Label Bias, Label Shift: Fair Machine Learning with Unreliable Labels

Jessica Dai
Brown University
Providence, RI 02912
jessica.dai@brown.edu

Sarah M. Brown
University of Rhode Island
Kingston, RI 02881
brownsarahm@uri.edu

Abstract

Most supervised learning problems assume that the data available for training is well-representative of the data on which the model will be deployed. Distribution shift is a fairly robust subfield within machine learning more broadly, but there has been little work that integrates findings from the distribution shift literature to fair machine learning, despite an increasing interest in and awareness of long-term dynamics created by fair machine learning systems. In settings where fairness is a concern, there are additional reasons for training data to be unrepresentative of the true data: the available data may have been generated through a process reflecting discrimination, such as the systematic mislabeling of positive examples from a specific group. In this work, we build upon work in both fairness and distribution shift to examine the performance of fair machine learning models when the reliability of labels is uncertain and dynamic, focusing on *label bias* as the bias model, and *label shift* as the mechanism of distribution shift. First, we present a framework for approaching and understanding distribution shift problems in the context of fairness. Then, motivated by real-world needs, we consider two scenarios for distribution shift: (i) no access to new labeled data, only new model inputs; and (ii) access to new (potentially-biased) labels. Our experimental results suggest that the combination of distribution shift and label shift may be a plausible failure mode for fair algorithms, indicating the relationship between distribution shift and learned fair models is an important area of continued study.

1 Introduction

When decisions with consequences for the trajectory of human lives are made or recommended algorithmically, it is imperative that stakeholders—decisionmakers, regulators, and those about whom decisions are being made—can be confident about the performance of the algorithm. Though the necessity of socially responsible machine learning systems has been obvious for several years, only recently has fair machine learning scholarship begun to explore many of the fundamental assumptions of the field. While a plethora of definitions, algorithms, and metrics for what constitutes "fair" machine learning have been developed (see, for example survey paper [10]), those approaches tend to focus on the algorithm in isolation, addressing statistics calculated only on the input and output of the individual algorithm at a single timestep. Recent work in fair machine learning, however, has emphasized the importance of (re)considering the algorithm in its broader context. A key component of contextualizing these algorithms is building confidence in the robustness of algorithm performance. An algorithm used responsibly cannot solely be *fair* (however fairness may be defined) at time of development—it must also be trusted to consistently and reliably perform well.

In this work, we consider the performance of a deployed "fair" algorithm under conditions of distribution shift. While long-term dynamics of fair decisionmaking systems have recently garnered

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

substantial interest [18, 5], the simpler case of a single shift between training and deployment has received surprisingly little attention.

One of the most common—and well-studied—reasons for real-world data to diverge from the training data is distribution shift, where $P(X)_{\text{train}} \neq P(X)_{\text{test}}$, or $P(Y)_{\text{train}} \neq P(Y)_{\text{test}}$. In contexts where fairness is relevant, however, there is a second major reason for potential unrepresentativeness of training data: the reliability of the available data itself, where features, labels, or sampling for particular subgroups can be systematically incorrect [7, 28, 3]. In this work, we focus on *label shift* as the mechanism of distribution shift—that is, $P(Y)_{\text{train}} \neq P(Y)_{\text{test}}$—and *label bias* as the model for bias—that is, the available labels for particular subpopulations are systematically incorrect. Distribution shift is a critical issue in the context of fairness-relevant domains, and the combination of *label shift* with *label bias* arises naturally: machine learning for the purposes of disease diagnosis, for example, is a common case study for distribution shift problems, and one can imagine that the generation of the training labels for such models can be skewed by systematic misdiagnosis for certain population subgroups (for example, women and people of color [21]).

In this work, we first propose a framework for approaching and understanding problems where the unreliability of data may be twofold—due to systemically biased data (label bias) *and* to potential changes in distribution (label shift). Then, we provide results for a series of experiments motivated by real-world needs, asking *to what extent are existing algorithms robust to label bias and label shift?* While in most settings, existing fair approaches are surprisingly robust to both label bias and label shift, a handful of failure modes highlight the importance of careful construction of models for bias.

