Label Flipping for Group Fairness

Shashank Thandri
sthandr@purdue.edu
Purdue University
West Lafayette, Indiana, USA

Romila Pradhan
rpradhan@purdue.edu
Purdue University
West Lafayette, Indiana, USA

ABSTRACT

As algorithmic systems based on machine learning and artificial intelligence become increasingly prevalent in high-stakes decision-making, fairness has emerged as a critical societal issue. Individuals belonging to diverse demographic groups routinely receive conflicting algorithmic decisions largely due to the inherent errors and biases in the underlying training data used to build the systems, thus resulting in violations of group fairness. We study system unfairness or bias as a manifestation of erroneous data labels and address the problem of determining the order in which the labels of erroneously labeled data points must be corrected such that a system trained over the modified data exhibits lower unfairness. To obtain such an ordering of data points, we propose solutions based on the notion of entropy of individual data points and the estimated impact of correcting a label on system fairness. We further utilize the information-theoretic concept of the value of perfect information to compute the maximum expected utility of correcting a label on system fairness. We experimentally evaluated our solutions on several real-world datasets and demonstrated that flipping a small fraction of training data labels drastically reduces model bias while exhibiting bias reduction and efficiency trade-offs for the different solutions.

VLDB Workshop Reference Format:
Shashank Thandri and Romila Pradhan. Label Flipping for Group Fairness. VLDB 2025 Workshop: 14th International Workshop on Quality in Databases (QDB'25).

VLDB Workshop Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/s-thandri/labelflipping.

1 INTRODUCTION

Technologies based on artificial intelligence (AI) and machine learning (ML) are ubiquitous in the modern world. Business organizations routinely incorporate AI-based systems, such as chatbots, advertising, and recommendation systems [14, 31], in their workflows to improve the efficacy and efficiency of solutions and lower the costs of production [16, 32]. However, the unprecedented growth of AI has led to the continued aggravation of surrounding issues such as fairness and bias [16, 32] — AI-based systems have been shown to unfairly favor certain groups over others leading to a violation of human rights and other legal implications [13, 15, 40].

Proceedings of the VLDB Endowment. ISSN 2150-8097.

A number of fairness metrics and bias mitigation techniques have emerged as a result of the efforts to quantify the discrimination and debias these systems [29, 36]. Bias mitigation techniques are primarily categorized as *pre-processing* (modifying the data before model training), *in-processing* (modifying the learning algorithm), and *post-processing* (modifying the model predictions). Pre-processing techniques have shown to be among the most effective solutions that are also easy to implement since only the underlying training data needs to be updated.

The recent emphasis on data-centric AI, in line with preprocessing techniques, has steered the focus away from building sophisticated ML models to ensuring high-quality training data to develop robust and fair AI-based systems [4]. To ensure that models do not perpetuate existing biases, it is imperative to carefully consider the underlying training data. Data quality concerns stem from a range of underlying issues including missing values [38], data inconsistencies [27], violations of functional dependencies [3], label annotations errors [40], presence of outliers [38] etc.

This work focuses on resolving model fairness issues arising due to *label annotation errors*. Although recent approaches for label annotation consider weak supervision [10], the task of labeling is largely a human-oriented process [40]. Annotations, therefore, are not guaranteed to be error-free and manifest themselves in the form of incorrect and discriminatory decisions. We address the problem of resolving machine learning model unfairness through correcting potentially mislabeled training data instances. Existing work in this space focuses on label flipping as a way to mitigate data bias for individual fairness [40]; however, their current optimization formulation does not account for group fairness. Label flipping for resolving group fairness has been proposed in data massaging [20] that flips the labels for data points that are either the most favored or the most discriminated against; however, this approach is specific to the fairness definition of statistical parity and is not applicable to other fairness metrics.

We address the problem of determining the training data points whose labels should be flipped to reduce model unfairness. Given a machine learning model trained on a dataset, we propose solutions that present an ordering according to which the labels of training data points should be flipped such that a model trained on the modified training data has higher fairness compared to the original model. The task of determining which labels should be flipped is challenging because of a number of reasons. First, to evaluate a training data point for potentially flipping its label, the model must be retrained after the flip and its impact on fairness reported. Second, we need a mechanism to quantify if one data point is more suited to flipping its label than another. Third, training data typically consists of a large number of data points and evaluating each of the data points for a label flip is computationally expensive.

