CERIAS Tech Report 2024-5

by Chowdhury Mohammad Rakin Haider
Center for Education and Research
Information Assurance and Security
Purdue University, West Lafayette, IN 47907-2086

IDENTIFYING INDUCED BIAS IN MACHINE LEARNING

by

Chowdhury Mohammad Rakin Haider

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Computer Science

West Lafayette, Indiana

May 2024

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Christopher Clifton, Chair

Department of Computer Science

Dr. Jean Honorio

Department of Computer Science

Dr. Ming Yin

Department of Computer Science

Dr. Bharat Bhargava

Department of Computer Science

Approved by:

Dr. Kihong Park

To my grandmother, Fazilatunnesa Barakat, and the valiant, Abrar Fahad.

ACKNOWLEDGMENTS

First and foremost, all the gratitude goes to Allah. I thank him for allowing me to travel on this path and approach the destination of my brief journey at Purdue.

I am grateful to my parents for their lifelong support in my academic endeavors. I am especially thankful to my loving wife for being with me throughout this challenging episode of my life. Special thanks to my sisters, especially Nazifa Karima, who helped us immensely in our first days here.

I want to express my heartfelt gratitude to Prof. Chris Clifton for his unwavering support and guidance as my PhD supervisor. Even during his hectic schedule, he always made himself available for discussion and suggestions. Without a doubt, he is a great academic mentor. Moreover, he cares deeply about his students' well-being. His encouraging words always helped me find interest in my research and continue putting in extra effort even when things were not working. I cannot thank him enough for his advice and support during my PhD.

I want to thank the advisory committee members, Prof. Jean Honorio and Prof. Ming Yin, for their thoughtful suggestions on my thesis. I want to especially acknowledge Prof. Ming's suggestions in re-writing and presenting our IUI2024 paper. During my brief time working with her, I gained a wealth of knowledge and experience. In addition, I would like to thank Prof. Shahbaz for being flexible in grading schedules when I was a TA in his course.

Finally, I want to thank the members of the Bangladeshi community at Purdue with whom we spent most of our time. I am thankful to Dr. S M Ferdous, Dr. Bushra Ferdousi, and Abdullah-Al-Mamun for making our days at Purdue more enjoyable. I shall cherish the foods, Adda, and 29 card games at your place for a long time.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$\mathcal{D}$	Dataset
$\mathbf{x}$	Feature vector
s	Sensitive attribute
p	Privileged group; short form of $s = p$.
u	Unprivileged group; short form of $s = u$.
y	Class label
n	Number of features
N	Number of samples in $\mathcal{D}$
$y^{\pm}$	Positive and negative class labels respectively; short form of $y = \pm$.
$\hat{y}$	Predicted label
$\hat{y}^{\pm}$	Positive and negative predictions respectively; short form of $\hat{y} = \pm$.
$DI(\mathcal{D})$	Disparate impact of dataset $\mathcal{D}$
α	Probability of positive class labels, $\mathbb{P}(y^{+})$
β	Ratio of positive and negative samples, $\frac{\mathbb{P}(p)}{\mathbb{P}(u)}$

ABBREVIATIONS

AI	Artificial Intelligence
APGD	Auto-PGD
BIM	Basic Iterative Method
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
CW	Carlini and Wagner Attack
DI	Disparate Impact
FAB	Fast Adaptive Boundary Attack
FairML	Fairness-aware Machine Learning
FBD	Fair Balanced Dataset
FGSM	Fast Gradient Sign Method
k-NN	k-Nearest Neighbor
MAR	Missing At Random
MCAR	Missing Completely At Random
MICE	Multivariate Imputation by Chained Equations
ML	Machine Learning
MNAR	Missing Not At Random
NB	Naive Bayesian
NN	Neural Network
PGD	Projected Gradient Descent
PIMA	Pima Indians Diabetes Database
PPD	Predictive Power Disparity
PR	Prejudice Remover
RBC	Reduction Based Classifier
SFBD	Synthetic Fair Balanced Dataset
SP	Statistical Parity
SVM	Support Vector Machine
TRADES	TRadeoff-inspired Adversarial DEfense via Surrogate-loss minimization

ABSTRACT

The last decade has witnessed an unprecedented rise in the application of machine learning in high-stake automated decision-making systems such as hiring, policing, bail sentencing, medical screening, etc. The long-lasting impact of these intelligent systems on human life has drawn attention to their fairness implications. A majority of subsequent studies targeted the existing historically unfair decision labels in the training data as the primary source of bias and strived toward either removing them from the dataset (de-biasing) or avoiding learning discriminatory patterns from them during training. In this thesis, we show label bias is not a necessary condition for unfair outcomes from a machine learning model. We develop theoretical and empirical evidence showing that biased model outcomes can be introduced by a range of different data properties and components of the machine learning development pipeline.

In this thesis, we first prove that machine learning models are expected to introduce bias even when the training data doesn't include label bias. We use the proof-by-construction technique in our formal analysis. We demonstrate that machine learning models, trained to optimize for joint accuracy, introduce bias even when the underlying training data is free from label bias but might include other forms of disparity. We identify two data properties that led to the introduction of bias in machine learning. They are the group-wise disparity in the feature predictivity and the group-wise disparity in the rates of missing values. The experimental results suggest that a wide range of classifiers trained on synthetic or real-world datasets are prone to introducing bias under feature disparity and missing value disparity independently from or in conjunction with the label bias. We further analyze the trade-off between fairness and established techniques to improve the generalization of machine learning models such as adversarial training, increasing model complexity, etc. We report that adversarial training sacrifices fairness to achieve robustness against noisy (typically adversarial) samples. We propose a fair re-weighted adversarial training method to improve the fairness of the adversarially trained models while sacrificing minimal adversarial robustness. Finally, we observe that although increasing model complexity typically improves generalization accuracy, it doesn't linearly improve the disparities in the prediction rates.

This thesis unveils a vital limitation of machine learning that has yet to receive significant attention in FairML literature. Conventional FairML literature reduces the ML fairness task to as simple as de-biasing or avoiding learning discriminatory patterns. However, the reality is far away from it. Starting from deciding on which features collect up to algorithmic choices such as optimizing robustness can act as a source of bias in model predictions. It calls for detailed investigations on the fairness implications of machine learning development practices. In addition, identifying sources of bias can facilitate pre-deployment fairness audits of machine learning driven automated decision-making systems.

1. INTRODUCTION

The last decade has witnessed an unprecedented rise in the development and application of artificial intelligence and machine learning. Widespread applications of ML have led to its use in making high-stakes decisions with long-lasting impacts on human life. Artificial intelligence (AI), especially machine learning (ML), is frequently being used in making consequential decisions such as hiring [1], credit lending [2], policing and bail sentencing [3], patient monitoring [4, 5], etc. Although the initial motivation behind automatic decision-making systems was to alleviate human labor, errors, and biases, mounting evidence [3–5] suggest that machine learning models tend to further exacerbate bias in their predictions. One of the most striking examples of unfairness in ML was reported in NothPointe's criminal risk assessment tool, commonly known as COMPAS [3]. The authors revealed that COMPAS unfairly predicts higher risk scores for black defendants. It triggered a broad spectrum of studies on *fairness-aware machine learning* (FairML).

FairML focuses on fairness concerns of ML models in real-world applications. Conventional wisdom dictates that real-world systems are interweaved with social biases and discriminations. Datasets collected from such biased systems typically include historically biased decisions. Researchers posit that biased decision labels eventually drive ML models toward unfairness. Preliminary works studied the direct or indirect relations between sensitive attributes and decision labels. If the sensitive attributes directly influence the decision making it is called direct bias. For example, the disenfranchisement of black women until the 1965 Voting Rights Act. On the other hand, indirect bias occurs when an attribute, commonly known as redlining attribute [6], acts as a proxy of the sensitive attributes. Rejecting loan applications from African American-dominated neighborhoods at a higher rate is a well-known instance of the redlining effect where *zipcode* is used as a proxy of the racial attribute. Thus redlining attributes contribute to unfair behavior even though the original sensitive attribute is not directly involved in the decision-making process. From the FairML perspective, direct bias indicates including the sensitive attributes in model training and predictions while indirect bias involves using proxy-sensitive attributes in model development. In [7], it was formally shown that removing sensitive attributes is not sufficient since redlin-

ing attributes contain sensitive information. Since it is often difficult to identify and remove redlining attributes, fairness interventions attempt to avoid learning discriminatory patterns in the presence of redlining attributes. In short, common fairness interventions assume that label bias in the dataset is responsible for unfairness in machine learning and the goal of FairML is to avoid learning the existing discriminatory patterns from these decision labels.