2 Related work

Distribution shift is a well-known problem space in machine learning. [20] provide a canonical taxonomy of the main types of shift, including covariate (a change in $P(X)$), label (a change in $P(Y)$), concept (a change in $P(Y|X)$ or $P(X|Y)$), and "other"; they also provide explanations for causes of shift, such as systemic under-sampling. Of primary interest in our work is *label shift*, where $P(Y)$ changes and $P(X|Y)$ stays the same. Specifically, label shift occurs in $y \rightarrow x$ problems—for example, disease diagnosis or image classification. Recent strategies for detection and mitigation of label shift without requiring retraining of the entire classifier are proposed by [12], who require that the probability mass function of y in the test setting is known, and [17]'s Black Box Shift Estimation (BBSE), which requires only the predicted outputs at test time. [23] provide an empirical evaluation of several methods for detecting dataset shift, though their work is not exclusive to label bias.

In *fair machine learning*, one well-developed line of work involves the idea that (available) data does not reflect "ground truth"; generally, unrepresentativeness is contextualized as one of the reasons available data and algorithms trained on such data is unfair. [3] identify several distinct reasons that data collected can be unrepresentative, including *label bias*, where labels for subsets of the population are incorrect; *feature bias*, where feature values for subsets of the population are incorrect; and *sample bias*, where some subsets of the population are systematically under- or over- sampled. Our work draws from the approach of [14], who find that fairness-adjusted classifiers that may *appear* fair on the available data can actually be unfair on the true data, due to the systematic censoring of training data; they successfully apply methods typically used in covariate shift to improve the calculation of fairness metrics. Other work that develops the idea of available vs true datasets, including many preprocessing approaches to achieving fairness, includes [7], [6], [8], and [28].

We focus on *label bias* as the mechanism by which bias is introduced to the data. In the binary classification case, this can be modeled by the systematic flipping of labels belonging to particular population subgroups; for example, positive examples from the disadvantaged group may be mislabeled as negative, or negative examples from the advantaged group may be mislabeled as positive. Recent work has emphasized that the often-cited "fairness-accuracy" tradeoff [19, 29, 10] may disappear once accuracy is measured against some hidden true dataset rather than the original available data: [2] find that the lowest-risk equal-opportunity-constrained classifier on the available biased data is *also* the lowest-risk classifier on the true data. [9] place bounds on the relationship between the observed and true value of metrics calculated with possibly incorrect labels. [13] first rewrite common notions of fairness as constraint functions, then use those constraints to define the relationship between true and available labels. They find, with theoretical and empirical results, that a simple reweighing approach to training on the biased labels approximates training on the true

labels. Similarly, [27] contribute a semi supervised learning method, where outcomes on unlabeled data are used as a regularizer on the weights learned from the labeled data. Of note for all of these results relating label bias to the fairness-accuracy tradeoff is an assumption on the true underlying data—namely, that any observed discrepancy in base rates $P(\textit{label}|\textit{group})$ is due to label bias, and that the true underlying base rates are equal across groups.

Work *applying ideas from distribution shift to fairness settings* is relatively new, and to date has mostly focused on covariate shift and/or domain adaptation and transfer learning, where the goal is to re-purpose a pre-trained fair classifier for an entirely new task or for an entirely different protected attribute rather than to ensure the continued performance of the pre-trained classifier on the same task [25, 24, 4]. [26] introduce *fairness warnings*, which are designed for domain adaptation—the warnings identify which changes to the features of the data might cause the "fair" algorithm to fail on the new setting. The warnings are generated by a meta-model which learns what shifts may cause fairness to fail. The spirit of our work is similar, in that we hope to provide insight towards auditing and understanding the fairness performance of a pre-trained classifier, but (i) we focus on a change in the label distribution rather than in the means of the features, and (ii) we are primarily concerned about distribution shifts where the classifier continues to be used for the same task, rather than repurposed for a separate domain.

3 Problem setting

In this work, we focus on label bias, though the general framework proposed in this section can be adapted for other scenarios where some underlying "true" data distribution is transformed to the "biased" dataset available for training. Our first contribution is a general conceptual framework for understanding distribution shift in a fairness context.

3.1 General framework

The modelling of *bias* in a dataset and the modelling of *distribution shift* from train time to test time share several concepts. For example, one cause for shift cited in [20] is systematic undersampling based on a feature value or a label value; this is also a common explanation for why a dataset reflects bias against a particular group.