To address these challenges, we first make two key observations on the training data. First, the trained model has varying levels of *uncertainty* attached to the label prediction for each training data point. Flipping the label of a data point that the model is less certain about might result in a more accurate model than one that the model is more certain about. Second, even though the model may be less certain about a data point, flipping its label might not impact the model's fairness. We incorporate these observations to propose several solutions to rank the order in which the labels of training data points should be flipped.

Our main **contributions** are summarized as follows:

- We formalize the problem of ordering label flipping of training data points to improve the fairness of machine learning models.
- We propose ranking strategies to generate an effective ordering in which the labels of training data points should be flipped.
- We conduct extensive experimental evaluation on real-world datasets to demonstrate the effectiveness of our solutions.

The rest of the paper is organized as follows. Section 2 introduces the data notation and preliminaries, In Section 3, we present our proposed label flipping strategies. Section 4 presents our experimental evaluation. We discuss the related works in Section 5, and we conclude in Section 6.

2 PRELIMINARIES

We present our notation and relevant background information on classification and algorithmic fairness.

Classification. We consider the problem of binary classification. Consider a training dataset $\mathsf{D} = d_i^n = \{x_i, y_i\}_{i=1}^n \in \mathsf{Dom}(\mathbf{X}) \times \mathsf{Dom}(Y)$ where $x_i \in \mathsf{Dom}(\mathbf{X})$ with $\mathbf{X}$ denoting a set of features, and $y_i \in \mathsf{Dom}(Y) = \{0, 1\}$ denoting a binary label to be predicted. The objective of binary classification is to train a classifier $h : \mathsf{Dom}(\mathbf{X}) \rightarrow \hat{Y}$ on D such that each data point $\mathbf{x}$ has an associated predicted label $\hat{y} = h(\mathbf{x}) \in \{0, 1\}$. We evaluate the performance of h on $\mathsf{D}_{test} \in \mathsf{Dom}(\mathbf{X}) \times \mathsf{Dom}(Y)$. Classifiers use a learning algorithm that trains on D to learn the optimal parameters $\theta^* \in \mathbb{R}^p$ that minimize the empirical loss $\mathsf{L}(\mathsf{D}, \theta) = \frac{1}{n} \sum_{i=1}^n \mathsf{L}(d_i, \theta)$. We consider learning algorithms that use a loss function L that is strictly convex and is twice-differentiable. In this paper, we focus on logistic regression, which is one of the simplest such classifiers.

Algorithmic group fairness. Given a binary classifier $h : \mathbf{X} \rightarrow \hat{Y}$ and a protected attribute $S \in \mathbf{X}$ (such as age), we denote a favorable prediction by $\hat{Y} = 1$ and an unfavorable prediction by $\hat{Y} = 0$. We assume the domain of S, $\mathsf{Dom}(S) = \{0, 1\}$ where $S = 1$ indicates a privileged and $S = 0$ indicates a protected group (e.g., males and non-males, respectively). Group fairness mandates that individuals belonging to different groups must be treated similarly. The notion of similarity in treatment is captured by different associative notions of fairness such as demographic parity, predictive parity and equalized odds [8, 29, 36]. We focus on demographic parity (a.k.a. *statistical parity*), which is a widely used notion of group fairness. A classifier h satisfies statistical parity if both the protected and the privileged groups have the same probability of being predicted the favorable outcome i.e., $P(\hat{Y} = 1|S = 0) = P(\hat{Y} = 1|S = 1)$. We denote the chosen fairness metric by f and quantify the fairness in the predictions of a model trained on D by f_{D}. For example, $f_{\mathsf{D}} = P(\hat{Y} = 1|S = 0) - P(\hat{Y} = 1|S = 1)$ quantifies demographic parity difference. If $f_{\mathsf{D}} < 0$, the model is biased against the protected group, while $f_{\mathsf{D}} > 0$ indicates the model is biased against the privileged group. A lower $|f_{\mathsf{D}}|$ indicates a fairer model.

Problem statement. Given a binary classifier h trained on D and fairness metric f evaluated on D_{test}, we address the problem of determining the order in which labels in D should be flipped such that a model trained on the modified data results in a lower $|f|$.

3 LABEL FLIPPING ALGORITHMS

The naïve approach of ranking training data points evaluates data points according to their impact on model fairness upon label flip. For each data point in the training data, this method flips the label one at a time and evaluates the change in model fairness before and after the flip. The labels of data points are then flipped in decreasing order of their effect on model fairness. This approach has a high computational complexity $O(|\mathsf{D}| * M_{train})$ where M_{train} is the time taken to retrain a model upon flipping a single label.

