When Fair Classification Meets Noisy Protected Attributes

Avijit Ghosh
Northeastern University
Boston, USA
ghosh.a@northeastern.edu

Pablo Kvitca
Northeastern University
Boston, USA
kvitca.p@northeastern.edu

Christo Wilson
Northeastern University
Boston, USA
cbw@ccs.neu.edu

ABSTRACT

The operationalization of algorithmic fairness comes with several practical challenges, not the least of which is the availability or reliability of protected attributes in datasets. In real-world contexts, practical and legal impediments may prevent the collection and use of demographic data, making it difficult to ensure algorithmic fairness. While initial fairness algorithms did not consider these limitations, recent proposals aim to achieve algorithmic fairness in classification by incorporating noisiness in protected attributes or not using protected attributes at all.

To the best of our knowledge, this is the first head-to-head study of fair classification algorithms to compare attribute-reliant, noise-tolerant and attribute-unaware algorithms along the dual axes of predictivity and fairness. We evaluated these algorithms via case studies on four real-world datasets and synthetic perturbations. Our study reveals that attribute-unaware and noise-tolerant fair classifiers can potentially achieve similar level of performance as attribute-reliant algorithms, even when protected attributes are noisy. However, implementing them in practice requires careful nuance. Our study provides insights into the practical implications of using fair classification algorithms in scenarios where protected attributes are noisy or partially available.

CCS CONCEPTS

• **Social and professional topics** → **User characteristics**; • **General and reference** → **Surveys and overviews**; • **Computing methodologies** → **Machine learning algorithms**.

KEYWORDS

fairness; classification; protected attributes; evaluation

ACM Reference Format:
Avijit Ghosh, Pablo Kvitca, and Christo Wilson. 2023. When Fair Classification Meets Noisy Protected Attributes. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '23), August 8–10, 2023, Montréal, QC, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3600211.3604707

1 INTRODUCTION

In October 2022, the White House released the Blueprint for an AI Bill of Rights [56]. This document, like other statements of AI principles [21, 30, 47, 49, 57], calls for protections against unfair discrimination (colloquially, *fairness*) to be deeply integrated into all AI systems. Researchers and journalists have led the way in this area, both in terms of identifying unfairness in real world systems [6, 11, 14, 44], and in the development of machine learning (ML) classifiers that jointly optimize for predictive performance and fairness [18, 26, 34, 37] (for a variety of different definitions of fairness [4, 27, 58, 63]).

Despite the widespread acknowledgment that fairness is a key component of trustworthy AI, formidable challenges remain to the adoption of fair classifiers in real world scenarios—chief among them being questions about demographic data itself. Many *classical fair classifiers* assume that protected attributes are available at training time and/or testing time [18] and that this data is accurate. However, demographic data may be noisy for a variety of reasons, including imprecision in human-generated labels [15], reliance on imperfect demographic-inference algorithms to generate protected attributes [23], or the presence of an adversary that is intentionally poisoning demographic data [24]. To attempt to deal with these issues, researchers have proposed *noise-tolerant fair classifiers* that aim to achieve distributional fairness by incorporating the error rate of demographic attributes in the fair classifier optimization process itself [13, 48, 60].

In other instances demographic data may not be available at all, which violates the assumptions of both classical and noise-tolerant fair classifiers. This may occur when demographic data is unobtainable (e.g., laws or social norms impede collection [5, 10]), prohibitively expensive to generate (e.g., when large datasets are scraped from the web [16, 35, 41]), or when laws disallow the use of protected attributes to train classifiers (e.g., direct discrimination [62]). For cases such as these, researchers have proposed *demographic-unaware fair classifiers* that use the latent representations in the feature space of the training data to reduce gaps in classification errors between protected groups, either via assigning higher weights to groups of training examples that are misclassified [28], or by training an auxiliary adversarial model to computationally identify regions of misclassification [39].

Motivated by this explosion of fundamentally different fair classifiers, we present an empirical, head-to-head evaluation of the performance of 14 classifiers in this study, spread across four classes: two *unconstrained classifiers*, seven classical fair classifiers, three noise-tolerant fair classifiers, and two demographic-unaware classifiers. Drawing on the methodological approach used by Friedler et al. [22] in their comparative study of classical fair classifiers, we evaluate the accuracy, stability, and fairness guarantees (defined as the equal odds difference) of these 14 classifiers across four datasets as we vary noise in the protected attribute (sex). To help explain the performance differences that we observe, we calculate and compare the feature importance vectors for our various trained classifiers. This methodological approach

AIES '23, August 8–10, 2023, Montréal, QC, Canada

ACM ISBN 979-8-4007-0231-0/23/08…$15.00
https://doi.org/10.1145/3600211.3604707

enables us to compare the performance of these 14 algorithms under controlled, naturalistic circumstances in an apples-to-apples manner.

Based on our head-to-head evaluation we make the following key observations:

- Two classical fair classifiers, one noise-tolerant fair classifier, and one demographic-unaware fair classifier performed consistently well across all metrics on our experiments.
- The best classifier for each case study showed some variability, confirming that the choice of dataset is an important factor when selecting a model.
- One demographic-unaware fair classifier was able to achieve equal odds for males and females under a variety of ecological conditions, confirming that demographics are not always necessary at training or testing time to achieve fairness.

We release our source code and data[1] so that others can replicate and expand upon our results.

We argue that large-scale, head-to-head evaluations such as the one we conduct in this study are critical for researchers and ML practitioners. Our results act as a checkpoint, informing the community about the relative performance characteristics of classifiers within and between classes. For researchers, this can highlight gaps where novel algorithms are still needed (e.g., noise-tolerant and demographic-unaware classifiers, based on our findings) and provide a framework for rigorously evaluating them. For practitioners, our results highlight the importance of thoroughly evaluating many classifiers from many classes before adopting one in practice, and we provide a roadmap for choosing the best classifiers for a given real-world scenario, depending on the availability and quality of demographic data.

Our study proceeds as follows: in § 2 we present a brief overview of the history of fair models and head-to-head performance evaluation. Next, in § 3, we introduce the 14 classifiers and the metrics we use to evaluate them for predictive performance and fairness. In § 4 we present our experimental approach, including the datasets we use for our four case studies. In § 5 we present the results of our experiments and we discuss our findings in § 6.

2 RELATED WORK

We discuss different classes of fair classifiers, their known shortcomings, and how they have been evaluated in the past.

2.1 Fair Classifiers

Dwork et al. [18] were one of the first to operationalize the idea of fairness in machine learning classifiers, through their key observation that awareness of demographics is crucial for building models that rectify unfair discrimination and historical inequity. Their work takes the idea of awareness literally, by incorporating protected attributes directly into the model and jointly optimizing for accuracy and fairness. Many subsequent works have built on this foundation by developing versions of classical ML classifiers that incorporate fairness constraints (e.g., decision trees, random forests, SVMs, boosting, etc. [46]).

[1]The code and data for replicating this paper can be found at https://github.com/evijit/Awareness_vs_Unawareness

Collectively, we refer to this class of algorithms as classical fair classifiers. They are now widely available to practitioners [9, 42, 52] and have been adopted into real-world systems [19].

