for individual fairness. *Advances in Neural Information Processing Systems* 34 (2021), 25944–25955.

[85] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. On Fairness and Calibration. In *31st Conference on Neural Information Processing Systems*. Neural information processing systems foundation, 5681–5690. arXiv:1709.02012 http://arxiv.org/abs/1709.02012

[86] John Pruitt and Tamara Adlin. 2005. *The Persona Lifecycle: Keeping People in Mind Throughout Product Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[87] Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating Training Data Influence by Tracing Gradient Descent. In *34th Internation Conference on Neural Information Processing Systems*. 19920–19930. https://doi.org/10.5555/3495724.3497396

[88] Brianna Richardson, Jean Garcia-Gathright, and Samuel F. Way. 2021. Towards fairness in practice: A practitioner-oriented rubric for evaluating fair ml toolkits. *Conference on Human Factors in Computing Systems - Proceedings* (may 2021). https://doi.org/10.1145/3411764.3445604

[89] Brianna Richardson and Juan E. Gilbert. 2021. A Framework for Fairness: A Systematic Review of Existing Fair AI Solutions. (dec 2021). https://doi.org/10.48550/arxiv.2112.05700 arXiv:2112.05700

[90] Pouria Rouzrokh, Bardia Khosravi, Shahriar Faghani, Mana Moassefi, Diana V.Vera Garcia, Yashbir Singh, Kuan Zhang, Gian Marco Conte, and Bradley J. Erickson. 2022. Mitigating Bias in Radiology Machine Learning: 1. Data Handling. *https://doi.org/10.1148/ryai.210290* 4, 5 (aug 2022). https://doi.org/10.1148/RYAI.210290

[91] Ville Satopää, Jeannie Albrecht, David Irwin, and Barath Raghavan. 2011. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. *Proceedings - International Conference on Distributed Computing Systems* (2011), 166–171. https://doi.org/10.1109/ICDCSW.2011.20

[92] Prasanna Sattigeri, Soumya Ghosh, Inkit Padhi, Pierre Dognin, and Kush R. Varshney. 2022. Fair Infinitesimal Jackknife: Mitigating the Influence of Biased Training Data Points Without Refitting. In *Advances in Neural Information Processing Systems*. https://doi.org/10.48550/arxiv.2212.06803 arXiv:2212.06803

[93] Richard Sawyer, Nancy Cole, and James Cole. 1976. Utilities and the Issue of Fairness in a Decision Theoretic Model for Selection. *Journal of Educational Measurement* 13 (09 1976), 59 – 76. https://doi.org/10.1111/j.1745-3984.1976.tb00182.x

[94] Andrea Schioppa, Polina Zablotskaia, David Vilar, and Artem Sokolov. 2021. Scaling Up Influence Functions. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 8 (dec 2021), 8179–8186. https://doi.org/10.48550/arxiv.2112.03052 arXiv:2112.03052

[95] Jakob Schoeffer and Niklas Kuehl. 2021. Appropriate Fairness Perceptions? On the Effectiveness of Explanations in Enabling People to Assess the Fairness of Automated Decision Systems. In *2021 Conference on Computer Supported Cooperative Work and Social Computing (CSCW '21 Companion)*, Vol. 1. Association for Computing Machinery. https://doi.org/10.1145/3462204.3481742 arXiv:2108.06500

[96] Peter Schulam and Suchi Saria. 2019. Can You Trust This Prediction? Auditing Pointwise Reliability After Learning. *AISTATS 2019 - 22nd International Conference on Artificial Intelligence and Statistics* (jan 2019). https://doi.org/10.48550/arxiv.1901.00403 arXiv:1901.00403

[97] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Inc, New York, NY, USA, 59–68. https://doi.org/10.1145/3287560.3287598

[98] Michelle Seng, Ah Lee, and Jatinder Singh. 2021. The Landscape and Gaps in Open Source Fairness Toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3411764.3445261

[99] Shubham Sharma, Yunfeng Zhang, Jesus M. Rios Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R. Varshney. 2020. Data augmentation for discrimination prevention and bias disambiguation. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (Feb. 2020), 358–364. https://doi.org/10.1145/3375627.3375865

[100] Andrew Silva, Rohit Chopra, and Matthew Gombolay. 2022. Cross-Loss Influence Functions to Explain Deep Network Representations. In *the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Valencia, Spain. https://www.researchgate.net/publication/346614859_Using_Cross-Loss_Influence_Functions_to_Explain_Deep_Network_Representations

[101] Jacob Steinhardt, Pang Wei Koh, and Percy Liang. 2017. Certified Defenses for Data Poisoning Attacks. *Advances in Neural Information Processing Systems* 2017-December (jun 2017), 3518–3530. https://doi.org/10.48550/arxiv.1706.03691 arXiv:1706.03691

[102] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. 2018. Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation. *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (dec 2018), 303–310. https://doi.org/10.1145/3278721.3278725 arXiv:1710.06169

[103] Kush R. Varshney. 2022. *Trustworthy Machine Learning.* Independently Published, Chappaqua, NY, USA.