To expedite this process, we present several label flipping algorithms that determine the order in which training data labels should be flipped so as to reduce model bias by the most.

3.1 Entropy-based ranking

This section presents a ranking algorithm based on the concept of *entropy* [34] that quantifies the average information content in a data point. Entropy is a way to measure the level of uncertainty in probabilistic objects. In supervised settings, data point d_i is a probabilistic object whose class label ranges over all of its possible class values $\mathsf{Dom}(Y)$. We define the entropy of d_i as:

$$H_i = \sum_{k=1}^{|\mathsf{Dom}(Y)|} p_i^k \log p_i^k$$

where p_i^k is the model's prediction probability for class k. A data item that has a low entropy has a higher degree of certainty (i.e., the model predicted some class as being correct with a high probability) compared to a data item having classes that are almost equally likely to be predicted correct. A low entropy also covers the case when the model incorrectly predicts a class with a high probability. Using this uncertainty measure, we identify the next data point to flip label as one that has the highest entropy. While this approach is relatively fast ($O(|\mathsf{D}|)$ complexity), data points are evaluated individually without considering their impact on other data points.

3.2 Decision-theoretic expected utility

Entropy-based ranking, although computationally inexpensive, determines data points to flip labels one data item at a time and does not consider potential dependencies among data points, therefore, offering no guarantee on improving fairness of the learned model. Our objective is to identify the best data item flipping which would benefit overall fairness of the model. To this end, we design a decision-theoretic ranking method that identifies data points most likely to improve model fairness upon label flip.

We define the *utility* of our model in terms of its fairness. The higher the fairness (lower the unfairness), the higher the model's utility. In the presence of labeling errors, we rely on the *value of perfect information* [33] which measures the expected gain in

the utility function earned by whether or not the label is flipped. We present Expected Utility (denoted by ExpU), a framework that integrates the utility function with the concept of value of perfect information as:

$$\mathsf{ExpU}(d_i) = (p_i f_{\mathsf{D}} + (1 - p_i) f_{\mathsf{D}'}) - f_{\mathsf{D}} \quad (1)$$

where f_{D} is the fairness of the model on the original dataset and p_i is the prediction probability for the original label of d_i, and $f_{\mathsf{D}'}$ is the model fairness when the label of d_i is flipped.

We identify the next data point to flip label as one that has the highest expected utility. Data points identified thus result in lowest unfairness irrespective of whether or not the label is flipped. However, this approach is computationally expensive since we need to retrain the model with flipping the label of each data point.

3.3 Impact approximation

This section presents a ranking algorithm that evaluates data points based on their impact on overall model fairness in case of a label flip. However, unlike the naïve approach of retraining the model after a label flip, we *estimate* the impact of the flip on model fairness. To estimate this impact, we leverage the concept of *influence functions* [21] which approximates the change in model parameters upon an infinitesimal change in the underlying training data.

Given that θ^* minimizes empirical risk i.e., $\theta^* = \mathrm{argmin}_{\theta\in\Theta} \mathsf{L}(\mathsf{D}, \theta) = \mathrm{argmin}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \mathsf{L}(d_i, \theta)$, we denote the gradient of the loss function by $\nabla_\theta \mathsf{L}(\theta)$ and its Hessian matrix by $\mathsf{H}_\theta = \nabla^2_\theta \mathsf{L}(\mathsf{D}, \theta) = \frac{1}{n}\sum_{i=1}^{n} \nabla^2_\theta \mathsf{L}(d_i, \theta)$. Since $\mathsf{L}(\mathsf{D}, \theta)$ is convex and twice-differentiable (Section 2), H_θ is positive definite and therefore, H^{-1} exists.

The influence of flipping the label of training data point $d_i = (x_i, y_i)$ by a small amount ϵ on model parameters θ is computed as:

$$\mathsf{Inf}_\theta(d_i) = \left(\frac{-\mathsf{H}^{-1}}{n}\right)\left(\nabla_\theta \mathsf{L}((x_i, y_i^\delta), \theta^*) - \nabla_\theta \mathsf{L}((x_i, y_i), \theta^*)\right) \quad (2)$$

where y_i^δ denotes the flipped label of data point d_i. More details on influence functions can be found in [22].