While classical fair classifiers are an important advance over their unconstrained predecessors, they rely on a strong assumption that data about protected attributes is accurate. Unfortunately, this may not be true in practice. For example, in contexts like finance and employment candidate screening, demographic data may not be available due to legal constraints or social norms [10, 62], yet the need to fairly classify people remains paramount. To bridge this gap, practitioners may infer peoples' protected attributes using human labelers [8] or algorithms that take names, locations, photos, etc. as input [1]. However, work by Ghosh et al. [23] demonstrates that these inference approaches produce noisy demographic data, and that this noise obviates the fairness guarantees provided by fair models.

With these limitations in mind, researchers have begun developing what we refer to as noise-tolerant fair classifiers that, as the name suggests, jointly optimize for accuracy and fairness in the presence of uncertainty in the protected attribute data. Approaches include robust optimization that adjusts for the presence of noise in the fairness constraint [60], adjusting the "fairness tolerance" value for binary protected groups [40], using noisy attributes to post-process the outputs for fairness instead of the true attributes under certain conditional independence assumptions [7], estimating de-noised constraints that allow for near optimal fairness [13], or a combination of approaches [48].

Noise-tolerant fair classifiers, like classical fair classifiers, still rely on the assumption that protected attributes are available at training time. As we discuss in § 1, however, there are many real-world contexts when this assumption may be violated. The strongest such impediment is legal, i.e., any inclusion of protected attributes in the classifier would be considered illegal direct discrimination.

A different approach for achieving fairness through awareness that is amenable to these strong constraints is embodied by what we refer to as demographic-unaware fair classifiers. These algorithms do not take protected attributes as input, but they attempt to achieve demographic fairness anyway by relying on the latent representations of the training data [28, 39]. Thus, this approach to classification still incorporates a general awareness of unfair discrimination and historical inequity without being directly aware of demographics.

While demographic-unaware fair classifiers are an attractive solution in contexts where protected attributes are unavailable, practical questions about the efficacy of these algorithms remain. First, because these techniques are unsupervised, it is unclear what groups are identified for fairness optimization. Under what circumstances are demographic-unaware fair classifiers able to achieve fairness for social groups that have been historically marginalized or are legally protected? Conversely, are the groups constructed by demographic-unaware fair classifiers arbitrary and thus divorced from salient real-world sociohistorical context? Second, assuming that demographic-unaware fair classifiers do identify and act on meaningful groups of individuals, how does their performance (in terms of predictions and fairness) compare to classical and noise-tolerant fair classifiers? In this study, our goal

is to begin answering these questions about relative performance across all four classes of fair classifiers.

2.2 Head-to-Head Evaluation

It is standard practice for ML researchers to compare the performance of their novel algorithms against competitors. However, these comparisons are rarely comprehensive, i.e., they focus on comparisons with a narrow set of comparable algorithms to demonstrate advances over the state-of-the-art. While these evaluations are crucial for assessing the benefits of new algorithms, they do not paint a complete picture of performance across a variety of different algorithms, spanning both time and fundamental approaches.

Benchmark studies address this gap by focusing on the evaluation of a large set of models under expansive and carefully controlled conditions [22, 29]. These studies provide important context for the ML field, e.g., by identifying models that do not work well in practice, models that have equivalent performance characteristics under a wide range of circumstances, and areas where new models may be needed. To the best of our knowledge, existing benchmark studies focus solely on classical fair classifiers, which motivates us to update their results. Thus, in this study we adopt the methodological approach for evaluation developed by Friedler et al. [22] and build upon their work by evaluating four different classes of classifiers (both fairness constrained and unconstrained).

3 ALGORITHMS AND METRICS

In this section, we introduce the 14 classifiers that we evaluated in this study and the metrics we used to evaluate them.

3.1 Classifiers

We group the classifiers that we evaluated in this study into four classes: (1) unconstrained classifiers that solely optimize for accuracy; (2) classical fair classifiers that require access to protected attributes at training (and sometimes testing) time, and assume that this data are accurate; (3) noise-tolerant fair classifiers that also require access to protected attributes but account for uncertainty in the data; and (4) demographic-unaware fair classifiers that jointly optimize for accuracy and fairness but without access to any protected attribute data. The set of classifiers we have selected is not exhaustive. Instead, we aim to include representative classifiers from the various types of approaches that exist within each class. We discuss the classifiers from each class that we selected for our study below, with further details on related approaches in each subsection.

3.1.1 Unconstrained Classifiers. We chose two classifiers that do not have any fairness constraints, i.e., they only aim to maximize predictive accuracy.

- **Logistic Regression (LR)** is the simplest classifier we evaluate. While LR is demographic-aware because it takes all features (including protected attributes) as model inputs at both train and test time, it is not designed to achieve any fairness criteria.
- **Random Forest (RF)** is an ensemble method for classification built out of decision trees. Like LR, we train RF classifiers on all input features including protected attributes.

3.1.2 Classical Fair Classifiers. We chose seven classifiers from the literature that take protected attributes as input and attempt to achieve demographic fairness. These classifiers vary with respect to how they implement fairness, i.e., by pre-processing data, in-process during model training, or by post-processing the trained model. In particular, there exist many techniques for fairness optimization in this class, such as: reweighting of samples via group sizes [12, 20, 32] or via mutual independence of protected and unprotected features in the latent representations [64, 65], adding fairness constraints during the learning process [2, 3, 34, 63], or by changing the output labels to match some fairness criterion [33, 50]. The seven classifiers we choose below are representative of these different approaches.

- **Sample Reweighting (SREW)** is a pre-processing technique that takes each (group, label) combination in the training data and assigns rebalanced weights to them. The goal of this procedure is to remove imbalances in the training data, with the ultimate aim of ensuring fairness before the classifier is trained [32].
- **Learned Fair Representation (LFR)** is a pre-processing technique that converts the input features into a latent encoding that is designed to represent the training data well while simultaneously hiding protected attribute information from the classifier [64].
- **Adversarial Debiasing (ADDEB)** is an in-process technique that trains a classifier to maximize accuracy while simultaneously reducing an adversarial network's ability to determine the protected attributes from the predictions [65].
- **Exponentiated Gradient Reduction (EGR)** is an in-process technique that reduces fair classification to a set of cost-sensitive classification problems, essentially treating the main classifier itself as a black box and forcing the predictions to be the most accurate under a given fairness constraint [2]. In this case, the constraint is solved as a saddle point problem using the exponentiated gradient algorithm.
- **Grid Search Reduction (GSR)** uses the same set of cost-sensitive classification problems approach as EGR, except in this case the constraints are solved using the grid search algorithm [2, 3].
- **Calibrated Equalized Odds (CALEQ)** is a post-processing technique that optimizes the calibrated classifier score output to find the probabilities that it uses to change the output labels, with an equalized odds objective [50].
- **Reject Option Classifier (ROC)** is a post-processing technique that swaps favorable and unfavorable outcomes for privileged and unprivileged groups around the decision boundaries with the highest uncertainty [33].

Note that the CALEQ and ROC algorithms have access to protected attributes at both train and test time, while the other classifiers only have access to protected attributes at training time.