[104] Michael Veale and Reuben Binns. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society* 4, 2 (2017). https://doi.org/10.1177/2053951717743530

[105] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. *Conference on Human Factors in Computing Systems - Proceedings* 2018-April (feb 2018). https://doi.org/10.1145/3173574.3174014 arXiv:1802.01029

[106] Sam Waller, Mike Bradley, Ian Hosking, and P. John Clarkson. 2015. Making the case for inclusive design. *Applied Ergonomics* 46 (2015), 297–303. https://doi.org/10.1016/j.apergo.2013.03.012 Special Issue: Inclusive Design.

[107] Hao Wang, Berk Ustun, and Flavio P Calmon. 2019. Repairing without Retraining: Avoiding Disparate Impact with Counterfactual Distributions. In *Proceedings of the 36th International Conference on Machine Learning.* http://github.com/ustunb/ctfdist.

[108] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. 2020. Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation. In *Conference on Computer Vision and Pattern Recognition.* 8916–8925. arXiv:1911.11834

[109] Jameson Watts and Anastasia Adriano. 2020. Uncovering the Sources of Machine-Learning Mistakes in Advertising: Contextual Bias in the Evaluation of Semantic Relatedness. *Journal of Advertising* 50, 1 (2020), 26–38. https://doi.org/10.1080/00913367.2020.1821411

[110] William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)* 28, 4 (nov 2010). https://doi.org/10.1145/1852102.1852106

[111] Linda F. Wightman. 1998. LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series.

[112] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. 2019. Predictive Inequity in Object Detection. (feb 2019). arXiv:1902.11097 http://arxiv.org/abs/1902.11097

[113] Wenying Wu, Panagiotis Michalatos, Pavlos Protopapaps, and Zheng Yang. 2020. Gender Classification and Bias Mitigation in Facial Images. *WebSci 2020 - Proceedings of the 12th ACM Conference on Web Science* (jul 2020), 106–114. https://doi.org/10.1145/3394231.3397900

[114] Yongkai Wu, Lu Zhang, and Xintao Wu. 2019. On convexity and bounds of fairness-aware classification. *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019* (may 2019), 3356–3362. https://doi.org/10.1145/3308558.3313723

[115] Mohammad Yaghini, Andreas Krause, and Hoda Heidari. 2019. A Human-in-the-loop Framework to Construct Context-aware Mathematical Notions of Outcome Fairness. (nov 2019). arXiv:1911.03020 http://arxiv.org/abs/1911.03020

[116] Shen Yan, Hsien Te Kao, and Emilio Ferrara. 2020. Fair Class Balancing: Enhancing Model Fairness without Observing Sensitive Attributes. *International Conference on Information and Knowledge Management, Proceedings* (oct 2020), 1715–1724. https://doi.org/10.1145/3340531.3411980

[117] Ke Yang, Joshua R. Loftus, and Julia Stoyanovich. 2020. Causal intersectionality for fair ranking. *Leibniz International Proceedings in Informatics, LIPIcs* 192 (jun 2020). https://doi.org/10.48550/arxiv.2006.08688 arXiv:2006.08688

[118] Yao-Yuan Yang, Chi-Ning Chou, and Kamalika Chaudhuri. 2022. Understanding Rare Spurious Correlations in Neural Networks. (feb 2022). https://doi.org/10.48550/arxiv.2202.05189 arXiv:2202.05189

[119] Jiangxing Yu, Hong Zhu, Chih Yao Chang, Xinhua Feng, Bowen Yuan, Xiuqiang He, and Zhenhua Dong. 2020. Influence Function for Unbiased Recommendation. *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (jul 2020), 1929–1932. https://doi.org/10.1145/3397271.3401321

[120] Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. 2020. Training individually fair ML models with sensitive subspace robustness. In *International Conference on Learning Representations.*

[121] Richard Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning.*

[122] Rui Zhang and Shihua Zhang. 2022. Rethinking Influence Functions of Neural Networks in the Over-Parameterized Regime. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 8 (2022), 9082–9090. https://doi.org/10.1609/aaai.v36i8.20893 arXiv:2112.08297