We use influence functions to *estimate* the impact I_i of data point d_i by approximating the change in the fairness metric f due to the label flip d_i through the chain rule of differentiation as:

$$\mathsf{I}_i = -\nabla_\theta f(\theta^*) \mathsf{Inf}_\theta(d_i) \quad (3)$$

Having estimated the impact of individual data points, we rank them in decreasing order of their estimated impact on fairness and flip the labels of data points in that order.

3.4 Additional considerations

Note that as training data points are relabeled, the size of the dataset does not change. As detailed below, there a few additional considerations in the relabeling process:

Stopping criterion. Flipping the labels of data points changes the distribution of favorable and unfavorable outcomes for the sensitive groups in the training dataset. At some point, this change effects a reversal of fairness violation i.e., the model exhibits unfairness toward the privileged group. To avoid this behavior, we flip labels only until the fairness of the updated model is below a threshold τ. In Section 4, we do not enforce this stopping criterion and present the result of label flipping for the entire dataset.

Effect on model accuracy. Our ranking strategies are based solely on fairness and consequently, do not guarantee any improvement in model accuracy as a result of the label flips.

4 EXPERIMENTAL EVALUATION

In this section, we present our experimental setup (including the datasets, metrics, and competing methods) along with the effectiveness and efficiency of the methods over the datasets.

4.1 Experimental Setup

Datasets. We evaluated the label flipping algorithms on four real-world datasets popular in the fair ML literature:

GermanCredit [18]. This dataset contains financial and demographic information for 1,000 individuals. The classification task predicts whether individuals are good credit risks; the sensitive attribute is age (age>45 considered privileged).

AdultCensus [28]. This dataset contains demographic and employment information of 48k individuals and is used to predict whether an individual's annual income exceeds 50k. The sensitive attribute is sex (sex=male considered privileged).

COMPAS [25]. This dataset contains demographic information and criminal history of 7,214 defendants. The prediction task assesses a criminal defendant's likelihood to re-offend in the future; the sensitive attribute is race (race=Caucasian considered privileged).

SQF [1]. This dataset contains demographic information of individuals in New York City considered for questioning in accordance with NYC's stop, question, and frisk policy. The classification task predicts whether an individual would be stopped and questioned; the sensitive attribute is race (race=Caucasian considered privileged).

ACSIncome [11]. This dataset is similar to AdultCensus and contains demographic and financial information for over 1.6 million individuals. The classification task predicts if an individual earns more than 50k annually and the sensitive attribute is sex (sex=Male considered privileged). Due to the size and complexity of this dataset, we evaluated our solutions on individual states (e.g., Nevada, Iowa).

Fairness Metrics. Without loss of generality, we evaluated fairness or bias of the model predictions in terms of statistical parity [29]; our algorithms are also applicable to other associative fairness metrics e.g., predictive parity, equalized odds etc.

Competing methods. We consider the following ranking algorithms to determine which labels to flip.

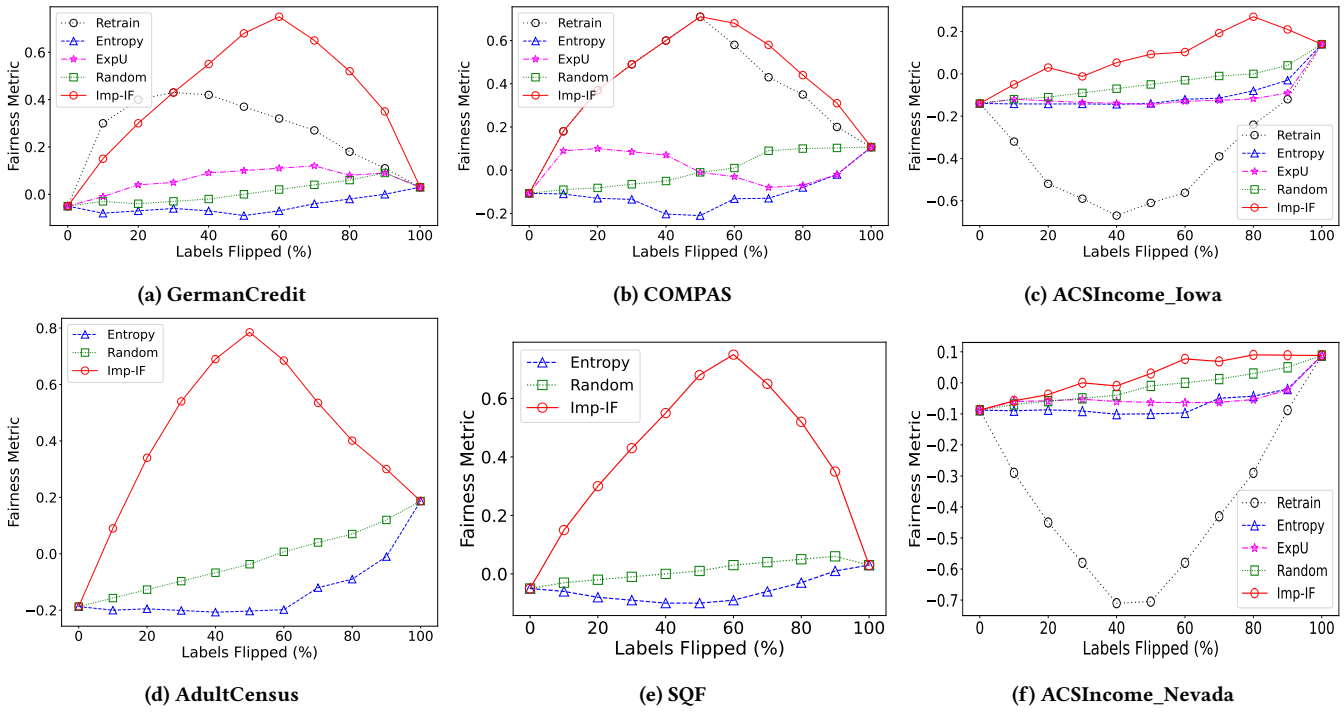
Random. This method randomly selects training data points to flip their labels.