However, we are not *only* concerned with similarities in the way distribution shift and "unfair datasets" may be structured. Rather, we are concerned about shifts in the true data distribution, in addition to the relationship between the true data and the available data. For clarity, let $t = 0$ represent the time of training, and $t = 1$ represent some new timestep. To return to our motivating example of disease diagnosis from the introduction, *label shift only* would imply that the prevalence of a particular disease is different at $t = 1$ than at $t = 0$. *Label bias only*, on the other hand, suggests that at any given time, for women that *should* have a positive diagnosis, some percentage of them are misdiagnosed as negative. Our model investigates the setting where *both* of the above occur.

3.2 Unified model of label bias

First, we unify several related models for label bias. While all of the work cited in Section 2 involve systematic mislabellings of particular population subgroups, the specific way label bias is modelled varies. Our model, which encompasses most previously-proposed models for label bias, is as follows. The true underlying distribution $\mathcal{D}$ is defined by (X, S, Z), where X is the vector of permissible attributes; S is the sensitive attribute(s); and Z is the hidden *true* label. Let $S = \mathbf{A}$ indicate membership in the *privileged* class, $S = \mathbf{B}$ indicate membership in the *disadvantaged* class and $Z = 1$ indicate the *desirable* outcome. Furthermore, assuming binary classification, let $\zeta_A = P(Z = 1|S = \mathbf{A})$ and $\zeta_B = P(Z = 1|S = \mathbf{B})$. Note that a key assumption in most current label bias work is that $\zeta_A = \zeta_B$.

Though selection bias is generally likely or plausible, for simplicity we assume that the dataset we have available does not undersample or oversample any group based on the *true* labels. Rather, bias is introduced into the dataset through the labels only: that is, the bias-generating process is the process of generating biased labels. Let $Y = \mathcal{G}(X, S, Z)$ represent this process, where Y is the label available in the dataset, and the function $\mathcal{G}$ represents how the biased labels were generated.

The simplest version of $\mathcal{G}$ involves *flipping* the labels of certain subsets of $\mathcal{D}$ depending only on S and Z. We can define

$$Y = \mathcal{G}(X, S, Z) = \begin{cases} 1 - Z & \text{with probability } \rho_A \text{ if } S = \mathbf{A}, Z = 0 \\ 1 - Z & \text{with probability } \rho_B \text{ if } S = \mathbf{B}, Z = 1 \\ Z & \text{otherwise.} \end{cases}$$

In other words, *negative* examples from the privileged class are flipped with probability ρ_A, and *positive* examples from the disadvantaged class are flipped with probability ρ_B. In [2], $\rho_A = 0$, as flipping only occurs for positive examples from Group B; in [9], $\rho_B = 0$, as flipping only occurs for negative examples from Group A; in [27], $\rho_A = \rho_B \neq 0$. [1]

3.3 Modeling dataset shift

We assume that we have access to some classifier $\mathcal{F}$, which has been trained to be fair with respect to the *true* labels (i.e. Z) at $t = 0$. There are several parameters that may change from $t = 0$ to $t = 1$: ζ_A, ζ_B, ρ_A, and ρ_B. We propose two scenarios of interest, motivated by real-world auditing needs.

Scenario 1: no access to new labels. That is, we have no access to new labels at $t = 1$, so the nature of the bias model at $t = 1$ is unknown, but the *true* distribution of labels Z (i.e. $\mathcal{D}$) has changed. Practically, this means that ζ_A and/or ζ_B has changed. (Recall that since we are only considering label shift, we assume $P(B)_{t=0} = P(B)_{t=1}$. If $P(B)$ had changed, that would be a special case of covariate shift.) In this scenario, we are primarily interested in the impact of label bias in the training data on classifier performance, specifically under label shift in the test data.

Scenario 2: access to new (biased) labels. Suppose new data is collected for the purposes of validating the model's continued performance at $t = 1$; this new collected data has labels $Y_{t=1}$. Here, we have two possible mechanisms for how shift may have occured.

1. **True distribution changes**. Like in scenario 1, $\mathcal{D}_{t=1} \neq \mathcal{D}_{t=0}$; ζ_A and/or ζ_B have changed.
2. **Bias model changes**. We now have labels $Y_{t=1}$. It may be the case that while the *true* distribution of labels has remained the same, label shift *appears* to have occurred because $\mathcal{G}$ has changed; practically, this means that ρ_A and/or ρ_B may have changed. Changes in ρ technically fall under concept shift (a change in the relationship between X and Y), and are beyond the scope of typical distribution shift work. However, we still consider changes in ρ in this work, as we believe that this is a simple, plausible, and relevant occurrence, and as there are few systematic approaches to resolving concept shift.