3.1.3 Noise-tolerant Fair Classifiers. We chose three classifiers from the literature that take protected attributes as input and attempt to achieve demographic fairness even in the presence of

noise. Other than the three classifiers that we chose, we are aware of only one other approach: by Celis et al. [13], who suggests using de-noised constraints to achieve near-optimal fairness.[2]

- **Modified Distributionally Robust Optimization (MDRO)** by Wang et al. [60] is an extension of the Distributionally Robust Optimization (DRO) algorithm [28] that adds a maximum total variation distance in the DRO procedure. By assuming a noise model for the protected attributes, it aims to provide tighter bounds for DRO.
- **Soft Group Assignments (SOFT)**, also by Wang et al. [60], is a theoretically robust approach that first performs "soft" group assignments and then performs classification, with the idea being that if an algorithm is fair in terms of those robust criteria for noisy groups, then they must also be fair for true protected groups [31].
- **Private Learning (PRIV)** is an approach by Mozannar et al. [48] that uses differential privacy techniques to learn a fair classifier while having partial access to protected attributes. The approach requires two steps. The first step is to obtain locally private versions of the protected attributes (like Lamy et al. [40]). Second, following Awasthi et al. [7], PRIV tries to create a fair classifier based on the private attributes. For this study, we select the privacy level hyperparameter to be a medium value (zero).

3.1.4 Demographic-unaware Fair Classifiers. We chose two classifiers from the literature that attempt to achieve fairness without taking protected attributes as input.

- **Adversarially Reweighted Learning (ARL)** harnesses non-protected attributes and labels by utilizing the computational separability of these training instances to divide them into subgroups, and then uses an adversarial reweighting approach on the subgroups to improve classification fairness [39].
- **Distributionally Robust Optimization (DRO)** is an algorithm that attempts to minimize the worst case risk of all groups that are close to the empirical distribution [28]. In the spirit of Rawlsian distributive justice, the algorithm tries to control the risk to minority groups while being oblivious to their identities.

These two classifiers operate under similar principles: they both try to reduce the gap in errors between protected groups by reducing the classification errors between latent groups in the training set. They do however have one difference: while DRO just increases the weights of the training examples that have higher errors, ARL trains an auxillary adversarial network to identify the regions in the latent input space that lead to higher errors and tries to equalize them, a phenomenon Lahoti et al. [39] call *computational identifiability*.

3.2 Evaluation Metrics

To compare the above 14 classifiers head-to-head, we studied their predictive power and their ability to achieve a fairness condition. We also measured the stability of these quantities when noise in the protected attributes was and was not present (described in § 4.2).

To assess predictive performance we computed accuracy, defined as:

$$\text{Accuracy} = \frac{\text{number of correct classifications}}{\text{test dataset size}}. \quad (1)$$

Accuracy is continuous between zero and one with the ideal value being one, which indicates a perfectly predictive classifier.

Many measures of fairness exist in the literature [46]. For the purposes of this study, however, we needed to choose a metric that is supported by all the 14 classifiers so that our comparison is apples-to-apples. The classical and noise-tolerant fair classifiers have support for achieving any user-specified fairness constraint, while the demographic-unaware fair classifiers try to minimize the gap in utility between the protected groups. Based on this limitation, and for the sake of brevity, we choose the Average Odds Difference between two demographic groups as our fairness metric, and subsequently choose Equal Odds Difference (EOD) over both groups as our regularization constraint for the classical and noise-tolerant fair classifiers. EOD is defined as:

$$\text{EOD} = \frac{(\text{FPR}_{\text{unpriv}} - \text{FPR}_{\text{priv}}) + (\text{TPR}_{\text{unpriv}} - \text{TPR}_{\text{priv}})}{2} \quad (2)$$

where TPR is the true positive rate and FPR is the false positive rate. Priv and Unpriv denote the privileged and unprivileged groups, respectively. The ideal value of EOD is zero, which indicates that both groups have equal odds of correct and incorrect classification by the trained classifier.

In this study, when we evaluate fairness, we do so for binary sex attributes. We adopted this approach because the datasets we use in our evaluation all include this attribute (see § 4) and four classifiers in our evaluation (e.g., CALEQ, ROC, EGR, GSR) only support fairness constraints over two groups. Whenever necessary, we consider males to be the privileged group and females to be the unprivileged group. Note that optimizing for fairness between two groups is the simplest scenario that fair classifiers will encounter in practice—if they perform poorly on this task, then they are unlikely to succeed in more complex scenarios with multiple, possibly intersectional, groups.

4 METHODOLOGY

In this section, we describe the approach we used to empirically evaluate the 14 classifiers that we chose for our study.

4.1 Case Studies

To observe how the classifiers perform on real-world data we chose four different datasets. The classification tasks are described below. Each dataset had binary sex as part of the input features.

(1) **Public Coverage [17].** The task is to predict whether an individual (who is low income and not eligible for Medicare) was covered under public health insurance. We used census data from California for the year 2018.
(2) **Employment [17].** The task is to predict whether an individual (between the ages of 16 and 90), is employed. For this task too, we looked at census data from California for the year 2018.
(3) **Law School Admissions [61].** The task is to predict whether a student was admitted to law school.

[2] Celis et al. [13]'s source code only supported Statistical Parity and False Discovery constraints, not EOD, which is why we omitted their classifier from our analysis.

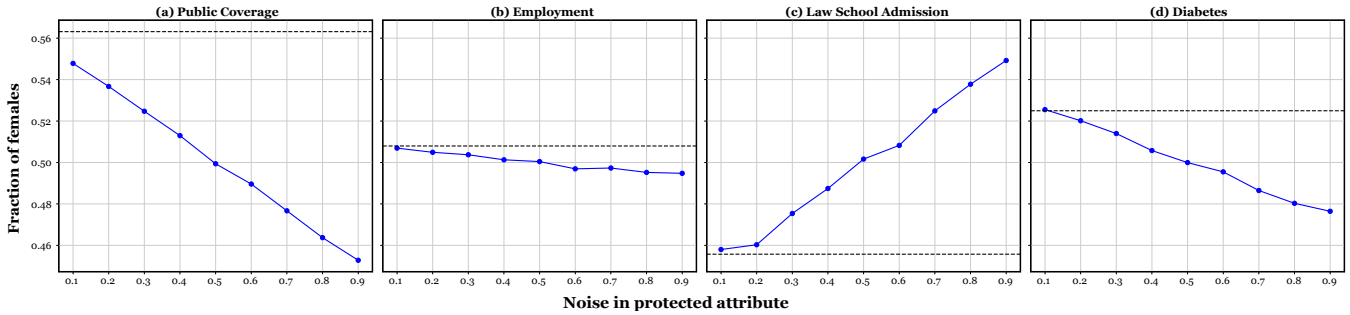


Figure 1: Fraction of females in our datasets after adding synthetic noise. The dashed line indicates the true fraction of females.

(4) **Diabetes [54].** The task is to predict whether a diabetes patient was readmitted to the hospital for treatment after 30 days.

For each of these case studies, we split the dataset into train and test sets in an 80:20 ratio, trained every classifier on the same training set, and then used the trained classifiers to generate predictions on the same testing set. We verified via two-tailed Kolmogorov–Smirnov tests [36, 53] and Mann–Whitney U tests [45] that the test set distribution for every feature was the same as the training set distribution. Finally, we calculated the metrics in § 3.2 on these predictions and compared the results from each classifier head-to-head. We repeated this procedure ten times to assess the stability of accuracy and EOD for each classifier.