Retrain. This naïve method iteratively selects the data point which, when assigned a flipped label and a model retrained, results in the most reduction in bias.

Entropy. This method (described in Section 3.1) selects data point that the model is least certain about.

Imp-IF. This method (described in Section 3.3) ranks data points according to their *estimated* effect on fairness upon flipping labels; data points with higher influences are prioritized for flipping.

ExpU. This is our decision-theoretic method (described in Section 3.2) that identifies a data point that would reduce model bias by the most with either label.



(a) GermanCredit (b) COMPAS (c) ACSIncome_Iowa

(d) AdultCensus (e) SQF (f) ACSIncome_Nevada

Figure 1: Comparison of different label flipping strategies. ExpU and Imp-IF consistently reduce model unfairness whereas the other methods do not always improve model fairness.

Performance metrics. We compare the competing methods according to two primary performance metrics:
Effectiveness: To evaluate the effectiveness of the ranking algorithms, we flipped the labels of data points individually in the order specified by a method and recorded the fairness metric upon each label flip. A value closer to 0 indicates that the modified data results in fairer and less biased decisions.
Efficiency: We report the average time taken by a method to determine the next data point whose label should be flipped.

4.2 Effectiveness of label flipping strategies

In this section, we evaluate the effectiveness of the competing methods in improving the fairness of the downstream machine learning model. Our goal in this experiment is to assess the methods in terms of the fraction of data points whose labels should be flipped to reduce model unfairness. We demonstrate in Figure 1, the gradual improvement in fairness (demographic parity in y-axis) for increasing the fraction of flipped labels (x-axis) for all the methods. Initially, across datasets demographic parity is negative indicating that the learned model is biased against the protected group. The closer the line for a method is to 0, the fairer the learned model is. We were able to run all the methods for most of the methods; however, computationally expensive methods, Retrain and ExpU, took more than 2 hours to complete and are not included here for the larger datasets (AdultCensus and SQF).

As seen for GermanCredit, COMPAS, AdultCensus, SQF in Figures 1(a)-1(e), we observe that all methods except Entropy exhibit an upward trend indicating an improvement in model fairness as more labels are flipped. This behavior is not surprising as Entropy determines the data points flipping the labels of which would reduce uncertainty but does not guarantee any impact on model fairness. While Random improves model fairness, the improvement is pretty slow, requiring close to 50% of label flips for achieving parity. On the other hand, Retrain is able to identify data points that should be relabeled but is prohibitively expensive (and we are unable to leverage it for our larger datasets AdultCensus and SQF). Notably, ExpU has much faster improvement in model fairness for GermanCredit and COMPAS: this is because the method computes expected model fairness whether there is a label flip and selects the data point that results in the lowest fairness. Irrespective of whether the label is flipped, the model will have higher fairness than before. We observe the steepest improvement in model fairness by Imp-IF: this behavior is consistently demonstrated across datasets. For most of the datasets, Imp-IF reaches parity (i.e., 0% fairness metric value) by flipping the labels of less than 5% of datasets, which is a promising result because by estimating the impact of individual training data points on model fairness, Imp-IF is able to effectively identify fewer than 5% of training data points that are potentially mislabeled and if corrected, resolve the issue of model unfairness.