In this case, new information from the potentially-biased labels $Y_{t=1}$ raises the additional question of how reliable performance metrics calculated at $t = 1$ may be, in particular, whether it might be possible that $\mathcal{F}$ *appears* fair on $Y_{t=1}$, but in reality is no longer fair on the *true* labels $Z_{t=1}$.

4 Experiments

Datasets We currently consider only a single, binary sensitive attribute.[2] We run experiments on both synthetic data and a modified version of the 1994 Adult Census data. Recall that ζ_G represents $P(Z = 1|S = \mathbf{G})$, where G is group A or B and where Z is the true label. For each dataset type (synthetic or Adult), we generate versions with several possible ζs; a core assumption in most label bias papers is that $\zeta_A = \zeta_B$, so we train models only on the datasets where $\zeta_A = \zeta_B$. Then, for each version of the dataset, we further generate a series of "label-biased" Y-values.

To test the simplest instantiation of our model, we generate a synthetic dataset where $P(S = \mathbf{B}) = 0.3$, and each example has exactly two attributes: a sensitive attribute `a`, and a real-valued score `x0` where the score `x0` is drawn from $\mathcal{N}(5, 1)$ for positively-classified examples (i.e. $Z = 1$) and from $\mathcal{N}(2, 1)$

[1] [13] have a much more complex notion of which labels are likely to flip, corresponding to a particular measure of fairness; however, their model suggests that the specific *fairness metric* determines the nature of label bias, which we find unintuitive.

[2] We are aware of the limitations associated with this choice, and hope this work is a starting point for analyzing distribution shift under more complex (and realistic) versions of label bias.

for negatively-classified examples ($Z = 0$). We create datasets with $\zeta \in [0.2, 0.3, 0.4]$, and for each of those datasets add label-biased versions for $\rho_A = \rho_B \in [0.05, 0.1, 0.15, 0.2, 0.25]$ as well as $\rho_A = 0, \rho_B \in [0.05, 0.1, 0.15, 0.2, 0.25]$. Since ρ represents the probability that labels will be flipped, these sets of parameters test two different models: one where both negative examples from A and positive examples from B are flipped, and one where only positive examples from B are flipped.

For the Adult Census dataset, we use `race` as the sensitive attribute, dropping `sex` from training. We start with the IBM Fairness 360 preprocessing; for reference, this version of the dataset has $P(S = \mathbf{B}) = 0.14$, $P(Y = 1|\mathbf{B}) = 0.158$, $P(Y = 1|\mathbf{A}) = 0.262$ [1]. To construct datasets with our desired ζs, we drop examples from the dataset until our desired statistic is reached. For this dataset, we create datasets with $\zeta \in [0.1, 0.15, 0.2, 0.25]$, and add label-biased versions with $\rho_A = 0, \rho_B \in [0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4]$.

Algorithms For both dataset types, we run the following algorithms using their implementations in the IBM Fairness 360 package.

- **Baseline**—Logistic Regression. For both datasets, Logistic Regression reaches around 80% accuracy. Logistic Regression is also the base classifier for the following methods.
- **Preprocessing**—*Disparate Impact Remover* [7], which edits feature values in order to achieve minimal disparate impact; and *Reweighing* [15], which weights examples in each (group, label) combination.
- **Postprocessing**—*Equal Opportunity* [11], which probabilistically flips predicted labels in order to achieve equalized odds; *Calibrated Equal Opportunity* [22], which is similar in method to [11] but optimizes instead over *calibrated* classifier outputs; and *Reject Option Classification* [16], which flips predicted labels for unprivileged groups originally given unfavorable classifications and privileged groups originally given favorable classifications.

For all of these algorithms, after training the classifier, we also search for the Logistic Regression threshold that gives maximizes balanced accuracy (across positive and negative classes).

Metrics For each experiment, we calculate the average odds difference ($\frac{1}{2}((FPR_B - FPR_A) + (TPR_B - TPR_A))$), balanced accuracy, equal-opportunity difference ($TPR_B - TPR_A$), and statistical-parity difference ($P(\hat{Y} = 1|B) - P(\hat{Y} = 1|A)$).