4.2 Synthetic Noise

While studying the performance of these classifiers on a variety of real-world datasets is important, in order to get a more thorough understanding of the theoretical fairness and predictivity limits of the classifiers we subjected them to robust synthetic stress tests. As discussed in § 2.1, in the real world, practitioners may not have access to the protected attribute information of people in their dataset. As a result, practitioners may use inference tools to find proxies for protected attributes, which can lead to unexpected, unfair outcomes [23]. To characterize what might happen in such a scenario, we perform the following synthetic experiments:

(1) For each dataset, with a given probability (ranging from 0.1 to 0.9), we randomly flip the protected attribute labels (binary sex in this case) in the dataset. We refer to this probability value as *noise*.
(2) With the synthetically generated dataset from Step 1, we then proceed to split the dataset 80:20, train all 14 algorithms on the same training set, and then calculate predictions on the same test set. The noisy (flipped) labels are passed as inputs to the classifiers at this step.
(3) Next, with the predicted outcomes from Step 2, we calculate accuracy and EOD. Note that we calculate EOD with the *true* protected attributes, i.e., we measure the output bias in terms of the original sex labels from the given dataset.
(4) We repeat Steps 1–3 ten times for each value of noise, to ensure statistical fairness and assess the stability of our metrics per classifier.

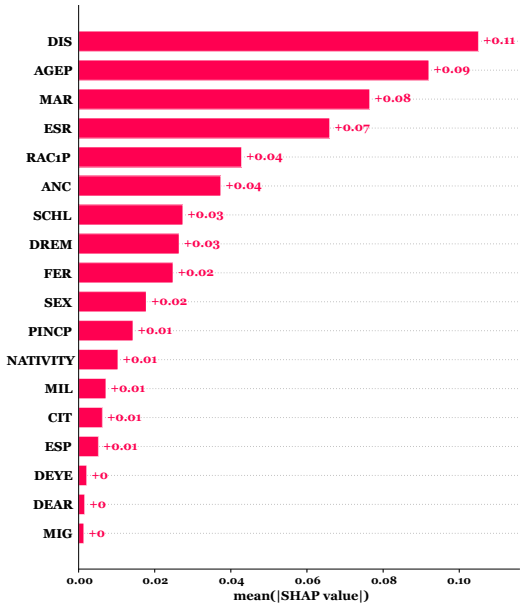


Figure 2: KernelShap feature explanations calculated for the Logistic Regression (LR) classifier when trained on the Public Coverage dataset with no added noise. We used the same approach to calculate feature importances for every classifier-dataset pair at different noise levels.

Figure 1 shows the fraction of females in the noised datasets at each level of noise. The fraction of females goes up or down with noise depending on what the true fraction of females in the different datasets were to begin with.

4.3 Calculating Feature Importance

To help explain the variations in performance that we observed in our results, we calculated feature importance for each of our trained models. Although there are several black-box model explanation tools in the research literature—such as LIME [51], SHAP [43], and Integrated Gradients [55]—we required an explanation method that was model agnostic. The method that we settled on was KernelShap.[3] According to the documentation, KernelShap uses

[3] https://shap-lrjball.readthedocs.io/en/latest/generated/shap.KernelExplainer.html

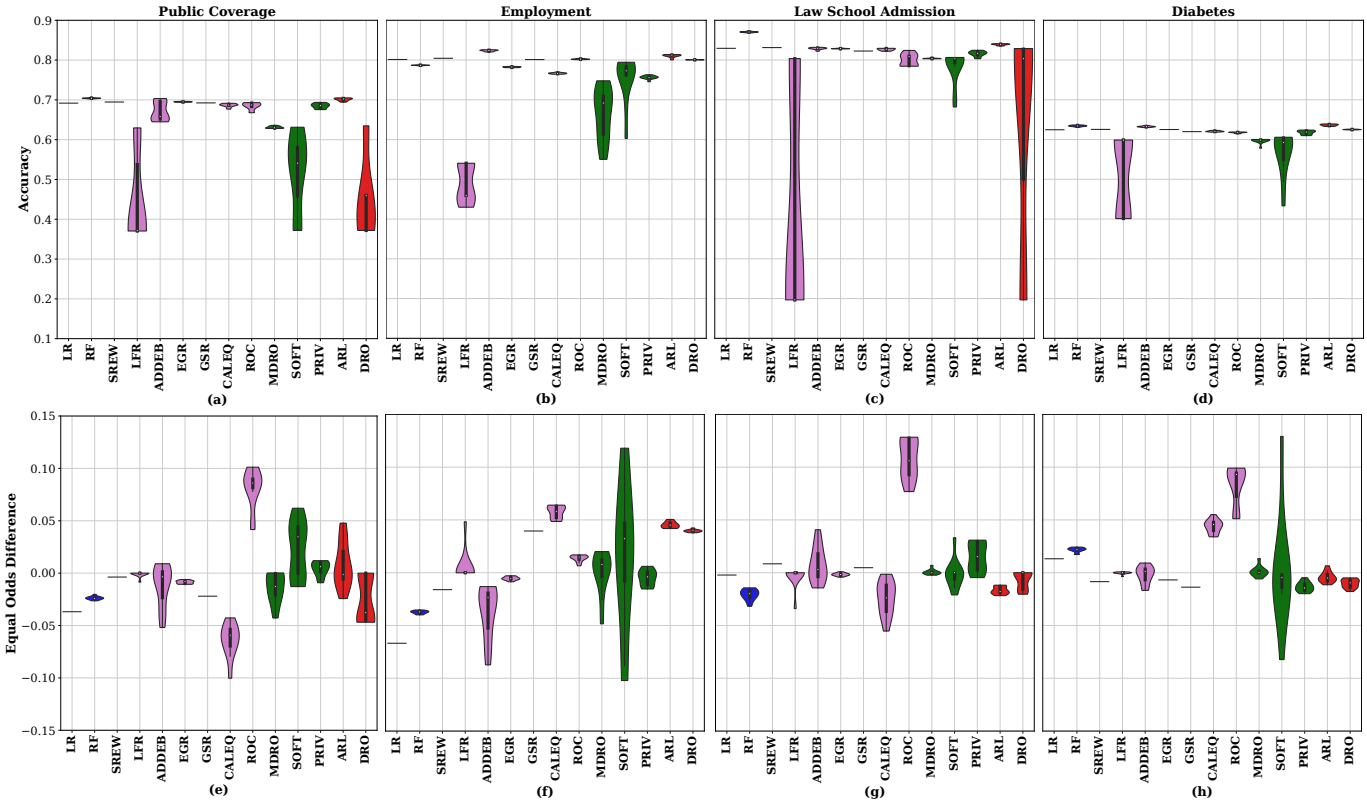


Figure 3: Accuracy and EOD for our 14 classifiers, calculated over four datasets with ten runs each. No noise was added to the protected attribute in these tests. Violins are color coded by class: blue for unconstrained classifiers, purple for classical fair classifiers, green for noise-tolerant fair classifiers, and red for demographic-unaware fair classifiers. LR, SREW, and GSR are deterministic algorithms and therefore appear as fixed points.

a special weighted linear regression model to calculate local coefficients, to estimate the Shapley value (a game theoretic concept that estimates the individual contribution of each player towards the final outcome). As opposed to retraining the model with every combination of features as in vanilla SHAP, KernelShap uses the full model and integrates out different features one by one. It also supports any type of model, not just linear models, and was thus a good candidate for our study.