Especially for the larger datasets (Figures 1(d) and 1(e)), we observe that Imp-IF exhibits the most rapid improvement in fairness compared to the other computationally inexpensive methods. Across datasets, as the labels of more data points are flipped, we observe that Imp-IF keeps increasing fairness until it reaches parity

	Imp-IF	Retrain	Entropy	ExpU	Random
GermanCredit	0.7	73	0.08	69.92	0.02
SQF	9.19	-	0.23	-	0.03
COMPAS	2.44	576	0.1	565	0.01
AdultCensus	17.06	-	0.44	-	1.44
ACSI_Nevada	4.89	1531	0.15	1537	0.02
ACSI_Iowa	6.05	2053	0.21	1993	0.02

Table 1: Time taken for each solution (in seconds)

and then results in fairness favoring the unprivileged group (fairness metric becomes positive). As discussed in Section 3, label flips do not guarantee an improvement in accuracy. Consistent with prior studies on the trade-off between accuracy and fairness, as we reach parity (statistical disparity=0), the model indeed displays a deterioration in accuracy ($4.8 - 44\%$) across ranking strategies. Optimizing for fairness and accuracy simultaneously is a line of research that we defer for future work.

Takeaways: (1) Improving the quality of training data by identifying labeling errors is an effective strategy to reduce model unfairness. (2) Valuation-based label flipping Imp-IF rapidly improves model fairness by flipping the labels of very few data points.

4.3 Efficiency of label flipping strategies

Our goal in this set of experiments is to evaluate the scalability of the different labeling strategies to larger datasets. In Table 1, we show the time taken by the different methods in selecting the appropriate data points for label flipping. We observe that our gold standard Retrain is prohibitively expensive and cannot be applied to large datasets. The label flipping strategy employing expected utility ExpU that guarantees lower unfairness has comparable time performance as Retrain. On the other hand, Imp-IF is orders of magnitude faster than these two methods. The time for Imp-IF includes the one-time expensive offline computation for the Hessian matrix, which contributes the most to Imp-IF time, and is hence slower than Entropy which is computed over the learned model's predicted probabilities.

Takeaways: (1) Imp-IF efficiently scales up to large datasets in the online computation of data points whose label should be flipped. (2) The one-time offline computation of Imp-IF, while slower than Entropy and Random, is orders of magnitude faster than naïve model retraining (Retrain) which has an extreme computation cost and cannot be used on large datasets.

5 RELATED WORK

The study in this paper is related to the following research areas: algorithmic fairness and data quality (including label errors). These areas were studied extensively, but our approach of exploring label flipping strategies to ensure fairer algorithmic systems is novel.

Algorithmic Fairness. Intensive research on algorithmic fairness has resulted in a number of fairness metrics and bias mitigation techniques [29, 36]. Discriminatory behavior, quantified through *fairness* metrics in the algorithmic literature, is broadly categorized as *individual fairness*, *group fairness* and *causal fairness*. Individual fairness [12, 38] states that similar individuals must be treated similarly, group fairness [30, 36] mandates parity between individuals belonging to different sensitive groups (e.g., males vs. non-males, Asians vs. non-Asians). Causal fairness studies whether features have a causal effect on the fairness of outcomes [7, 24]. These notions of fairness are orthogonal to each other; in this work, we focus on group fairness (further detailed in Section 2). Bias mitigation techniques can broadly be categorized as *pre-processing*, *in-processing*, and *post-processing* techniques. Pre-processing techniques [6, 19, 20, 23, 26, 39] modify the underlying data before training a model, in-processing techniques [37] modify the model's learning algorithm to satisfy fairness constraints, and post-processing techniques [13, 15, 40] update the model predictions after training. While in-processing has been shown to largely resolve fairness issues, it requires knowledge of and tweaking the model's formulation. In contrast, pre-processing assumes biased data to be the root of model unfairness, does not need the knowledge of model internals, and has been shown to alleviate unfairness. Our work deals with flipping training data labels and, therefore, is most related to pre-processing bias mitigation techniques.