Fairness-interventions are broadly categorized as pre-processing [6, 8–11], in-processing [12–16] and post-processing techniques [17–20]. Pre-processing focuses on data de-biasing through sampling, relabeling, reweighing, etc. Similarly, in-processing techniques emphasize on avoiding historical biases through fairness constraints and modification of traditional ML algorithms. In contrast, post-processing techniques consider ML algorithms as black boxes and ensure fairness by balancing out the predictions among sensitive groups. Applications of causal reasoning to reduce the causal effect of sensitive attributes on the predictions have also been explored in literature [21–23]. FairML literature acknowledges the fairness vs accuracy trade-off [24, 25]. Consequently, we see applications of game theory [26, 27] and adversarial mechanism [28] as fairness interventions that try to optimize this trade-off. Authors have also proposed models that consider one or more desired properties, i.e., robustness [29], explainability [30], etc., in conjunction with fairness.

Irrespective of the stage of fairness intervention, a universal assumption in FairML literature is that biased decision labels lead to biased model outcomes. However, label bias is not the only source of bias in machine learning. In this work, we investigate the sources of bias beyond label bias. These sources are often systemic in nature, i.e., introduced by or inherent to the ML algorithm itself or the components of the ML development pipeline. Although *Systemic bias* is recognized as a paradigm of bias in social justice literature [31, 32], it is rarely considered in FairML. Systemic discrimination is typically defined as setting up a system, institution, or policy in such a way that it inherently disadvantages the protected group. A quintessential example of systemic bias is the higher stop-and-frisk rate at African American majority neighborhoods results in higher incarceration of African Americans even though the crime rate in those neighborhoods is similar to that of White ones. Without a doubt, the COMPAS dataset is biased against African Americans. Stereotypes against African American communities might have directly or indirectly skewed historical court de-

cisions against them (label bias). At the same time, systemic bias, i.e., strict policing, poor financial situation, lack of rehabilitation, etc., have contributed towards imbalanced arrest and sentencing rates; resulting in a skewed majority of African American samples in the COMPAS dataset. Instead of focusing on label bias, as was done in previous works, we pay attention to systemic bias in machine learning systems. We define systemic bias in machine learning as the biases that are inherent to the machine learning models themselves or introduced by parts of the development process. As a result, in ML, we refer to systemic bias as introduced bias or inherent bias. We investigate the extent of introduced bias in ML. In other words, we investigate whether label bias is a necessary condition of unfair ML. We identify several conditions where introduced bias appears in ML and establish them as a frequent phenomenon in real-world ML systems. We provide theoretical and empirical evidence of introduced bias under group-wise disparity in feature predictivity and group-wise disparity in the rates of missing values. In addition, we show that adversarial robustness against disparately effective evasion attacks can result in increased between-group disparity of model robustness. Finally, we study the disparity in classification complexity and its implications on model fairness.

We argue that systemic bias appears due to disparities in data collection, feature design, or other steps in the ML algorithm pipeline, in short disparities not directly related to label bias. These biases contribute to producing biased ML models. Given a dataset with fixed data properties, systemic biases pose an upper bound of model fairness beyond which may not be achievable with simple computational solutions. Therefore, it is imperative to study and identify conditions where systemic bias introduces or intensifies bias in machine learning. Similar to the case of societal, institutional, or political systemic bias, systemic bias in ML often requires system-wide changes and carefully designed practices. We hypothesize that machine learning models can be expected to introduce/ amplify bias in model outcomes under certain conditions irrespective of whether the training data contains label bias.

1.1 Problem Formulation

In this thesis, we examine whether sources of bias other than label bias exist which either exacerbates the existing AI systems's inequality or drives the model toward unfair predictions. In this regard, we examine systemic bias as a potential source of bias in ML. Typically, systemic bias is defined as the inherent tendency of a system to support unfair outcomes. It either introduces or amplifies existing discrimination by systematically deteriorating the status of the disadvantaged group. The effect of systemic bias on the unprivileged group is even more insidious since it provokes justification for unfair treatment towards them. From the perspective of machine learning, systemic disparities may appear innocuous but can introduce unfair outcomes in the long run. Therefore, it is also called introduced bias or induced bias. While ML models simply mimic the label bias, introduced bias drives the ML models towards further unfairness in their predictions. In this work, we take the first attempt to study introduced bias in machine learning. We show that certain systemic disparities during the ML development process introduce bias in the model outcome independently of label bias. In our analysis, we assess the extent of introduced bias in ML by measuring the disparate group-wise prediction rates when systemic disparities are imposed on the training data. We explore the following instances where introduced bias can appear in ML.

- We first study whether the between-group disparity in the efficacy of the collected features is an instance of introduced bias.
- Secondly, we examine the existence of introduced bias due to the disparity in group-wise rates of missing values.
- Next, we explore whether achieving robustness against noisy samples is at odds with ML fairness. More specifically, we examine whether optimizing for robustness can sacrifice model fairness.
- Finally, we study the fairness implications of increasing model complexity and disparities in group-wise classification complexities.

Disparity in feature predictivity appears from improper feature engineering and selection. Similarly, the disparity in noise and missingness typically results from the discrepancies in the data collection mechanism. Missing value in data samples is a well-known phenomenon in machine learning. About 45% of the datasets in the renowned UCI dataset repository contain missing values [33]. Missing values can appear due to partial survey completion by the participants, missing by nature such as middle name, or non-response to optional items such as online profile, college application, etc. In addition to missing values, a group of individuals may be under-represented when representative samples are missing from the dataset. We consider both missing values and missing samples under the umbrella of missing data. Data imputation is a widely approved technique to handle missing data. Since the group-wise rates of missing values are beyond the control of ML developers and largely based on the underlying data collection mechanisms, we consider such disparity as an instance of systemic bias. We investigate whether such disparity in missing value appearance can introduce bias in machine learning. Furthermore, we study the interaction of model fairness with the stability of model predictions, typically known as model robustness. We argue that noisy samples may appear at a disparate rate among the sensitive groups. A lack of robustness against noisy samples in one group leads to unfair prediction rates against adversaries. Finally, we observe that fairness is not monotonically increasing with model complexity. As a result, established practices of over-parameterization of large-scale neural networks may lead to unfair model choices. Note that the disparities in both robustness and classification complexity are inherent to the classification task at hand and independent of the label bias in the dataset.

Identifying sources of bias necessitates removing confounding factors from the analysis. As a result, in the formal analysis, we start with fairness guarantees such as equal base rates across the demographic sub-groups, no selection bias, i.e., balanced collection of samples with no majority or minority groups, etc. We analyze whether in the absence of other types of disparities introducing certain disparities in the dataset leads the joint-optimal models to produce unfair prediction rates. Similarly, during the experimental evaluations, it is necessary to avoid other types of disparities in the training process to identify the extent of bias introduced by a specific disparity. As a result, we exploit synthetically generated

data that offers the flexibility to enforce fairness guarantees. For each specific evaluation, we generate the synthetic samples such that they contain only one type of disparity. It allows us to measure the bias introduced in the final models by a specific type of disparity. We further extend the empirical evaluations to real-world datasets so that we can analyze the frequency of introduced bias in real-world applications. Here, we first train baseline models on the unaltered publicly available real-world dataset. Then, we induce the type of disparity under consideration. In this way, we can determine the level of introduced bias by comparing the bias in model performance with or without the specific disparity.

In short, this thesis explores the hypothesis that systemic biases lead to exacerbating AI systems' inequality in outcomes and explores conditions where such disparities can introduce bias in ML. A thorough understanding of the effects of these biases will contribute to designing better ML systems and development practices.