Figure 2 shows an example distribution of feature importances calculated for the LR algorithm when trained on the Public Coverage dataset at noise level zero (i.e., no noise). In a similar fashion, we used KernelShap to calculate feature importance values for trained classifier outputs at noise levels 0, 0.2, 0.4, 0.6 and 0.8 for all 14 models.

Research by Kumar et al. [38] has shown that different explanation methods often do not agree with each other. We do not claim that the feature importances we calculated using KernelShap are guaranteed to agree with those produced by other tools. Nonetheless, we are specifically interested in the relative importance of the sex feature towards the final outcome as compared to the other input features. Shapley value-based explanations give us a reasonable sense of relative feature importance, as has been empirically shown in previous work [25].

5 RESULTS

In this section, we present the results of our experiments. We begin by examining the baseline performance of the 14 classifiers when there is no noise, followed by their performance in the presence of synthetic noise. Finally, we delve into feature importance explanations to help explain the relative performance characteristics of the classifiers.

5.1 Baseline Characteristics

Figure 3(a–d) shows the accuracy and fairness outcomes for all 14 classifiers when there was no noise in the datasets. We executed each classifier ten times without fixing a random seed and present the resulting distributions of metrics using violin plots. We observe that most of the classifiers achieved comparable accuracy to each other on each dataset, and that most classifiers exhibited stable accuracy over the ten executions of the experiments. Learned Fair Representation (LFR), Soft Group Assignment (SOFT), and Distributed Robust Optimization (DRO) were the exceptions: the former two exhibited unstable accuracy on all four datasets, the latter on two datasets.

As shown in Figure 3(e–h), EOD was considerably more variable over runs than accuracy. The unconstrained classifiers (LR and RF) were relatively stable and, in some cases, achieved roughly

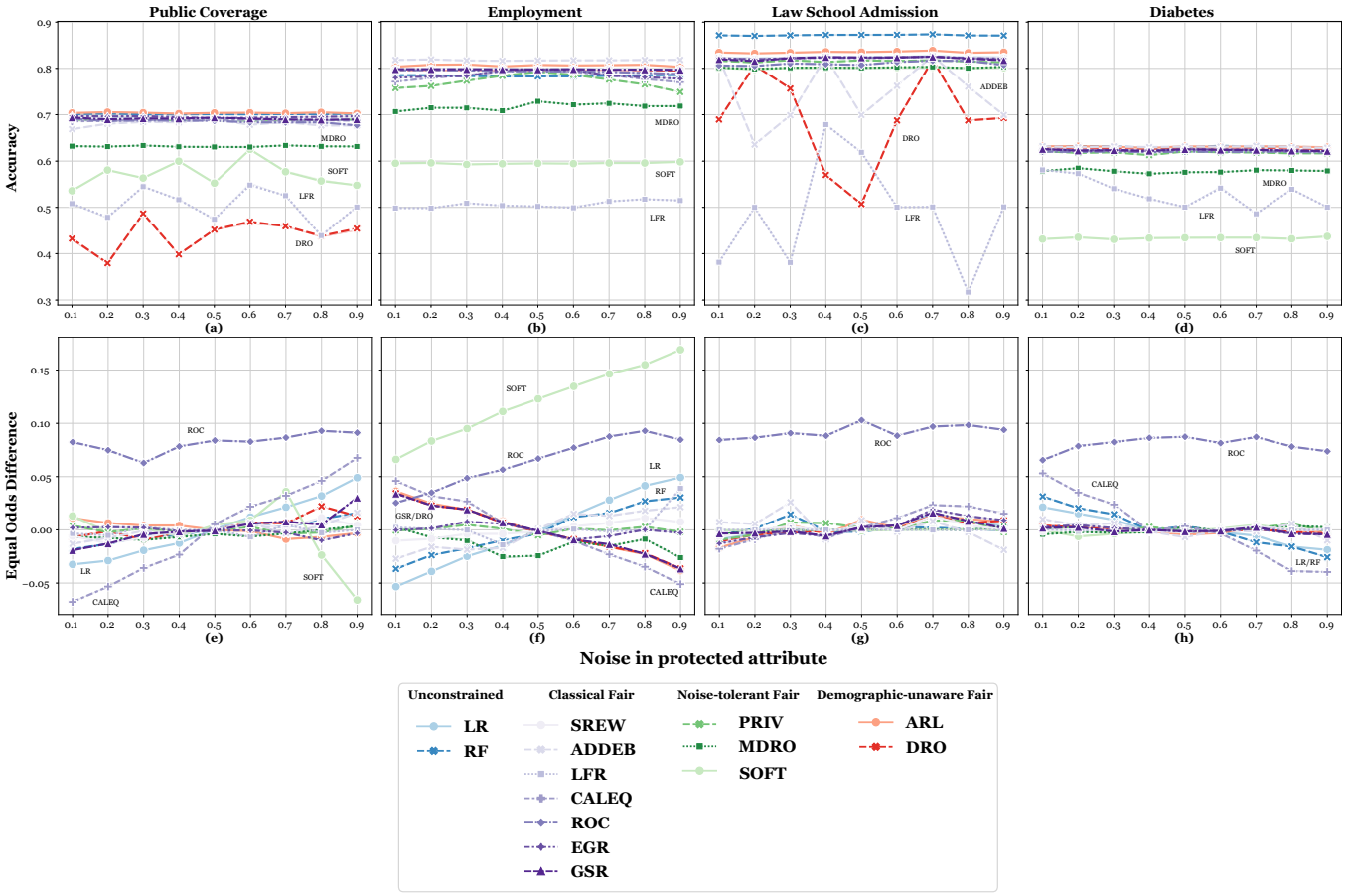


Figure 4: Accuracy and EOD for our 14 classifiers, calculated over four datasets as we increase noise in the protected attribute (sex). Each point is the average of ten runs for a given classifier, dataset, and noise level. Classifiers are color coded according to the legend. We highlight classifiers whose performance significantly diverges from the consensus with annotated labels.

equalized odds (e.g., on the Law School and Diabetes datasets). The classical fair classifier group contained the two least fair classifiers in these experiments (CALEQ and ROC), while the other pre-processing and in-processing algorithms performed relatively better. Adversarial Debiasing (ADDEB) was slightly unstable but the distribution centered around zero. Among the noise-tolerant fair classifiers, Soft Group Assignment (SOFT) was unstable on three out of four datasets, while the other two classifiers (MDRO and PRIV) were relatively more stable and more fair. The two demographic-unaware fair classifiers (ARL and DRO) were unstable on the Public Coverage dataset (Figure 3e) and did not achieve equalized odds on the Employment dataset (Figure 3f). However, ARL and DRO were stable and fair on the remaining two datasets.