5 Results & discussion

General sensitivity Interestingly, under the simple perturbations we apply to each of the datasets, these algorithms appear to be fairly robust to both true label shift (as discussed in scenario 1), and to testing on newly biased data (as discussed in scenario 2) in general. For the vast majority combinations of algorithm and metric, the pre-trained algorithm performs well for all combinations of possible training dataset versions and testing dataset versions; this is the case for both datasets. There are a few notable exceptions, which appear to be algorithm-dependent; however, even so, observed differences in metrics either depend solely on the bias model in the *training* data, or solely on the bias model in the *testing* data. In other words, there does not appear to be complex relationships between potentially-different levels of label bias in the training and testing datasets, even under label shift.

The impact of label bias in *training* In scenario 1 from our problem setting (section 3.3), we asked whether fair classifiers able to recover a hypothesis which reflects the *true* data distribution, even if they have been trained on the biased data, and furthermore, whether this hypothesis is robust to shifts in the true test data. While this is generally the case, results from the Calibrated Equalized Odds algorithm suggest that *the specific type of bias model* can have significant impacts on the learned hypothesis, and furthermore that these hypotheses are nevertheless robust to subsequent changes in the testing data. In Figure 2, note that the heatmaps for the two metrics—average odds difference and equal opportunity difference—are roughly split into two sections, one reflecting algorithms trained on data with a bias model of $\rho_A = 0$, and one reflecting a bias model of $\rho_A = \rho_B$. Furthermore, within these two regions, the *extent* of label bias (i.e. value of ρ) has a relatively insignificant impact. Rather, it is the choice in defining the label bias model—whether negative examples from group A are also flipped, rather than just positive examples from group B—that can have a drastic impact on the learned hypothesis.

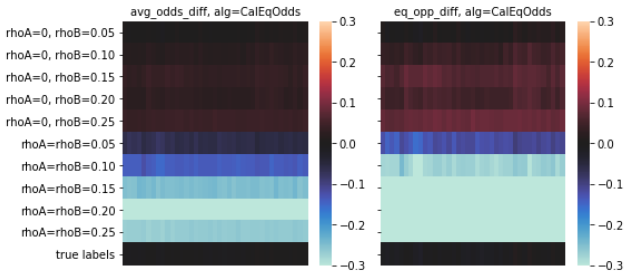

Figure 1: Calibrated Equalized Odds postprocessing algorithm, synthetic data. Vertical axis: training data, along with the specific way in which it was label biased. Horizontal axis: test data. In this plot, reflects *all* combinations of dataset versions noted above, including several different ζs and ρ combinations.

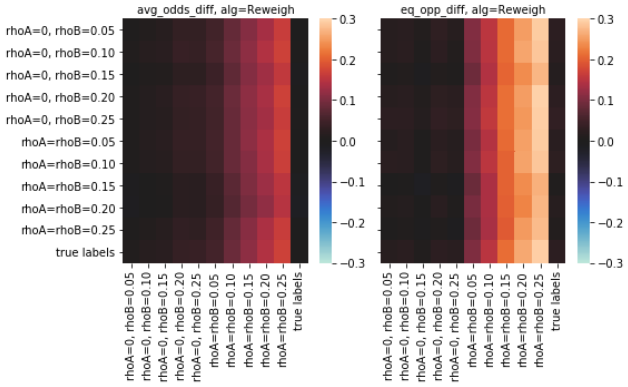

Figure 2: Reweigh preprocessing algorithm, synthetic data. Horizontal axis reflects datasets with constant ζs and various versions of label bias. While the selected results are for $\zeta_A = \zeta_B = 0.3$, the exact same pattern appears for $\zeta_A = \zeta_B = 0.2$ and $\zeta_A = \zeta_B = 0.4$.

The impact of label bias in *testing* In scenario 2 from our problem setting, we hoped to characterize the impact of *evaluating* pre-trained algorithms on potentially label-biased datasets. As above, label bias in testing data, even when also label-shifted, generally does not result in erroneous evaluations. However, with the Reweighing algorithm, there are some versions of label bias where, no matter what version of label bias was present in *training*, an algorithm tested on these versions may appear to no longer be fair even when it is actually fair with regards to the true labels. Most notably, it is exactly the versions of label bias that also impacted the Calibrated Equalized Odds algorithm at train time: specifically, where $\rho_A = \rho_B$.