1.2 Contributions

The contributions of this work are manifold. We summarize the key contributions as follows,

- We offer a concise survey of state-of-the-art fairness interventions and fairness metrics. (Chapter 2)
- We show that ML models can be expected to produce group-wise inequality *even when the training data offers certain fairness guarantees*. We start with data having no base rate disparity (equal favorable/unfavorable outcomes between groups), an absence of majority or minority group (balanced), and all examples are presumed correct. We show that even with such fair training data, an optimal accuracy ML model is *expected* to introduce disparate impact (different favorable outcome rates between groups) in predictions. In particular, when the optimal group-wise models are different we theoretically show that Bayes-optimal ML models can be expected to introduce such bias. We also analyze the cases where the allocable resources are either insufficient or surplus. We substantiate our theoretical claims with empirical estimates using multiple machine learning approaches and real-world datasets. (Chapter 3)

- Secondly, we prove that missing value disparity is sufficient to introduce bias in model behavior. We offer constructive proof of missing value disparity-induced bias involving simple mechanisms, showing why missing value handling leads to outcome (prediction) bias. We prove by construction that for a Bayes-optimal classifier, missing value disparity handled by dropping or mean imputation is sufficient to induce outcome bias when the group-wise optimal models are different. The use of a Bayes-optimal classifier implies that the construction is extendable to other optimal accuracy classifiers. It indicates that this introduced bias is inherent to techniques that maximize accuracy. We further show that disparity in missing value rates exacerbates unfairness in imbalanced datasets. We also offer empirical evidence of outcome bias induced by missing data in more complex situations, including real-world data that may contain several sources of bias, more complex missing value imputation methods (k-NN imputation [34], MICE [35], matrix completion [36]), and both traditional (SVM, Neural Network) and state-of-the-art fairness-aware classifiers. We report that in cases where the changes in prediction probabilities were too low to alter the predicted labels, missing value disparity drives the classifiers to disproportionately promote the relative rankings of one group and demote the other. Although we mainly focus on missing values in the training data, we found that induced bias became even more prominent when the test data included missing values. Having shown that missing value disparity is sufficient to induce bias, we further investigate the extent of this induced bias when coupled with other sources of bias. We hypothesize that missing data disparity will exacerbate existing outcome bias in situations where there are multiple sources of bias. We validate this with experiments on real-world datasets which commonly contain selection bias and label bias among other types of disparities. We observed missing value disparity-induced bias in model behavior across different types of classifiers, multiple causal MAR mechanisms for missing values, and several imputation mechanisms. We conclude that induced bias through missing value disparity is a common occurrence and threatens the fairness of machine learning models. (Chapter 4)

- Furthermore, we investigate the impact of adversarial training on the group-wise fairness of the classifiers. We study the interaction between model robustness and group-wise fairness in four facial image classification tasks. In each of these tasks, we obtain relatively robust classifiers by applying nine variations of adversarial training algorithms and evaluate their performance against state-of-the-art adversarial attacks. We compare the group-wise fairness of robust classifiers to that of the robustness-unaware baseline classifiers. We observe that adversarial training introduces bias in model predictions. In the worst case, adversarial training increased the difference between group-wise robust accuracy and group-wise false positive rates by a magnitude of 27.6% and 39% respectively. Moreover, the increased disparity is observed in 83.4% of classifier and adversarial attack combinations considered in our experiment. Similarly, in our experiment, adversarial training also increased the group-wise disparity in robustness bias defined in [37]. The experimental results indicate that, in search of robust models, state-of-the-art adversarial training often improves the group-wise robustness at different rates and thus eventually increases the disparity. Since both robustness and group fairness are desirable in our classifiers, we propose two re-weighting techniques to improve the group fairness of adversarially trained models while maintaining high robustness against evasion attacks. Both of them are easy to implement and add minimal overhead to the adversarial training. The re-weighting techniques place a higher emphasis on the more vulnerable group to reduce the disparity in robust accuracies. Empirical evaluation reveals that the proposed re-weighting techniques typically reduce the robust accuracy disparity without sacrificing much robustness. (Section 5.1)

- Finally, we analyzed the fairness implication of over-parameterization of deep neural networks. We observed that although generalization accuracy monotonically increases with model complexity, model fairness demonstrates a non-monotonic relation with model complexity. It indicates that without careful tuning of model complexity, unfairness can be introduced from model selections. (Section 5.2)

2. BACKGROUND

In this chapter, we introduce the prerequisites required for a comprehensive view of the research conducted under this dissertation. We first describe the fairness notions found in ML literature and the proposed fairness metrics used to assess the fairness of ML models. We then discuss the state-of-the-art fairness interventions proposed in the literature. Typically, fairness-aware algorithms optimize one or more of the fairness metrics in their process. Afterward, we present the underlying reasons that delineate the inevitability of missing values in the ML dataset and offer a categorization of them based on the causal relation of the missing values and the observed or unobserved data in the dataset. We also introduce state-of-the-art missing value handling mechanisms typically known as imputations. Finally, we introduce adversarial sample generation techniques, i.e., evasion attacks and adversarial training mechanisms developed to achieve robustness against evasion attacks.

2.1 Fairness Metrics

FairML literature has defined an abundance of fairness metrics to assess model fairness. Each fairness metric caters to certain equality requirements of the context. Three main family of fairness definitions are group fairness, individual fairness and sub-group fairness metrics. We describe these fairness notions below.

2.1.1 Group Fairness

Group fairness focuses on statistical equality among sensitive groups. According to [38], proposed group fairness definitions consider either base rates, group-conditioned accuracy, or group-conditioned calibration. Let $\mathcal{D} = \{\mathbf{x}^{(k)}, s^{(k)}, y^{(k)}\}_{k=1}^{N}$ be the training dataset, where $\mathbf{x}$, s and y are the set of n features, the sensitive attribute, and the target variable, respectively. Let $s \in \{p, u\}$, where p and u denote privileged and unprivileged group, respectively. Let's also assume that $\hat{y}$ is the predicted model outcome. We provide a short description of the most common group fairness definitions below. For a more comprehensive discussion of fairness metrics, see [38, 39].

- *Disparate Impact* is a fairness metric based on base rates. It defines fairness as equality between the probability of positive prediction in the unprivileged group and the privileged group.

$$\mathbb{P}(\hat{y}^+ \mid u) = \mathbb{P}(\hat{y}^+ \mid p) \tag{2.1}$$

Here, $\mathbb{P}(\hat{y}^+ \mid u)$ is short form of $\mathbb{P}(\hat{y}^+ \mid s = u)$. The fairness criteria is modelled as a ratio or a difference to obtain the fairness metric. If θ is the trained model,

$$DI(\mathcal{D}; \theta) = \frac{\mathbb{P}(\hat{y}^+ \mid u)}{\mathbb{P}(\hat{y}^+ \mid p)} \tag{2.2}$$

The closer $DI(\mathcal{D}; \theta)$ is to 1, θ is considered more fair. Disparate impact can also be defined as a characteristic of the dataset rather than the model. That means,

$$DI(\mathcal{D}) = \frac{\mathbb{P}(y^+ \mid u)}{\mathbb{P}(y^+ \mid p)} \tag{2.3}$$

Calders et al. [17] formulated the notion of statistical parity (SP) as the difference between group-wise favorable prediction rates or selection rates (SR) instead of their ratio. A model is considered fair if the difference between the group-wise favorable prediction rates is zero. Otherwise, it is either biased towards the privileged group (> 0) or the unprivileged (< 0). Formally,

$$SP(\theta) = \mathbb{P}(\hat{y}^+ \mid p) - \mathbb{P}(\hat{y}^+ \mid u) \tag{2.4}$$

- *Equal Opportunity* is satisfied if the true positive rates of privileged and unprivileged groups are equal. Formally,

$$\mathbb{P}(\hat{y}^+ \mid y^+, p) = \mathbb{P}(\hat{y}^+ \mid y^+, u) \tag{2.5}$$

- *Equalized Odds* is a fairness constraint that requires both the true positive rates and the false positive rates in both groups to be equal. It follows,

$$\mathbb{P}(\hat{y}^+ \mid y, p) = \mathbb{P}(\hat{y}^+ \mid y, u) \text{ s.t. } y \in \{y^+, y^-\} \tag{2.6}$$

- *Equalized Accuracy* is considered fair if the accuracy of prediction in each group is equal. That means, $\mathbb{P}(\hat{y} = y \mid p) = \mathbb{P}(\hat{y} = y \mid u)$.

Following the formulation in (2.4), we capture the equality (or inequality) in accuracy (ACC) and false positive rates (FPR), as the group-wise differences shown below,

$$SP(ACC; \theta) = \mathbb{P}(\hat{y} = y \mid p) - \mathbb{P}(\hat{y} = y \mid u) \tag{2.7}$$

$$SP(FPR; \theta) = \mathbb{P}(\hat{y}^+ \mid y^-, p) - \mathbb{P}(\hat{y}^+ \mid y^-, u) \tag{2.8}$$

Disparate impact can be classified as a metric based on base rates. On the other hand, equalized odds, equalized opportunity, and equalized accuracy are metrics based on group-conditioned accuracy. The majority of FairML literature focuses on group-fairness mechanisms.

2.1.2 Individual Fairness

Individual Fairness is defined in [14] as similar treatment towards similar individuals. Let, a and b be two individuals, a model f maps them to $f(a)$ and $f(b)$. According to [14], if $d(a, b)$ is the distance between the individuals, then individual fairness requires that $d_f(f(a), f(b)) < Ld(a, b)$ where $L > 0$ is the Lipschitz constant. That means, the outcome distribution from a and b is required to be less distinguishable than $d(a, b)$. Individual fairness have been addressed in representation learning [14], clustering [40], graph learning [41], etc. In [42] a post-processing mechanism is proposed to ensure individual fairness.