[123] Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. 2021. Towards Fair Classifiers Without Sensitive Attributes: Exploring Biases in Related Features. *WSDM 2022 - Proceedings of the 15th ACM International Conference on Web Search and Data Mining* (apr 2021), 1433–1442. https://doi.org/10.1145/3488560.3498493 arXiv:2104.14537v4

A ADDITIONAL DETAILS ON FAIRNESS INFLUENCE PREDICTION AND BLACK-BOX ESTIMATION

For both fairness influence prediction as well as black-box estimation, influence score estimates are obtained for points in the *new* subset. In the former case, they are predicted by the gradient boosting predictor, while in the latter they are estimated via the black-box method of [107]. From these estimated scores, significant proponents and opponents in *new* are identified, as was done in *train* with ground truth scores. Depending on the mitigation strategy, the subsequent action is as follows:

- Addition: Add proponents from *new* to *train*.
- Duplication: Duplicate points in *train* that are the nearest neighbors to the proponents in *new*.
- Removal and relabeling: Remove or relabel points in *train* that are the nearest neighbors to the opponents in *new*.
- Nearest inverted neighbor duplication: Duplicate the points in *train* that are the nearest inverted neighbor to each opponent in *new*. (Alternatively, the nearest inverted neighbor in *new* could be found and added to *train*.)

The above procedures were chosen to mirror those with ground truth scores and to compare predicted or black-box estimated scores with ground truth scores in as fair a manner as possible. In addition, they are motivated by the following real-world scenarios in which model building and model auditing are performed by two different parties. For fairness influence prediction, the model builder could provide the influence predictor to the auditor to assist the latter in the process of collecting more data, for the purposes of mitigating bias and otherwise improving the model. Or the auditor could have primary responsibility for bias mitigation and request the influence predictor from the model builder. The auditor would then pass back proponents or opponents from the *new* subset along with requested actions such as the ones listed above. For black-box estimation, this last scenario also applies, with the difference being that the auditor does not require any input from the model builder to estimate influence scores and proponents/opponents for the *new* subset.

B ADDITIONAL EXPERIMENTAL RESULTS

Table 3: Rank-Biased Overlap between Glass-box predictions and ground truth scores.

	Regressor Type		
Metric	Gradient Boost	K-Neighbors	Linear
EO	0.860 ± 0.032	0.509 ± 0.087	0.277 ± 0.139
FDR	0.902 ± 0.023	0.593 ± 0.081	0.087 ± 0.051
FPR	0.882 ± 0.033	0.422 ± 0.077	0.155 ± 0.066
SP	0.895 ± 0.008	0.419 ± 0.156	0.326 ± 0.095