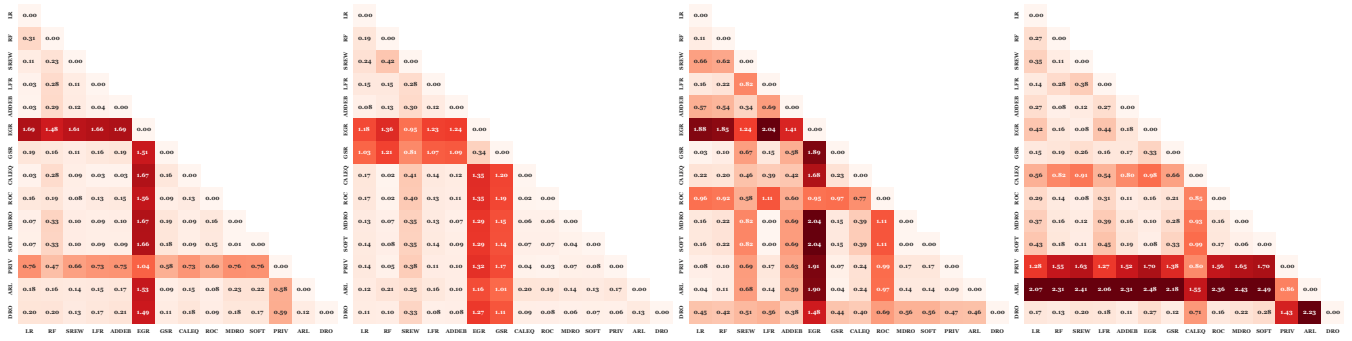
In summary, we observe that the accuracy and fairness performance of these classifiers was dependent on the dataset that they are trained and tested on, i.e., there was no single best classifier. Additionally, we can see that several classifiers are consistently unstable, which explains some of the results that we will present in the next section.

5.2 Characteristics Under Noise

Next, we present the results of experiments where we added noise to the protected attribute of the datasets. We added noise in increments of 0.1 starting from 0.1 and ranging up to 0.9. We added a given amount of noise to each dataset ten times and repeated the experiment, thus we plot the average values of accuracy and EOD for each classifier at each noise level.

Figure 4(a–d) shows the accuracy of the 14 classifiers' outputs as we varied noise. We observe that the MDRO, SOFT, and LFR classifiers had poor accuracy across all datasets and noise levels, while the DRO classifier had poor accuracy in two out of the four datasets. These observations mirror those from Figure 3, i.e., these classifiers exhibited poor average accuracy in the noisy experiments because they were unstable in general. The other classifiers tended to be both accurate and stable, irrespective of noise.

As shown in Figure 4(e–h), the EOD results were much more complex than the accuracy results. ROC generated unfair outputs over all four datasets, at every noise level. Its companion post processing algorithm, CALEQ, exhibited rising EOD with noise for the Public Coverage dataset (Figure 4e) and falling



(a) Public Coverage (b) Employment (c) Law School Admissions (d) Diabetes

Figure 5: Wasserstein distances between the average KernelShap feature importance distributions over different noise levels for the four datasets. Each square compares the average feature importances of two classifiers. Redder squares denote pairs of classifiers with more divergent feature importance distributions.

EOD for the Employment and Diabetes datasets (Figure 4f, h).[4] The unconstrained classifiers (LR and RF) moved in the same direction for every dataset, either rising (Figure 4e, f) or falling (Figure 4h) with noise. The SOFT classifier also exhibited some variable behavior: on the Employment dataset EOD rose with noise (Figure 4f), and on the Public Coverage (Figure 4e) dataset it failed to achieve equal odds at higher noise levels. The remaining classifiers tended to achieve equal odds irrespective of the noise level.

Figure 4 only depicts average values for accuracy and EOD, which is potentially problematic because it may hide instability in the classifiers' performance. To address this we present Figure 7 in the Supplementary Material, which shows the distribution of accuracy and EOD results for each classifier on each dataset at the 0.1, 0.5, and 0.9 noise levels. We observe that, overall, no classifier became consistently less stable as noise increased. Rather, the stability patterns for each classifier mirrored the patterns that we already observed in Figure 3.

In summary, the classifiers that had problematic performance in the baseline experiments (see Figure 3) continued to have issues in the presence of noise. Additionally, the unconstrained classifiers exhibited inconsistent fairness as noise varied. Surprisingly, the noise-tolerant classifiers did not uniformly outperform the other fair classifiers.

5.3 Feature Importance

Finally, we delve into model explanations as a means to further explore the root causes of the classifier performance characteristics that we observed in the previous sections. First, we calculated feature explanations using KernelShap for every classifier at five noise levels—0, 0.2, 0.4, 0.6 and 0.8—using the method we described in § 4.3. Next, we averaged the explanation distributions for each classifier to form a feature importance vector per classifier. Finally, we repeated this process for each dataset. For each dataset, we calculated Wasserstein distances [59] between the feature explanation distributions for each algorithm pair and present the results in Figure 5. Additionally, we plot the rank of the sex feature

[4]Note that a higher value of EOD (Equation 3.2) signifies that females received more positive predictions than males.

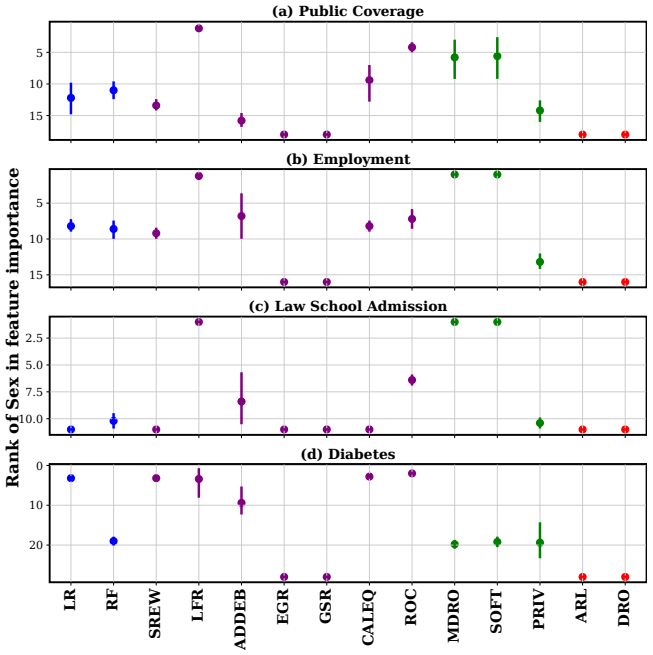


Figure 6: Rank of Sex in the average absolute KernelShap feature importances for the different algorithms in our case studies.

in terms of mean absolute feature importance for each classifier and present the results in Figure 6 (we also show the range of ranks if they vary over noise).

Figure 5 reveals that, with few exceptions (EGR in Public Coverage, EGR and GSR in Employment, EGR and ROC in Law school, and CALEQ, PRIV and ARL in Diabetes), most classifiers had similar feature explanation distributions. We do not observe any clear patterns among the exceptional classifiers, i.e., no classifier consistently diverged from the others across all datasets. Further, we do not observe clear correlations between accuracy, EOD, and feature distribution similarity, suggesting that different classifiers took different paths to reach the same levels of performance.

Figure 6 is more informative than Figure 5. Four of the classifiers that exhibited consistently poor performance—LFR, MDRO, and SOFT (Figure 3a–d), and ROC (Figure 3e–h)—learned to weight the sex feature higher than other features, which may point to the root cause of their accuracy and fairness issues. Similarly, the unconstrained classifiers (LR and RF) exhibited changing EOD with noise levels in three out of four datasets (Figure 4e, f, h), but not for Law School Admissions (Figure 4g), and we observe that they learned a relatively low weight for sex among the available features for the Law School dataset. CALEQ also learned a relatively low weight for sex on the Law School dataset and was subsequently unaffected by noise (Figure 4g), but showed variable trends in EOD for the other three datasets (Figure 4e, f, h) on which it learned a relatively higher weight for sex.