2.1.3 Sub-group Fairness

Group fairness mechanisms consider only a handful of groups with respect to a sensitive attribute. However, the intersectionality of sensitive attributes can lead to a combinatorially large number of sub-groups in the population. For example, considering only race can split the population into Caucasians and non-Caucasians whereas adding sex in the set of sensitive attributes multiplies the number of sub-groups by at least two. On the other hand, individual fairness considers each individual as a group by themselves. Therefore, sub-group fairness is an attempt to find a middle ground between group fairness and individual fairness. Sub-group fairness defines fairness as equality among a rich class of subgroups. Here, subgroups are indicated by a set of indicator functions $\mathcal{G}$ where each indicator $g_k(\mathbf{x}) \in \mathcal{G}$ indicates sample $\mathbf{x}$'s membership of k^{th} subgroup. According to [24, 25], a classifier f satisfies γ-subgroup fairness with respect to fairness metric FM (e.g., statistical parity) if for all k,

$$\alpha(g_k)\beta_{FM}(f, g_k) \leq \gamma \tag{2.9}$$

$$\text{where, } \alpha(g_k) = \mathbb{P}(g_k(\mathbf{x}) = 1) \tag{2.10}$$

$$\beta_{FM}(f, g_k) = \; \mid FM(f) - FM(f \mid g_k(x) = 1) \mid \tag{2.11}$$

2.2 Related Work

The goal of fairness-aware machine learning is to develop models whose predicted outcomes are non-discriminatory with respect to sensitive attributes such as race, sex, marital status, etc. Fairness interventions are broadly categorized as pre-processing, in-processing, and post-processing techniques. Pre-processing techniques focus on de-biasing the dataset to train unfair models. In-processing techniques acknowledge that real-world datasets are extremely difficult to de-bias perfectly and instead focus on modifying traditional machine learning techniques to avoid perpetuating bias in model outcomes. Finally, post-processing techniques consider machine learning models as black-box and modify the outcomes to ensure equal distributions of the predictions among the sensitive groups. The following sub-sections contain details on each type of fairness intervention.

2.2.1 Pre-processing

It is typically assumed that bias in training data eventually leaks into trained models. Therefore, early work emphasized on pre-processing techniques, i.e., modifying the dataset to reduce the impact of historical bias [6, 8–10]. Among the three bias reduction techniques proposed by Kamiran et al., [6, 9, 10], massaging modifies the class labels of individuals closer to the decision boundary. It reflects the promotion of discriminated individuals and the demotion of favored ones. A similar method is recommended by [21]. In contrast, preferential sampling [10] and reweighing [9] are sampling-based de-biasing algorithms. They sample the individuals from the sensitive groups such that a desired group-fairness condition holds. Finally, Feldman et al. [8] altered the redlining attributes to make them non-predictive of the sensitive attribute. According to their definition, training data is fair if the sensitive attributes are not predictable from the features. Their proposed method modifies the redlining feature distributions, such that an individual's within-group relative ranking remains unchanged but the sensitive attributes become unpredictable from the features. A more recent pre-processing technique [11] involves oversampling to obtain a balanced dataset. They argue that de-biasing with data modification leads to the loss of demographic correlations with the features and consequently fuels a trade-off between accuracy and fairness. As a result, they propose generating synthetic samples of under-represented groups so that it eliminates bias in data.

2.2.2 In-processing

In-processing techniques tend to modify traditional machine learning algorithms to reduce unfairness in the predicted outcomes. Authors have proposed modification of existing algorithms [12, 17], fair regularization terms [13], and optimization with fairness constraints [14–16] as fairness-interventions. Agarwal et al. [26, 27] applied a game theoretic approach to model the fairness-accuracy tradeoff as a game between two opponents. The solution is designed as two players taking turns in optimizing one end of the trade-off.

2.2.3 Post-processing

Post-processing techniques [17–20] modify the predicted outcomes or model parameters to obtain non-discriminatory predictions. Calders et al. [17] increased $\mathbb{P}(y^+ \mid u)$ to reduce the difference in positive prediction probability among the sensitive groups. In contrast, if the overall positive prediction rate becomes higher than the rate of positive class labels in the training data, $\mathbb{P}(y^+ \mid p)$ is decreased. The modified prediction probabilities are later used to build Naive Bayesian models. In [18], the authors derived an equal odds predictor $\tilde{\theta}$ from an existing model θ by solving a linear program with conditional prediction probabilities of θ. On the other hand, [19] starts with two group-wise calibrated classifiers and modifies one of them to equalize weighted odds between the groups.

2.2.4 Others

Applications of causal reasoning have been proposed to remove both direct and indirect discrimination from either the dataset or the classifiers [21–23]. Fairness-aware machine learning models have been proposed for specific application areas as well, such as natural language processing [43], graph embedding [44–46], and computer vision [47].

3. INTRODUCED BIAS FROM FEATURE DISPARITY

In this section, we focus on the impacts of systemic bias in the form of disparities in feature distributions. More specifically, we expand on the scenario where the existing features, either designed, selected, or collected, work better for one group than the other. Such disparity in feature predictivity is likely to occur due to lower inclusion of unprivileged group members in model development or, in other cases, due to better resource availability for the privileged group. In section 3.3, we refer to several real-world studies that resulted in designing features more suited to the privileged community. We show that such feature disparity leads to biased model outcomes even when the underlying data is free from such bias. The existence of such systemic bias encourages the use of inclusive or participatory design mechanisms so that we can design or collect better predictors for all the sub-groups.

3.1 Formal Analysis

In this chapter, we show that an optimal classifier can produce unfair outcomes even if the training data is free from label bias but contains feature disparity. Before moving on to the analysis, we first define feature disparity as follows,

Definition 3.1.1. ***Feature Disparity:*** Feature disparity is a type of bias that appears from feature design or selection. If the feature set included in the dataset is collectively a better predictor for one group than the other, then we call it feature disparity. Given feature vectors $\mathbf{x}$, a binary sensitive attribute $s \in \{p, u\}$ and group-wise optimal classifiers $f^*(\mathbf{x}; \theta_s)$, feature disparity indicates,

$$E_{\mathcal{D}}[f^*(\mathbf{x}; \theta_p) = y \mid p] \neq E_{\mathcal{D}}[f^*(\mathbf{x}; \theta_u) = y \mid u]$$

Let $\mathbb{P}(p) = \beta\mathbb{P}(u)$. If one group has a majority in $\mathcal{D}$, i.e., $\beta \neq 1$, then the predictions will be dominated by the majority. The behavior of ML models in such scenarios was investigated in [11]. It is shown that the model fairness can be improved by generating synthetic instances of the minority group. To avoid confounding factors such as dataset imbalance and label bias, we assume that the dataset doesn't involve majority or minority groups (or the imbalance

has been rectified), it doesn't contain any discriminatory decision labels and has equal group-wise prior probabilities. We prove the existence of model-induced bias from feature disparity even with such balanced and fair datasets.

3.1.1 Fair Balanced Dataset (FBD)

We assume that the fair and balanced training dataset has identical group *outcomes* (i.e., equal base rates and sample count). While all labels are correct, the features are insufficient for a "perfect" (100% accuracy) classifier. The only distinction between the privileged and unprivileged groups is that the Bayes-optimal model for each group is different, and without loss of generality the Bayes-optimal model for the privileged group is more accurate. We also assume (in keeping with legal requirements in many countries) that we cannot use the sensitive attributes to explicitly separate the privileged and unprivileged groups when making predictions, i.e., "fairness-through-unawareness". We show that feature disparity leads to outcome bias in the joint optimal model.