Data errors. Correcting data errors has been a focus of data management and data cleaning research for a long time. Recent work has started to focus on label bias and the fact that erroneous labels exist in datasets due to human bias [9, 40]. Research has also been conducted to fix the issue of erroneous labels through a multitude of ways [5, 10, 17, 35]. Our work focuses on label flipping for achieving group fairness and differs from [40] that focuses on flipping the labels of data points in the setting of individual fairness. Our work can also be seen as related to the area of active learning that aims to acquire labels for unlabeled data in a fairness setting [2]; we seek to correct potentially mislabeled data.

6 CONCLUSION AND FUTURE WORK

We presented a novel pay-as-you-go approach to determine which data points should be relabeled such that a machine learning model learned on the updated data generates fairer decisions. To the best of our knowledge, the present work is the first to propose label flipping strategies for the task of improving group fairness of machine learning models. We first presented a strategy that assesses data points individually by considering their local characteristics, and then presented a decision-theoretic solution that evaluates data points by their impact on overall model fairness. To expedite the process, we presented a strategy that assesses data points by their *estimated* impact on model fairness computed through influence functions. Experimental evaluation on standard real-world datasets in the fair ML literature highlights the effectiveness of our ranking strategies in reducing model unfairness without necessitating extensive model retraining. Incorporating data valuation using influence functions provides a promising direction for improving model fairness by rectifying potentially mislabeled data. In the future, we intend to develop solutions to highlight systemic errors that occur in specific subpopulations. Future work also includes designing human-in-the-loop solutions to diagnose and correct data errors in end-to-end machine learning pipelines.

REFERENCES

[1] [n.d.]. Publications, reports - NYPD. https://www.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page

[2] Hadis Anahideh, Abolfazl Asudeh, and Saravanan Thirumuruganathan. 2022. Fair active learning. *Expert systems with applications* 199 (8 2022), 116981. https://doi.org/10.1016/j.eswa.2022.116981

[3] Fabio Azzalini, Chiara Criscuolo, and Letizia Tanca. 2022. E-FAIR-DB: Functional Dependencies to Discover Data Bias and Enhance Data Equity. *J. Data and Information Quality* 14, 4, Article 29 (Nov. 2022), 26 pages. https://doi.org/10.1145/3552433

[4] Denis Baylor, Eric Breck, Heng-Tze Cheng, Noah Fiedel, Chuan Yu Foo, Zakaria Haque, Salem Haykal, Mustafa Ispir, Vihan Jain, Levent Koc, Chiu Yuen Koo, Lukasz Lew, Clemens Mewald, Akshay Naresh Modi, Neoklis Polyzotis, Sukriti Ramesh, Sudip Roy, Steven Euijong Whang, Martin Wicke, Jarek Wilkiewicz, Xin Zhang, and Martin Zinkevich. 2017. TFX: A TensorFlow-Based Production-Scale Machine Learning Platform. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS, Canada) *(KDD '17)*. Association for Computing Machinery, New York, NY, USA, 1387–1395. https://doi.org/10.1145/3097983.3098021

[5] Carla E. Brodley and M. A. Friedl. 1999. Identifying mislabeled training data. *Journal of artificial intelligence research/The journal of artificial intelligence research* 11 (8 1999), 131–167. https://doi.org/10.1613/jair.606

[6] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. 2009 IEEE International Conference on Data Mining Workshops. https://doi.org/10.1109/ICDMW.2009.83

[7] Silvia Chiappa. 2019. Path-Specific Counterfactual Fairness. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (Jul. 2019), 7801–7808. https://doi.org/10.1609/aaai.v33i01.33017801

[8] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *CoRR* abs/1703.00056 (2017).

[9] Jessica Dai and Sarah Brown. [n.d.]. Label bias, label shift: fair machine learning with unreliable labels. https://dynamicdecisions.github.io/assets/pdfs/29.pdf

[10] Kyriaki Dimitriadou, Rahul Manghwani, and Timothy Hoad. 2019. ClusterClean: a weak Semi-Supervised approach for cleaning data labels. 2019 IEEE International Conference on Big Data (Big Data). https://doi.org/10.1109/BigData47090.2019.9006371

[11] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring adult: new datasets for fair machine learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS '21)*. Curran Associates Inc., Red Hook, NY, USA, Article 496, 13 pages.