The importance of understanding the bias model and algorithm These results are at once reassuring—as most algorithms and metrics are more robust than one might have expected—and yet at the same time, raise additional questions about the way in which the dataset is modeled. The phenomena discussed above are *not* universal behavior across all algorithms, emphasizing the importance of fully understanding why and how an algorithm might be "correcting" for fairness. Furthermore, given that all of the recent work on "label bias" have slightly different (though still related) models for bias, one might expect that any of those assumptions would be sufficient to capture the general behavior expected under label bias. Our results suggest that this is not necessarily the case, and additionally that the ad-hoc choice of one model over another can have significant impacts on algorithm performance. Finally, "label bias" is already a highly simplified and stylized interpretation of how bias can be captured in a dataset; further work involving such models must explore the implications of different parameter settings for the bias model.

Broader Impact

This work is intended as an exploration of the performance of fair classifiers when distribution shift has occurred. We hope this paper raises new, relevant questions in the field of fair machine learning that are not currently addressed in the literature: we believe that understanding and validating the continued performance of "fair" classifiers is an urgent and important problem. However, in our work, the modelling stage involves the introduction of strong assumptions about the data, especially what defines "bias" for the specific dataset. No dataset can fully capture the nuances of the real world, and no "model for bias" can fully capture the nuances of any dataset. There are inevitably limits to the way that "bias" has been conceptualized here; we do not intend for this work to contribute to a narrative that a "perfectly unbiased" dataset may exist in the real world. While our goal has been to provide context and intuition for the limits of fair algorithms, we also hope this does not provide a false sense of confidence in particular combinations of algorithm/metric/shift.

References

[1] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

[2] A. Blum and K. Stangl. Recovering from biased data: Can fairness constraints improve accuracy? In *1st Symposium on Foundations of Responsible Computing (FORC 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.

[3] S. Corbett-Davies and S. Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[4] A. Coston, K. N. Ramamurthy, D. Wei, K. R. Varshney, S. Speakman, Z. Mustahsan, and S. Chakraborty. Fair transfer learning with missing protected attributes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 91–98, 2019.

[5] A. D'Amour, H. Srinivasan, J. Atwood, P. Baljekar, D. Sculley, and Y. Halpern. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 525–534, 2020.

[6] S. Dutta, D. Wei, H. Yueksel, P.-Y. Chen, S. Liu, and K. R. Varshney. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. 2020.

[7] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[8] B. Fish, J. Kun, and Á. D. Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 144–152. SIAM, 2016.

[9] R. Fogliato, M. G'Sell, and A. Chouldechova. Fairness evaluation in presence of biased noisy labels.

[10] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338, 2019.

[11] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[12] T. J. T. Heiser, M.-L. Allikivi, and M. Kull. Shift happens: Adjusting classifiers. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 55–70. Springer, 2019.

[13] H. Jiang and O. Nachum. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 702–712, 2020.

[14] N. Kallus and A. Zhou. Residual unfairness in fair machine learning from prejudiced data. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[15] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[16] F. Kamiran, A. Karim, and X. Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pages 924–929. IEEE, 2012.

[17] Z. Lipton, Y.-X. Wang, and A. Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, pages 3122–3130, 2018.

[18] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt. Delayed impact of fair machine learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 6196–6200. AAAI Press, 2019.

[19] A. K. Menon and R. C. Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pages 107–118, 2018.

[20] J. G. Moreno-Torres, T. Raeder, R. Alaiz-RodríGuez, N. V. Chawla, and F. Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012.

[21] C. J. Najdowski and K. M. Bernstein. Race, social class, and child abuse: Content and strength of medical professionals' stereotypes. *Child abuse & neglect*, 86:217–222, 2018.

[22] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689, 2017.

[23] S. Rabanser, S. Günnemann, and Z. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In *Advances in Neural Information Processing Systems*, pages 1396–1408, 2019.

[24] C. Schumann, X. Wang, A. Beutel, J. Chen, H. Qian, and E. H. Chi. Transfer of machine learning fairness across domains.

[25] H. Singh, R. Singh, V. Mhasawade, and R. Chunara. Fair predictors under distribution shift. In *NeurIPS Workshop on Fair ML for Health*, 2019.

[26] D. Slack, S. A. Friedler, and E. Givental. Fairness warnings and fair-maml: learning fairly with minimal data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 200–209, 2020.

[27] M. Wick, J.-B. Tristan, et al. Unlocking fairness: a trade-off revisited. In *Advances in Neural Information Processing Systems*, pages 8783–8792, 2019.

[28] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

[29] I. Zliobaite. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*, 2015.