Let $\mathcal{D}$ be an FBD, i.e., $DI(\mathcal{D}) = 1$ and $y \perp\!\!\!\perp s$. If y^+ and y^- indicate positive and negative samples, $P(y^+) = P(y^+ \mid s) = \alpha$ and $P(y^-) = P(y^- \mid s) = 1 - \alpha$. Assuming normally distributed features, a feature with a higher distinction between values of the positive and the negative labeled samples is more predictive. In contrast, a non-predictive feature's distribution is independent of its class labels. A proxy measure of feature disparity is the separation of positive and negative sample distributions. Here, separation *sep* between two normal distributions $\mathcal{N}(\mu_1, \sigma)$, and $\mathcal{N}(\mu_2, \sigma)$ is,

$$sep = \frac{\mid \mu_1 - \mu_2 \mid}{\sigma} \tag{3.1}$$

Let the most predictive feature set for the privileged and unprivileged groups be $\{x_1, x_2, \ldots, x_r\}$ and $\{x_{r+1}, x_{r+2}, \ldots, x_{2r}\}$ respectively. Let, x_{i}^{sy} indicate the random variable cor-

responding to the feature x_{i} of the members of group s in class y and $\mu_{\mathrm{i}}^{sy} = E[x_{\mathrm{i}}^{sy}]$. If $\mathrm{i} \leq r$,

$$x_{\mathrm{i}}^{p+} \sim \mathcal{N}(\mu_{\mathrm{i}}^{p+}, \sigma_{\mathrm{i}}^2) \qquad x_{\mathrm{i}}^{p-} \sim \mathcal{N}(\mu_{\mathrm{i}}^{p-}, \sigma_{\mathrm{i}}^2)$$
$$x_{\mathrm{i}}^{u+}, x_{\mathrm{i}}^{u-} \sim \mathcal{N}(\mu_{\mathrm{i}}^{p_{avg}} + \delta, \sigma_{\mathrm{i}}^2) \qquad \mu_{\mathrm{i}}^{p_{avg}} = \frac{\mu_{\mathrm{i}}^{p+} + \mu_{\mathrm{i}}^{p-}}{2}$$

Here, $\delta(= 0)$ is a constant. Similarly, we define the distributions of $x_{r+1 \leq \mathrm{i} \leq 2r}$. Finally, $x_{2r < \mathrm{i} \leq n} \sim \mathcal{N}(\mu_{\mathrm{i}}, \sigma_{\mathrm{i}}^2)$. Without loss of generality, we assume $\mu_{\mathrm{i}}^{s-} < \mu_{\mathrm{i}}^{s+}$.

We consider a Bayesian classifier θ, the theoretical optimal classifier for normally distributed features. Estimating prediction probabilities with Bayesian boundary and multivariate normal distribution requires generalized chi-squared distribution estimation [48]. Except for limited cases such as lower-dimensional linear boundary, the lack of known closed-form probability estimates has led to several computational methods [48]. As such, we limit analytical results to $\sigma_{\mathrm{i}} \not\perp\!\!\!\perp y$ (linear boundary) with $r = 1$, $n = 2$ in Section 3.1.2. Consequently, the unprivileged group is less separable than the privileged if $\sigma_1 < \sigma_2$; the groups are equally separable when $\sigma_1 = \sigma_2$. We show that even in such a simplified case, the model is expected to produce outcome bias not present in the training data. When $n > 2$, let $\sigma_p = \sigma_{\mathrm{i}}$ such that $1 \leq \mathrm{i} \leq r$. Similarly, we defined $\sigma_u = \sigma_{\mathrm{i}}$ where $r+1 \leq \mathrm{i} \leq 2r$. In this setup, feature disparity can be imposed by setting $\sigma_p < \sigma_u$. We provide computational estimations of introduced bias for $n > 2$ in Section 3.2.3.

3.1.2 Less Separable Unprivileged Group

A Bayesian classifier θ computes $\mu_{\mathrm{i}}^{\theta}$ and $\sigma_{\mathrm{i}}^{\theta}$ for each feature x_{i}. For $\mathbf{x} \in \mathcal{D}$, the decision boundary of θ is defined as,

$$\mathbb{P}(\hat{y}^+ \mid \mathbf{x}) > \mathbb{P}(\hat{y}^- \mid \mathbf{x}) \tag{3.2}$$

Here, $\mathbb{P}(\hat{y}^+ \mid y^+, s)$ and $\mathbb{P}(\hat{y}^+ \mid y^-, s)$ are group-wise true positive rate (TPR_s) and false positive rate (FPR_s). Therefore, the group-wise selection rate (SR_s) is,

$$\mathbb{P}(\hat{y}^+ \mid s) = \alpha \mathbb{P}(\hat{y}^+ \mid y^+, s) + (1-\alpha)\mathbb{P}(\hat{y}^+ \mid y^-, s) \tag{3.3}$$

$$\text{where } \mathbb{P}(\hat{y}^+ \mid y, s) = \int\limits_{\substack{\mathbf{x} \text{ s.t.} \\ P(\hat{y}^+|\mathbf{x}) > P(\hat{y}^-|\mathbf{x})}} P(\mathbf{x} \mid y, s) d\mathbf{x} \tag{3.4}$$

Applying the conditional independence assumption of the Naive Bayesian classifier (NBC) we expand the decision boundary in Equation 3.2 as shown below.

$$\begin{aligned}
& P(\hat{y}^+ \mid \mathbf{x}) > P(\hat{y}^- \mid \mathbf{x}) \\
\Rightarrow & \frac{P(\hat{y}^+)P(\mathbf{x} \mid \hat{y}^+)}{P(\hat{y}^-)P(\mathbf{x} \mid \hat{y}^-)} \geq 1 \\
\Rightarrow & \frac{P(\hat{y}^+)\prod_{\mathrm{i}=1}^{n} P(x_{\mathrm{i}} \mid \hat{y}^+)}{P(\hat{y}^-)\prod_{\mathrm{i}=1}^{n} P(x \mid \hat{y}^-)} \geq 1 && \text{(Conditional independence)} \\
\Rightarrow & \log\frac{\alpha}{1-\alpha} + \sum_{\mathrm{i}=1}^{n} \log \frac{\exp\left(\frac{-(x_{\mathrm{i}}-\mu_{\mathrm{i}}^{\theta+})^2}{2(\sigma_{\mathrm{i}}^{\theta})^2}\right)}{\exp\left(\frac{-(x_{\mathrm{i}}-\mu_{\mathrm{i}}^{\theta-})^2}{2(\sigma_{\mathrm{i}}^{\theta})^2}\right)} \geq 0 \\
\Rightarrow & c_\alpha + \sum_{\mathrm{i}=1}^{n} \left(\frac{-(x_{\mathrm{i}}-\mu_{\mathrm{i}}^{\theta+})^2}{2(\sigma_{\mathrm{i}}^{\theta})^2} + \frac{(x_{\mathrm{i}}-\mu_{\mathrm{i}}^{\theta-})^2}{2(\sigma_{\mathrm{i}}^{\theta})^2} \right) \geq 0 \\
\Rightarrow & c_\alpha + \sum_{\mathrm{i}=1}^{n} \frac{(2x_{\mathrm{i}} - \mu_{\mathrm{i}}^{\theta-} - \mu_{\mathrm{i}}^{\theta+})(\mu_{\mathrm{i}}^{\theta+} - \mu_{\mathrm{i}}^{\theta-})}{2(\sigma_{\mathrm{i}}^{\theta})^2} \geq 0 \\
\Rightarrow & c_\alpha + \sum_{\mathrm{i}=1}^{n} \frac{(x_{\mathrm{i}} - \mu_{\mathrm{i}}^{\theta avg})(\mu_{\mathrm{i}}^{\theta+} - \mu_{\mathrm{i}}^{\theta-})}{(\sigma_{\mathrm{i}}^{\theta})^2} \geq 0 \\
\Rightarrow & c_\alpha + \sum_{\mathrm{i}=1}^{2} \frac{(x_{\mathrm{i}} - \mu_{\mathrm{i}}^{\theta^{avg}})(\mu_{\mathrm{i}}^{\theta+} - \mu_{\mathrm{i}}^{\theta-})}{(\sigma_{\mathrm{i}}^{\theta})^2} \geq 0 && \text{(Since, } \forall \mathrm{i} > 2r; \mu_{\mathrm{i}}^{\theta+} - \mu_{\mathrm{i}}^{\theta-} = 0 \text{ and } r = 1) \\
\therefore & \frac{(x_1 - \mu_1^{avg})\Delta}{2(\sigma_1^{\theta})^2} + \frac{(x_2 - \mu_2^{avg})\Delta}{2(\sigma_2^{\theta})^2} + c_\alpha \geq 0 && (3.5)
\end{aligned}$$

Here, $c_\alpha = \log \frac{\alpha}{1-\alpha}$. The range of feature x_i that secures positive prediction is,

$$b_\mathrm{i}(x_\mathrm{j}) \leq x_\mathrm{i} < \infty \qquad \mathrm{i}, \mathrm{j} \in \{1, 2\} \text{ s.t. } \mathrm{i} \neq \mathrm{j} \tag{3.6}$$
$$\text{Here, } b_\mathrm{i}(x_\mathrm{j}) = \mu_\mathrm{i}^{avg} - \frac{2(\sigma_\mathrm{i}^\theta)^2}{\Delta}\left(c_\alpha + \frac{(x_\mathrm{j} - \mu_\mathrm{j}^{avg})\Delta}{2(\sigma_\mathrm{j}^\theta)^2}\right)$$