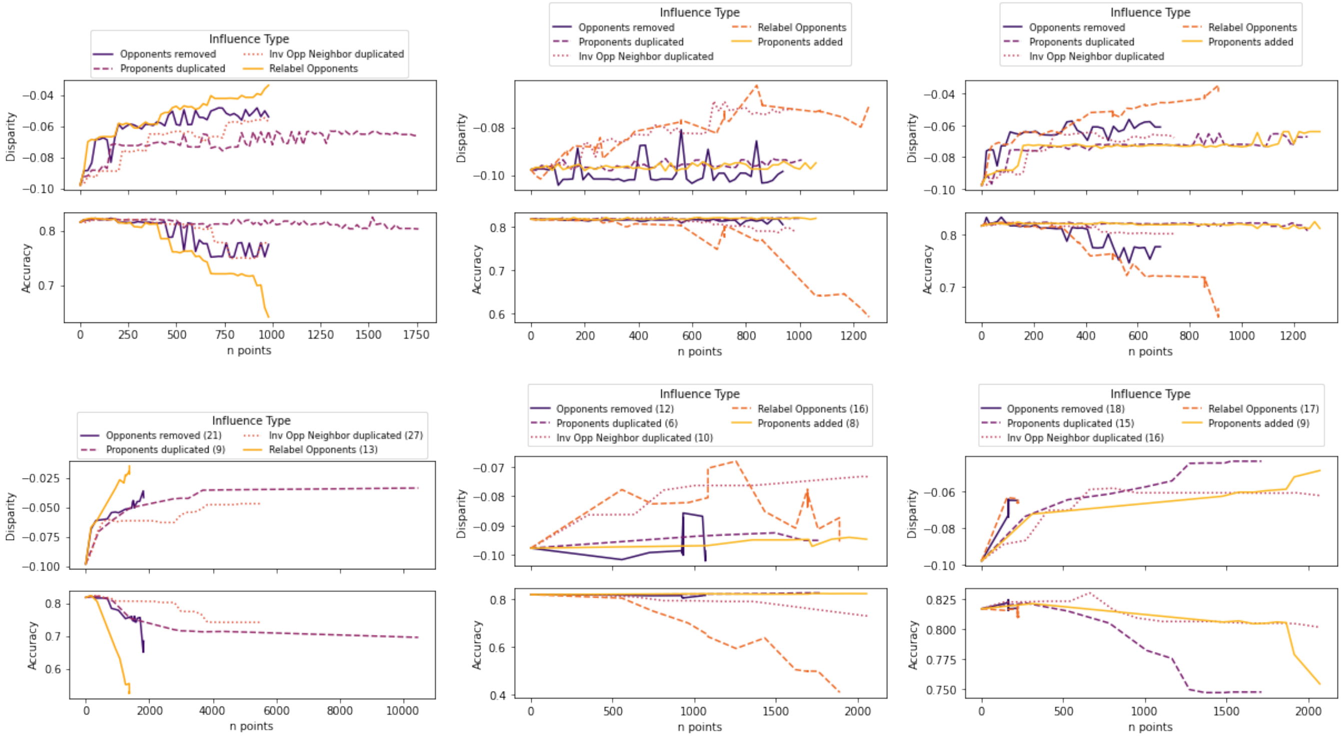

Figure 7: Results for experiments conducted on the COMPAS dataset, focused on mitigating FDR. From left to right, columns depict implementation of mitigation methods using ground truth scores, black-box influence estimators, and gradient boosting influence predictors, respectively. First row demonstrates single-step implementation of proposed methods, while bottom row depicts iterative experiments where influence scores are recomputed after each application of mitigation method with significant points. Each line segment represents an iteration and the total iterations for each experiment is shown in parentheses within the respective key. Results for other datasets can be seen in Suppl. Material.