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

[13] Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2014. Certifying and removing disparate impact. https://arxiv.org/abs/1412.3756

[14] John B. Ford, Varsha Jain, Ketan Wadhwani, and Damini Goyal Gupta. 2023. AI advertising: An overview and guidelines. *Journal of business research* 166 (11 2023), 114124. https://doi.org/10.1016/j.jbusres.2023.114124

[15] Vladimiro González-Zelaya, Julián Salas, David Megías, and Paolo Missier. 2023. Fair and Private Data Preprocessing through Microaggregation. *ACM transactions on knowledge discovery from data* 18, 3 (12 2023), 1–24. https://doi.org/10.1145/3617377

[16] Katherine Haan. 2023. How businesses are using artificial intelligence in 2024. (4 2023). https://forbes.com/advisor/business/software/ai-in-business/

[17] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. 2018. Using trusted data to train deep networks on labels corrupted by severe noise. https://proceedings.neurips.cc/paper_files/paper/2018/hash/ad554d8c3b06d6b97ee76a2448bd7913-Abstract.html

[18] Hans Hofmann. 1994. Statlog (German Credit Data). https://doi.org/10.24432/C5NC77

[19] Libin Jiang and Jean Walrand. 2010. A distributed CSMA algorithm for throughput and utility maximization in wireless networks. *IEEE/ACM transactions on networking* 18, 3 (6 2010), 960–972. https://doi.org/10.1109/tnet.2009.2035046

[20] Faisal Kamiran and Toon Calders. 2011. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems* 33, 1 (12 2011), 1–33. https://doi.org/10.1007/s10115-011-0463-8

[21] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. https://arxiv.org/abs/1703.04730

[22] Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70. 1885–1894.

[23] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Ioannis Kompatsiaris. 2018. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. WWW '18: Proceedings of the 2018 World Wide Web Conference. https://doi.org/10.1145/3178876.3186133

[24] Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4069–4079.

[25] Jeff Larson, Marjorie Roswell, and Vaggelis Atlidakis. [n.d.]. COMPAS. https://github.com/propublica/compas-analysis/

[26] Peizhao Li and Hongfu Liu. 2022. Achieving Fairness at No Utility Cost via Data Reweighing with Influence. https://proceedings.mlr.press/v162/li22p

[27] Lacramioara Mazilu, Norman W. Paton, Nikolaos Konstantinou, and Alvaro A. A. Fernandes. 2022. Fairness-aware Data Integration. *J. Data and Information Quality* 14, 4, Article 28 (Nov. 2022), 26 pages. https://doi.org/10.1145/3519419

[28] Meek Meek, Thiesson Thiesson, and Heckerman Heckerman. 1990. US census data (1990). https://doi.org/10.24432/C5VP42

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

[30] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6, Article 115 (jul 2021), 35 pages. https://doi.org/10.1145/3457607

[31] Andrej Miklošík, Nina Evans, and Athar Mahmood Ahmed Qureshi. 2021. The Use of Chatbots in Digital Business Transformation: A Systematic Literature review. *IEEE access* 9 (1 2021), 106530–106539. https://doi.org/10.1109/access.2021.3100885

[32] Jake Piazza. 2023. Apple, caught by surprise in generative AI boom, to spend $1 billion per year to catch up: Report. (102023).

[33] Stuart Russell and Peter Norvig. 2009. *Artificial Intelligence: A Modern Approach* (3rd ed.). Prentice Hall Press, USA.

[34] C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27, 3 (1948), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[35] Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? Improving data quality and data mining using multiple, noisy labelers. KDD '08: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. https://doi.org/10.1145/1401890.1401965

[36] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*. 1–7.

[37] Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. 2023. In-Processing Modeling Techniques for Machine Learning Fairness: A survey. *ACM transactions on knowledge discovery from data* 17, 3 (3 2023), 1–27. https://doi.org/10.1145/3551390

[38] Xiaomeng Wang, Yishi Zhang, and Ruilin Zhu. 2022. A brief review on algorithmic fairness. *Management system engineering* 1, 1 (11 2022). https://doi.org/10.1007/s44176-022-00006-z

[39] Bobby Yan, Skyler Seto, and Nicholas Apostoloff. 2022. FORML: Learning to Reweight Data for Fairness. http://arxiv.org/abs/2202.01719

[40] Hantian Zhang, Ki Hyun Tae, Jaeyoung Park, Cheng Xu, and Steven Euijong Whang. 2023. IFlipper: Label flipping for individual fairness. *Proceedings of the ACM on management of data* 1, 1 (5 2023), 1–26. https://doi.org/10.1145/3588688