We expand Equation 3.4 as follows,

$$\begin{aligned}
P(\hat{y}^+ \mid y, s) &= \int\limits_{\substack{x \text{ s.t.} \\ P(\hat{y}^+|x) > P(\hat{y}^-|x)}} P(x \mid y, s)dx \\
&= \int\limits_{\substack{x \text{ s.t.} \\ P(\hat{y}^+|x) > P(\hat{y}^-|x)}} \prod_\mathrm{i} P(x_\mathrm{i} \mid y, s)dx_\mathrm{i} \qquad \text{(Using the Naive Bayes assumption.)} \\
&= \int\limits_{-\infty}^{\infty} P(x_1 \mid y, s) \left(\int\limits_{\substack{x_2 = \\ b_2(x_1)}}^{\infty} P(x_2 \mid y, s)dx_2 \right) dx_1 \\
&= \int\limits_{-\infty}^{\infty} \frac{P(x_1 \mid y, s)}{2} \left[1 - erf\left(\frac{b_2(x_1) - \mu_2^{sy}}{\sqrt{2}\sigma_2}\right)\right] dx_1 \\
\therefore \mathbb{P}(\hat{y}^+ \mid y, s) &= \frac{1}{2} - \frac{1}{2} E_{x_1 \sim P(x_1^{sy})}\left[erf\left(\frac{b_2(x_1) - \mu_2^{sy}}{\sqrt{2}\sigma_2}\right)\right]
\end{aligned} \tag{3.7}$$

Here, erf is the error function [49]. According to [49] and [50],

$$E_{x \sim \mathcal{N}(\mu, \sigma^2)}\left[erf\left(mx + n\right)\right] = erf\left(\frac{m\mu + n}{\sqrt{1 + 2m^2\sigma^2}}\right) \tag{3.8}$$

Applying Bayesian classification theory, NBC is expected to yield,

$$\mu_\mathrm{i}^{\theta+} = \frac{1}{2}(\mu_\mathrm{i}^{p+} + \mu_\mathrm{i}^{u+}); \quad \mu_\mathrm{i}^{\theta-} = \frac{1}{2}(\mu_\mathrm{i}^{p-} + \mu_\mathrm{i}^{u-}); \quad \sigma_\mathrm{i}^\theta \geq \sigma_\mathrm{i}$$

By construction,

$$\begin{aligned}
\mu_2^{p_{avg}} &= \mu_2^{p+} \quad \mu_1^{p_{avg}} - \mu_1^{p+} = -\frac{\Delta}{2} \\
\mu_1^{u_{avg}} &= \mu_1^{p+} \quad \mu_2^{u_{avg}} - \mu_2^{p+} = -\frac{\Delta}{2}
\end{aligned}$$

To simplify the expressions, we define the following notations,

$$c_{denom} = 2\sqrt{2}\Delta\sqrt{(\sigma_1^\theta)^4\sigma_2^2 + (\sigma_2^\theta)^4\sigma_1^2} \quad c_2 = \frac{4(\sigma_1^\theta\sigma_2^\theta)^2}{c_{denom}}$$
$$c_1 = \frac{\Delta^2(\sigma_2^\theta)^2}{c_{denom}} \qquad c_3 = \frac{\Delta^2(\sigma_1^\theta)^2}{c_{denom}} \qquad c_{avg} = \frac{c_1 + c_3}{2}$$

Using 3.7 and 3.8, we get,

$$\mathbb{P}(\hat{y}^+ \mid y^+, s) = \frac{1}{2} + \frac{1}{2}erf\left[c_1\mathbb{1}_{s=p} + c_3\mathbb{1}_{s=u} + c_2\right] \tag{3.9}$$
$$\mathbb{P}(\hat{y}^+ \mid y^-, s) = \frac{1}{2} - \frac{1}{2}erf\left[c_1\mathbb{1}_{s=p} + c_3\mathbb{1}_{s=u} - c_2\right] \tag{3.10}$$
$$\tag{3.11}$$

Having $\sigma_1 < \sigma_2$ and $\mu_{\mathrm{i}}^{s+} - \mu_{\mathrm{i}}^{s-} = \Delta; \forall \mathrm{i}$, according to the definition of pooled variance, $\sigma_1^\theta < \sigma_2^\theta$. Since the error function is a strictly increasing function, from Equations 3.9 and 3.10, we conclude

$$P(\hat{y}^+ \mid y^+, p) > P(\hat{y}^+ \mid y^+, u) \tag{3.12}$$
$$\text{and, } P(\hat{y}^+ \mid y^-, p) < P(\hat{y}^+ \mid y^-, u) \tag{3.13}$$

Equations 3.12 and 3.13 conclude that feature disparity results in imbalanced TPR and FPR. We now show the introduced disparity in selection rates. More precisely, $SR_p > SR_u$ when $\alpha < 0.5$. Extending 3.3 with 3.9 and 3.10, it is equivalent to show that:

$$\alpha\left[erf(c_1 + c_2c_\alpha) - erf(c_3 + c_2c_\alpha)\right] > (1-\alpha)\left[erf(c_1 - c_2c_\alpha) - erf(c_3 - c_2c_\alpha)\right] \tag{3.14}$$

Since $\sigma_1 \neq \sigma_2$, it implies that $c_1 \neq c_3$. Therefore, from Figure 3.1, we conclude that:

$$erf(c_1 \pm c_2c_\alpha) - erf(c_3 \pm c_2c_\alpha) \propto \frac{d}{dx}erf(x)\Big|_{c_{avg} \pm c_2c_\alpha} \tag{3.15}$$

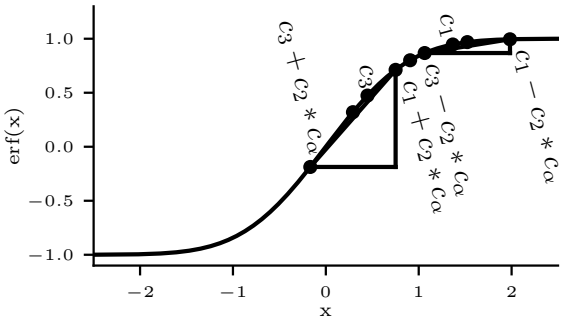


Figure 3.1. Erf Function with $\Delta = 10, \sigma_1 = 2$ and $\sigma_2 = 5$

From Equation 3.14, Equation 3.15 and the definition of the *erf* function,

$$\begin{aligned} \frac{\exp\left(-(c_{avg} - c_2 c_\alpha)^2\right)}{\exp\left(-(c_{avg} + c_2 c_\alpha)^2\right)} - \frac{\alpha}{1-\alpha} &< 0 \\ \therefore \mathrm{e}^{4c_{avg}c_2c_\alpha} - \mathrm{e}^{c_\alpha} &< 0 \\ \therefore (4c_{avg}c_2 - 1)c_\alpha &< 0 \end{aligned} \tag{3.16}$$

With $c_\alpha < 0$, Equation 3.16 holds only when $4c_{avg}c_2 > 1$. Since $\sigma_{\mathrm{i}} < \sigma_{\mathrm{i}}^{\theta}$, it implies that $4c_{avg}c_2 > 1$. Therefore, Equation 3.14 holds and we conclude that $SR_p > SR_u$ when $\alpha < 0.5$ Similarly, we can show that $SR_p < SR_u$ when $\alpha > 0.5$. (When $\alpha = 0.5$, the selection rates are expected to be equal. To summarize, an NB classifier trained on FBD $\mathcal{D}$ with disparate group-wise optimal model accuracy (due to feature disparity) can result in disparate selection rates among the groups.

3.1.3 Less Separable Unprivileged Group with Resource Constraints

In real-world scenarios, available resources are often not aligned with the pool of qualified candidates. The scarcity of resources compels selecting a predetermined portion of qualified candidates. In contrast, to increase utilization, surplus resources are often distributed with relaxed qualification considerations. The following section discusses the implications of resource constraints on model performance.

A typical approach to enforce resource constraints is decision boundary modification. Resource constraints can also be satisfied by ranking individuals with prediction probabilities from probabilistic classifiers and picking pre-established portion p_{res} of top-ranked ones. While the former method approximates the constraint, the latter guarantees exact resource allocation. Let the decision boundary that best approximates the resource constraints be,

$$\mathbb{P}(\hat{y}^+ \mid \mathbf{x}) > c_{res}\mathbb{P}(\hat{y}^- \mid \mathbf{x}) \tag{3.17}$$

Here, c_{res} is a constant that controls the rate of positive predictions from the model. Using the constrained decision boundary, we modify Equation 3.6, Equation 3.9, and Equation 3.10 as,

$$\frac{(x_1 - \mu_1^{avg})\Delta}{2(\sigma_1^{\theta})^2} + \frac{(x_2 - \mu_2^{avg})\Delta}{2(\sigma_2^{\theta})^2} + c_\alpha \geq \log c_{res} \tag{3.18}$$

$$\mathbb{P}(\hat{y}^+ \mid y^+, s) = \frac{1}{2} + \frac{1}{2}erf\left[c_1\mathbb{1}_{s=p} + c_3\mathbb{1}_{s=u} + c_2 - \frac{\log c_{res}}{c_{denom}}\right] \tag{3.19}$$

$$\mathbb{P}(\hat{y}^+ \mid y^-, s) = \frac{1}{2} - \frac{1}{2}erf\left[c_1\mathbb{1}_{s=p} + c_3\mathbb{1}_{s=u} - c_2 + \frac{\log c_{res}}{c_{denom}}\right] \tag{3.20}$$

Equation 3.19 is similar to 3.9 except the term inside erf function is decreased by $\frac{\log c_{res}}{c_{denom}}$. Since c_{res} is independent of s, the relationship in Equations 3.12 and 3.13 still hold.