Sex was the lowest ranked feature for the two demographic-unaware fair classifiers (DRO and ARL), which makes sense because they were not given these features as input. EGR and GSR also did not have access to sex while classifying the test dataset, so they also had sex as the lowest ranked feature.

5.4 Fairness-Accuracy Tradeoff

Three algorithms in our list - EGR, GSR, and PRIV, provide a mechanism to control the fairness-accuracy tradeoff via a hyperparameter – namely fairness violation *eps* in the case of EGR and GSR [2], and the privacy level ϵ in the case of PRIV [48]. Based on the experiments the authors of these algorithms did in their papers, we used different *eps* values between 0.01 and 0.20 and ϵ values between -2 and 2 and reran our experiments. We found that tweaking the tradeoff hyperparameter did not contribute meaningfully to the stability and noise resistance capabilities of these algorithms. Consequently we omit these results from the paper.

6 CONCLUSION

In this study, we present benchmark results—in terms of accuracy, fairness, and stability—for 14 ML classifiers divided into four classes. We evaluated these classifiers across four datasets and varying levels of random noise in the protected attribute. Overall, we found that two classical fair classifiers (SREW and EGR), one noise-tolerant fair classifier (PRIV), and one demographic-unaware fair classifier (ARL) performed consistently well across metrics on our experiments. In the future we recommend that ML researchers benchmark their own fair classifiers against these classifiers and that practitioners consider adopting them.

One surprising finding of our study was how well SREW and EGR performed in the face of noise in the protected attribute. Contrast this to noise-tolerant classifiers like MDRO—whose performance did not vary with noise but was inaccurate on some datasets—and SOFT—which was consistently inaccurate and had variable fairness in the face of noise. These results suggest that some classical fair classifiers may actually fare well in the face of noise, and that adopting more complex noise-tolerant fair classifiers may not always be necessary.

Another surprising finding of our study was how well ARL performed. As a demographic-unaware fair classifier it did not have access to the sex feature at training or testing time, yet it achieved fairness performance that was comparable to demographic-aware fair classifiers on three of our datasets, and its fairness performance was noise invariant on three datasets as well. We fit linear regression models on each dataset with sex as the independent variable, but these models did not uncover any obvious proxy features for ARL to use in place of the sex feature. This speaks to the strength of the ARL algorithm's adversarial approach to learning.

On one hand, our results confirm that demographic-unaware fair classifiers can achieve fairness for real-world disadvantaged groups under ecological conditions. This is positive news for practitioners who would like to adopt a fair classifier but lack (high-quality) demographic data. On the other hand, we still urge caution with respect to the adoption of demographic-unaware fair classifiers for practical reasons. First, determining whether a classifier like ARL will achieve acceptable performance in a given context requires thorough evaluation on a dataset that includes demographic data, as we have done here. Second, even if a demographic-unaware fair classifier performs well in testing, its performance may degrade after deployment if the context changes or there is distribution drift [25]. Monitoring the health of a classifier like ARL in the field requires demographic data. In short, adopting a demographic-unaware classifier does not completely obviate the need for at least some high-quality demographic data.

In general, the results of our study point to the need for further development in the areas of noise-tolerant and demographic-unaware fair classifiers. By releasing our source code and data, we hope to provide a solid foundation for evaluating these novel classifiers in the future.

Our study has several limitations. First, we only evaluate classifiers using binary protected attributes. It is unclear how their performance and consistency would change under more complex conditions. That said, we are confident that the classifiers that performed poorly will continue to do so in the presence of more complex fairness objectives. Second, our case studies and synthetic experiments, while thorough, are by no means completely representative of all real world datasets and contexts. We caution that our results should not be generalized indefinitely. Third, we did not evaluate all of the classical fair classifiers from the literature (see Friedler et al. [22] and Mehrabi et al. [46] for more). Our primary focus was on adding to the literature by benchmarking noise-tolerant and demographic-unaware fair classifiers. Finally, in this study we only evaluated one fairness metric—EOD—because it was the common denominator among all of the classifiers we selected. Future work could explore fairness performance further by choosing other fairness metrics along with subsets of amenable classifiers.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their helpful comments. We also thank Jeffrey Gleason for notes on the manuscript. This research was supported in part by NSF grant IIS-1910064. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

[1] 2014. Using publicly available information to proxy for unidentified race and ethnicity: A methodology and assessment. Consumer Financial Protection Bureau. https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*. PMLR, 60–69.

[3] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. 2019. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*. PMLR, 120–129.

[4] Sina Aghaei, Mohammad Javad Azizi, and Phebe Vayanos. 2019. Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 1418–1426.

[5] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 249–260.

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

[7] Pranjal Awasthi, Matthäus Kleindessner, and Jamie Morgenstern. 2019. Effectiveness of equalized odds for fair classification under imperfect group information. *arXiv preprint arXiv:1906.03284* (2019).

[8] Sid Basu, Ruthie Berman, Adam Bloomston, John Campbell, Anne Diaz, Nanako Era, Benjamin Evans, Sukhada Palkar, and Skyler Wharton. 2020. Measuring discrepancies in Airbnb guest acceptance rates using anonymized demographic data. AirBNB. https://news.airbnb.com/wp-content/uploads/sites/4/2020/06/Project-Lighthouse-Airbnb-2020-06-12.pdf.

[9] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. https://doi.org/10.48550/ARXIV.1810.01943

[10] Miranda Bogen, Aaron Rieke, and Shazeda Ahmed. 2020. Awareness in practice: tensions in access to sensitive attribute data for antidiscrimination. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 492–500.

[11] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. 77–91.

[12] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems* 30 (2017).

[13] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. 2021. Fair classification with noisy protected attributes: A framework with provable guarantees. In *International Conference on Machine Learning*. PMLR, 1349–1361.

[14] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of Data and Analytics*. Auerbach Publications, 296–299.

[15] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics* 10 (2022), 92–110.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[17] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems* 34 (2021), 6478–6490.

[18] 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.

[19] EY. 2020. Assessing and mitigating unfairness in credit models with Fairlearn. https://www.ey.com/en_ca/financial-services/assessing-and-mitigating-unfairness-in-credit-models. [Accessed: March 16th, 2023].

[20] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.

[21] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication* 2020-1 (2020).

[22] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*. 329–338.

[23] Avijit Ghosh, Ritam Dutt, and Christo Wilson. 2021. When fair ranking meets uncertain inference. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1033–1043.

[24] Avijit Ghosh, Matthew Jagielski, and Christo Wilson. 2022. Subverting Fair Image Search with Generative Adversarial Perturbations. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*. ACM, 637–650. https://doi.org/10.1145/3531146.3533128

[25] Avijit Ghosh, Aalok Shanbhag, and Christo Wilson. 2022. Faircanary: Rapid continuous explainable fairness. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 307–316.

[26] Naman Goel, Mohammad Yaghini, and Boi Faltings. 2018. Non-discriminatory machine learning through convex fairness criteria. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[27] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413* (2016).

[28] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*. PMLR, 1929–1938.

[29] Max Hort, Jie M. Zhang, Federica Sarro, and Mark Harman. 2021. Fairea: A Model Behaviour Mutation Approach to Benchmarking Bias Mitigation Methods. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*.

[30] IBM. 2022. AI Ethics: IBM's multidisciplinary, multidimensional approach to trustworthy AI. https://www.ibm.com/artificial-intelligence/ethics.