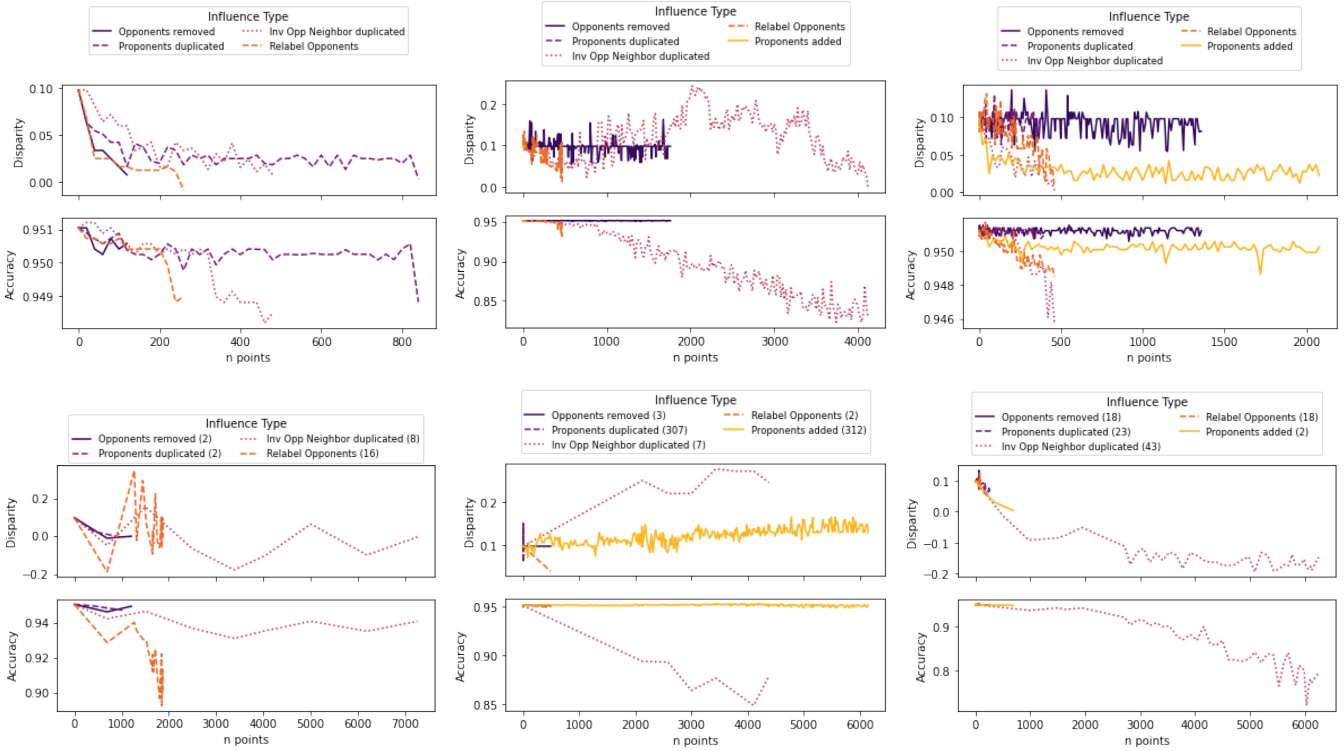

Figure 8: Results for experiments conducted on the Law dataset, focused on mitigating FPR. From left to right, columns depict implementation of mitigation methods using ground truth scores, black-box influence estimators, and gradient boosting influence predictors, respectively. First row demonstrate single-step implementation of proposed methods, while plots at the bottom depict iterative experiments where influence scores are recomputed after mitigation method employed with significant points. Each line segments represents an iteration and total iterations for each experiment is shown in parenthesis within the respective key.

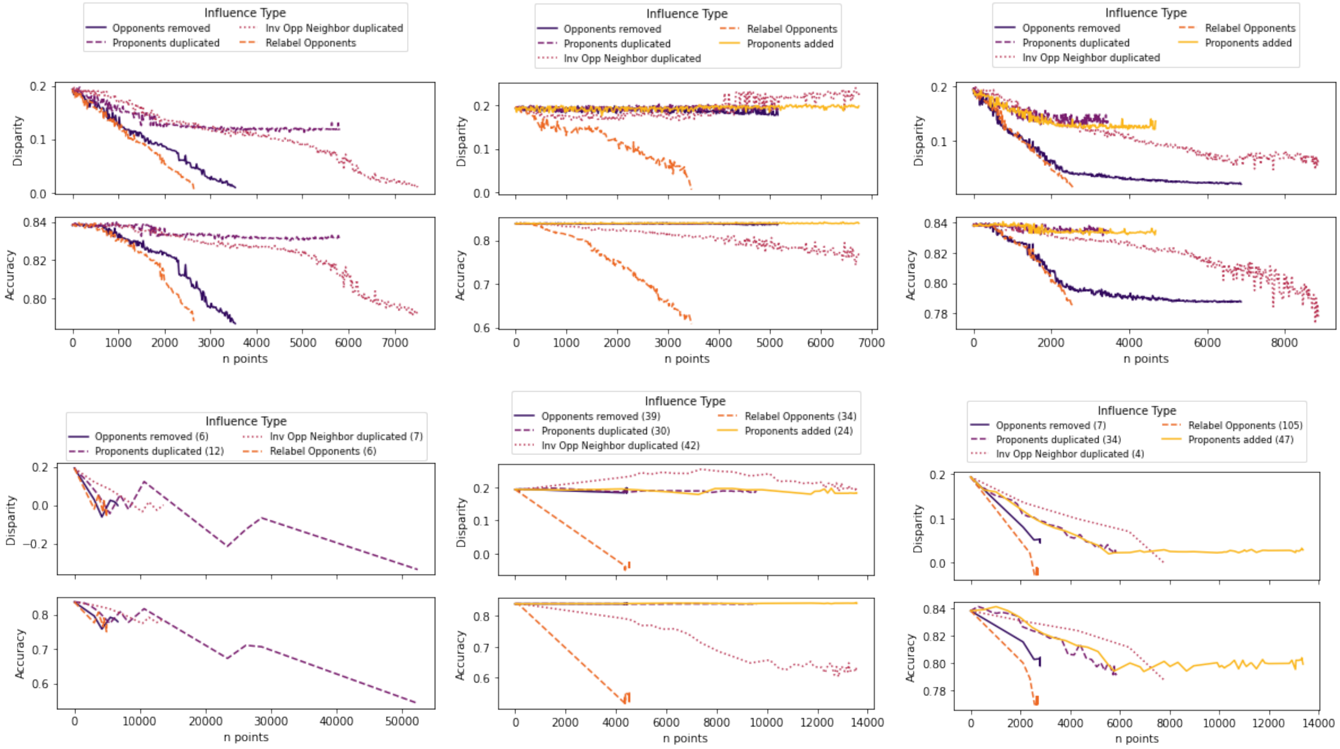

Figure 9: Results for experiments conducted on the Adult dataset, focused on mitigating SP. From left to right, columns depict implementation of mitigation methods using ground truth scores, black-box influence estimators, and gradient boosting influence predictors, respectively. First row demonstrate single-step implementation of proposed methods, while plots at the bottom depict iterative experiments where influence scores are recomputed after mitigation method employed with significant points. Each line segments represents an iteration and total iterations for each experiment is shown in parenthesis within the respective key.