Insufficient Resources

When resources are scarce, i.e., $c_{res} > 1$, from 3.19 and 3.20, we conclude that TPR_s and FPR_s are reduced from the unconstrained model. Since the unprivileged group is less separable than the privileged, qualified privileged individuals will tend to receive higher positive prediction probabilities than the qualified unprivileged ones. Therefore, qualified privileged members will be selected more than qualified unprivileged.

Surplus Resources

With surplus resources, i.e., $c_{res} < 1$, from 3.19 and 3.20, we can conclude that TPR_s and FPR_s are increased from the unconstrained model. In contrast to the limited resource scenario, unqualified unprivileged group members are expected to receive higher positive prediction probabilities than unqualified privileged members. It means that among similarly unqualified individuals from both groups, the unprivileged are more likely to receive a positive prediction. As a result, with surplus resource availability, false positives are more common in the unprivileged group. Assuming that unqualified candidates receiving surplus resources is detrimental (e.g., college admissions going to individuals who will drop out), the unprivileged group is again harmed.

3.2 Experimental Results

We now show that the theoretical results are reflected in common classifiers. First, we investigate the behavior of naive Bayesian classifiers on equally separable privileged and unprivileged groups. Then, we experiment with an unprivileged group that is less separable than the privileged group. Finally, we show results for a variety of common classifiers. The source code of the experiment is available at https://github.com/rakinhaider/Inherent-AI-Bias.

3.2.1 Fair Balanced Datasets

Real-world data typically contain historical unfairness. Instead, we first conduct experiments on synthetic fair balanced datasets (SFBD) so that we can analyze introduced bias in the absence of label bias. Later, we analyze feature disparity in conjunction with label bias using real-world datasets. Following Section 3.1, we generate r group-specific predictive attributes for each group and $n - 2r$ non-predictive group-independent attributes. The predictive attributes are sampled conditioned on class labels whereas the $n - r$ non-predictive attributes are unconditionally sampled from a random normal distribution. For example, in college admission prediction, privileged samples are generated by sampling x_{TS}^{p+} and x_{TS}^{p-}

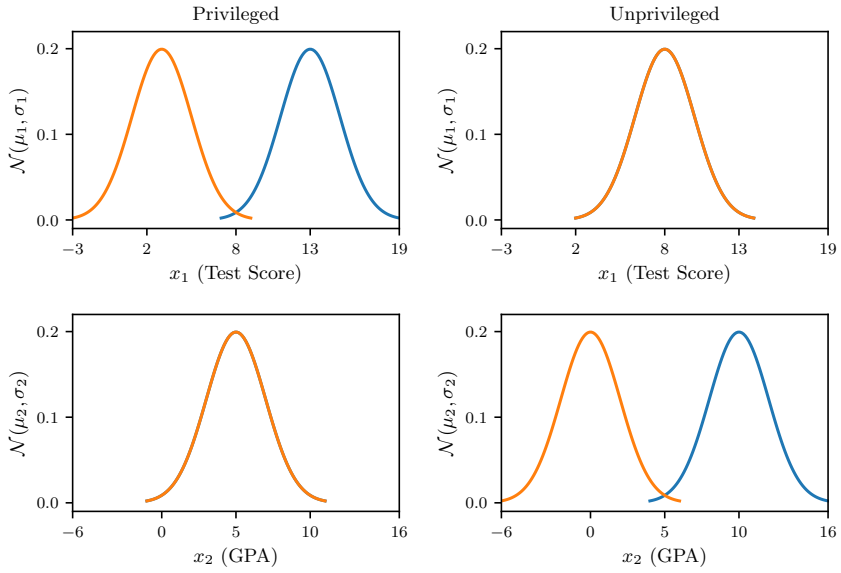


Figure 3.2. Example of feature distributions when the unprivileged group is equally separable to the privileged group, but optimal models are different. The blue (dotted) and orange (dashed) curves represent positive and negative sample distributions. The brown (dashdot) curve indicates two distributions (positive and negative) are overlapping.

from two distinct distributions. In contrast, x^{p+}_{GPA} and x^{p+}_{GPA} are sampled from the same distribution. In this work, each *SFBD* contains 10000 samples with either 2 ($r = 1, n = 2$) or 10 ($r = 3, n = 10$) attributes. We generate one SFBD for each $\alpha \in \{0.25, 0.5, 0.75\}$. By default, we use $\mu_{\mathrm{i}}^{p+} = 13, \mu_{\mathrm{i}}^{u+} = 10$ and $\Delta = 10$.

3.2.2 Equally Separable Unprivileged Group

We first let the unprivileged group be equally separable to the privileged group. For $n = 2$ and $n = 10$, $\sigma_{\mathrm{i}} = \sigma_{\mathrm{j}} = 2$ and $\sigma_{\mathrm{i}} = \sigma_{\mathrm{j}} = 4$, respectively such that $\mathrm{i} \in \{1, \ldots, r\}$ and $\mathrm{j} \in \{r + 1, \ldots 2r\}$. In the college admission scenario, this corresponds to $\sigma_{TS} = \sigma_{GPA}$. Figure 3.2 shows a distribution of the attributes where the groups are equally separable. Here, x_1^{p+} and x_1^{p-} are sampled from different distributions while $x_1^{u\pm}$ are sampled form the same distribution. A similar but opposite trend is observed for x_2. Since $\sigma_1 = \sigma_2$,

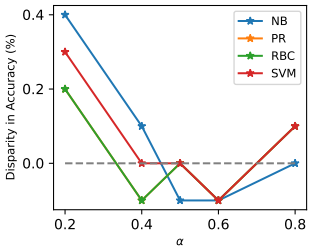
Table 3.1. Model performances on an equally separable unprivileged group with 2 attributes

Classifier	α	$SP(ACC;\theta)$	$SP(\theta)$	$SP(FPR;\theta)$
NB	0.2	0.40	0.00	0.00
	0.4	0.10	-0.20	-0.40
	0.5	-0.10	0.30	0.50
	0.6	-0.10	0.20	0.70
	0.8	0.00	-0.50	-2.10

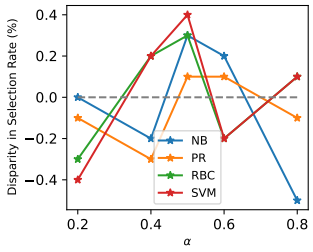
Table 3.2. Model performances on an equally separable unprivileged group with 10 attributes

Classifier	α	$SP(ACC;\theta)$	$SP(\theta)$	$SP(FPR;\theta)$
NB	0.2	-0.20	-0.20	-0.20
	0.4	-0.10	0.10	-0.20
	0.5	0.20	-0.80	-1.50
	0.6	0.10	-1.10	-1.40
	0.8	0.20	-0.30	-0.10

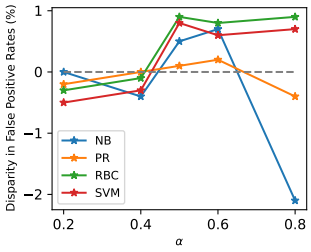
x_1 is equally predictive for the privileged as x_2 is for the unprivileged. Tables 3.1 and 3.2 contain the SR_s and FPR_s of joint-accuracy NB classifiers trained on the synthetic fair and balanced datasets with 2 and 10 attributes respectively. The group-wise optimal model accuracy is denoted by AC_s. We observe similar AC_s, SR_s and FPR_s for each s. The tables show in the absence of feature disparity model performance demonstrated minimal disparity between the sensitive groups. We further experiment with classifiers beyond NB classifier and report the disparities in Figures 3.3 and 3.4. We observe that the disparity between group-wise accuracy, selection rate, and false positive rates are less than 0.5%, 0.5%, and 2% respectively from all the classifiers. It further supports that the baseline datasets contain no label bias or are effectively de-biased such that even fairness-unaware algorithms yield approximately perfectly fair classifiers.



(a) Disparities in Group-wise Accuracy.