[31] Nathan Kallus, Xiaojie Mao, and Angela Zhou. 2022. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science* 68, 3 (2022), 1959–1981.

[32] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems* 33, 1 (2012), 1–33.

[33] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*. IEEE, 924–929.

[34] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 35–50.

[35] Kimmo Kärkkäinen and Jungseock Joo. 2019. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913* (2019).

[36] Andrey Kolmogorov. 1933. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.* 4 (1933), 83–91.

[37] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 world wide web conference*. 853–862.

[38] I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. 2020. Problems with Shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*. PMLR, 5491–5500.

[39] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. 2020. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems* 33 (2020), 728–740.

[40] Alex Lamy, Ziyuan Zhong, Aditya K Menon, and Nakul Verma. 2019. Noise-tolerant fair classification. *Advances in neural information processing systems* 32 (2019).

[41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[42] LinkedIn. 2021. LiFT: A Scalable Framework for Measuring Fairness in ML Applications. https://github.com/linkedin/LiFT. [Accessed: March 16th, 2023].

[43] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[44] Mykola Makhortykh, Aleksandra Urman, and Roberto Ulloa. 2021. Detecting race and gender bias in visual representation of AI on web search engines. In *International Workshop on Algorithmic Bias in Search and Recommendation*. Springer, 36–50.

[45] Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* (1947), 50–60.

[46] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).

[47] Microsoft. 2022. Microsoft Responsible AI Standard, v2. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4ZPmV.

[48] Hussein Mozannar, Mesrob Ohannessian, and Nathan Srebro. 2020. Fair learning with private demographic data. In *International Conference on Machine Learning*.

PMLR, 7066–7075.
[49] OECD. 2022. OECD AI Principles overview. https://oecd.ai/en/ai-principles.
[50] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. *Advances in neural information processing systems* 30 (2017).
[51] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
[52] Bird S., Dudik M., Edgar R., Horn D., Lutz R., Milan V., and Sameki M. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Proceedings of Machine Learning Research* 120 (2020), 1–8. https://doi.org/10.5555/3396126.3396130
[53] Nikolai V Smirnov. 1939. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscou* 2, 2 (1939), 3–14.
[54] Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. 2014. Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international* 2014 (2014).
[55] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*. PMLR, 3319–3328.
[56] The White House. 2022. Blueprint for an AI Bill of Rights: Making Automated Systems work for the American People. https://www.vox.com/recode/22455140/lemonade-insurance-ai-twitter.
[57] UNESCO. 2022. Draft text of the Recommendation on the Ethics of Artificial Intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000377897.
[58] Berk Ustun, Yang Liu, and David Parkes. 2019. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*. PMLR, 6373–6382.
[59] Cédric Villani. 2009. The wasserstein distances. In *Optimal transport*. Springer, 93–111.
[60] Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael Jordan. 2020. Robust optimization for fairness with noisy protected groups. *Advances in neural information processing systems* 33 (2020), 5190–5203.
[61] Linda F Wightman. 1998. LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series. (1998).
[62] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 666–677.
[63] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*. PMLR, 962–970.
[64] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International conference on machine learning*. PMLR, 325–333.
[65] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.

A SUPPLEMENTARY MATERIAL

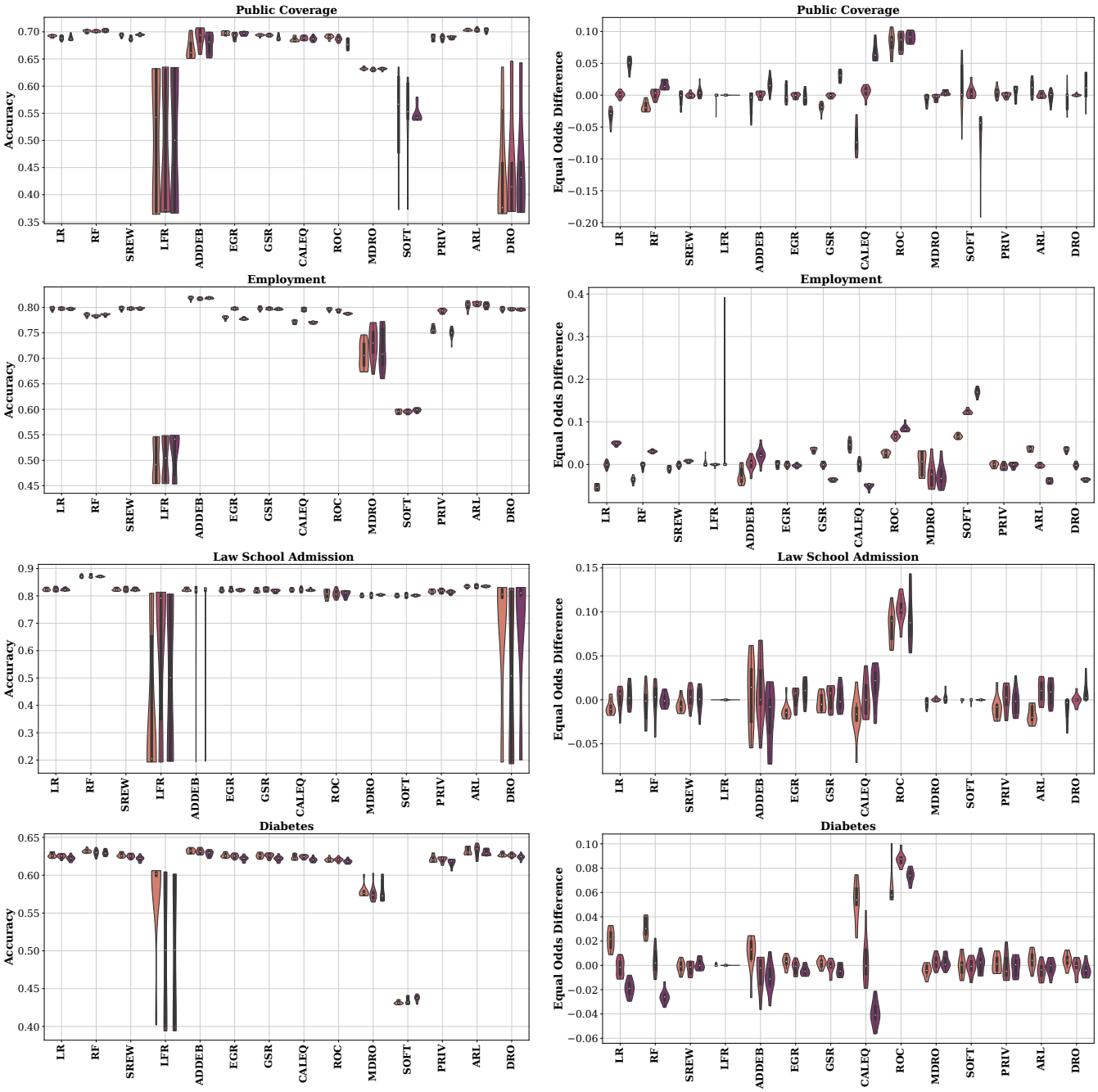


Figure 7: Plots showing the stability of our 14 classifiers over three different levels of noise in protected attributes (0.1, 0.5 and 0.9). For each dataset we present the stability of each classifiers' accuracy and EOD.