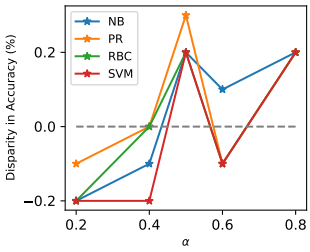


(b) Disparities in Group-wise Selection Rates.

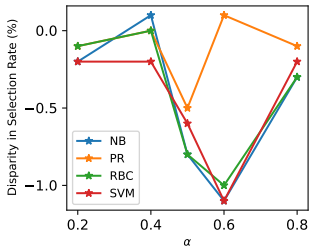


(c) Disparities in Group-wise False Positive Rates.

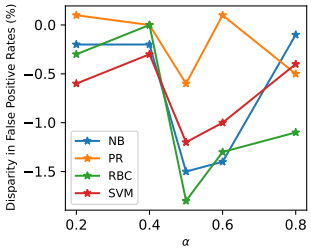
Figure 3.3. Equally separated groups with 2 attributes.



(a) Disparities in Group-wise Accuracy.



(b) Disparities in Group-wise Selection Rates.



(c) Disparities in Group-wise False Positive Rates.

Figure 3.4. Equally separated groups with 10 attributes.

3.2.3 Less Separable Unprivileged Group

We now show classifier performances when the dataset contains feature disparity. Note that we observed little to no outcome bias from classifiers trained on datasets without any feature disparity. Keeping everything else unchanged, we introduce feature disparity in the dataset. Since, keeping everything else unchanged, feature disparity is the only type of disparity added to the dataset, any outcome bias observed from the classifiers can be attributed to feature disparity. Following the construction discussed in Section 3.1.2, we introduce feature disparity by making the $\sigma_p < \sigma_u$ which eventually makes the unprivileged group less separable and difficult to classify. We examine the effects of this feature disparity on outcome fairness and its interaction with resource constraints.

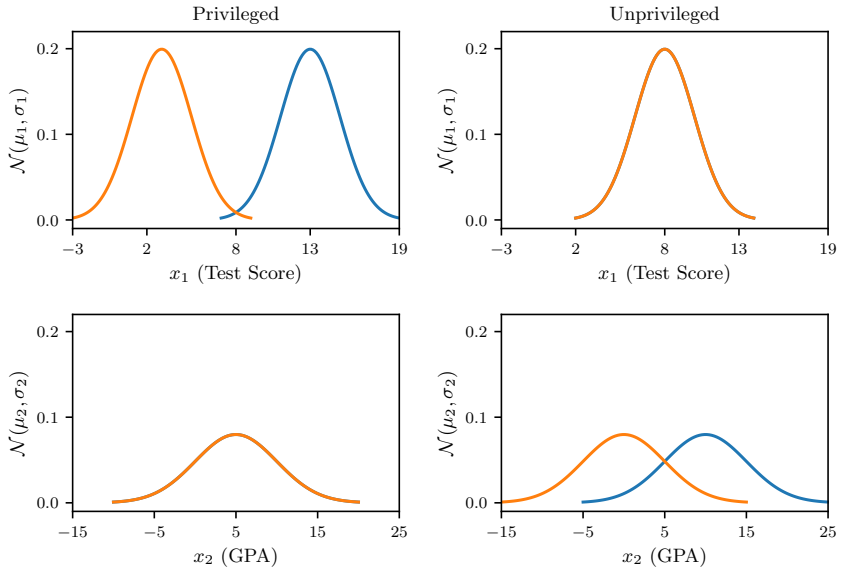


Figure 3.5. Example of feature distributions when the unprivileged group is less separable than the privileged group. The plot colors and patterns convey similar meanings as described in Figure 3.2.

Without Resource Constraints

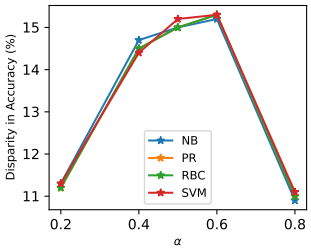
The less separable unprivileged group has a higher standard deviation in predictive attributes. The SFBD with $n = 2$ and less separable unprivileged group has $\sigma_{TS} = \sigma_{\mathrm{i}} = 2$ and $\sigma_{GPA} = \sigma_{\mathrm{j}} = 5$. For $n = 10$, we use $\sigma_{\mathrm{i}} = 4$ and $\sigma_{\mathrm{j}} = 7$. Here, $\mathrm{i} \in \{1, \ldots, r\}$ and $\mathrm{j} \in \{r+1, \ldots, 2r\}$. Figure 3.5 shows a distribution of x_1 (TS) and x_2 (GPA) for the privileged and unprivileged group. This is similar to Figure 3.2 except that $\sigma_1 < \sigma_2$. Clearly, x_1 is a better predictor for the privileged group than x_2 is for the unprivileged. We observe that all the classifiers achieve lower unprivileged model accuracy on the less separable unprivileged group in Table 3.3. We observe a significant disparity in selection rates of the NB classifier for $\alpha \neq 0.5$. The selection rate is higher for the privileged group when $\alpha < 0.5$, but favors the unprivileged group for $\alpha > 0.5$. This confirms the theoretical result of Section 3.1.2. The higher SR_u at $\alpha > 0.5$ is largely due to the high FPR_u. In the running example of college admission predictions, the trend of disparity in Table 3.3 indicates that highly se-

Table 3.3. Model performances for less separable unprivileged group

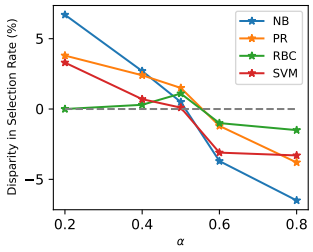
Classifier	α	$n = 2$			n=10		
		$SP(ACC;\theta)$	$SP(\theta)$	$SP(FPR;\theta)$	$SP(ACC;\theta)$	$SP(\theta)$	$SP(FPR;\theta)$
NB	0.2	11.30	6.70	-4.00	6.60	1.90	-6.80
	0.4	14.70	2.70	-16.20	8.60	0.60	-15.80
	0.5	15.00	0.50	-22.90	9.60	-1.20	-21.80
	0.6	15.20	-3.70	-32.60	8.90	-1.80	-25.70
	0.8	10.90	-6.50	-48.80	6.90	-1.60	-37.50
SVM	0.2	11.30	3.30	-4.30	6.70	1.40	-3.30
	0.4	14.40	0.70	-10.80	8.80	1.10	-6.80
	0.5	15.20	0.10	-14.50	9.80	-0.60	-11.30
	0.6	15.30	-3.10	-21.30	9.00	-2.70	-15.10
	0.8	11.10	-3.30	-32.10	7.30	-1.50	-19.90
PR	0.2	11.20	3.80	-4.90	6.70	1.70	-2.90
	0.4	14.50	2.40	-9.90	8.50	1.10	-6.60
	0.5	15.00	1.50	-13.50	9.50	0.00	-10.10
	0.6	15.30	-1.20	-20.60	8.80	-0.30	-12.00
	0.8	11.00	-3.80	-36.90	6.80	-2.00	-23.10
RBC	0.2	11.20	0.00	-8.20	6.60	0.10	-4.90
	0.4	14.50	0.30	-14.00	8.50	0.20	-8.90
	0.5	15.00	1.10	-17.40	9.60	-1.10	-13.10
	0.6	15.30	-1.00	-23.40	8.80	-1.70	-16.60
	0.8	11.00	-1.50	-34.30	6.80	-0.30	-20.00

lective prestigious colleges offer more admissions to privileged students. With low selectivity, typically found in predatory colleges, the unprivileged individuals are admitted at a 9.44% higher rate, but a significant portion of acceptance offers are received by under-qualified students, with the risk of higher drop-out rate and additional student loan burden among the unprivileged group.

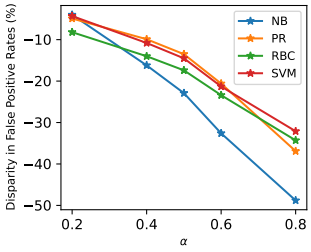
A generalized experiment, with more (predictive and non-predictive) features dampens the disparity in selection rate since increasing the number of predictive features increases model accuracy. However, false positive rate disparity still exists and increases with α. Table 3.3reports the disparity in performance for $n = 10$ features as well. Table 3.3 contains the disparities found in the prediction rates of four classifiers including two fairness-aware



(a) Disparities in Group-wise Accuracy.

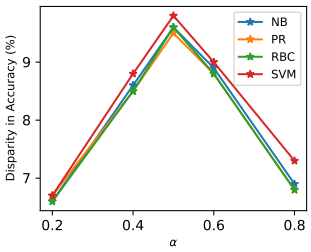


(b) Disparities in Group-wise Selection Rates.

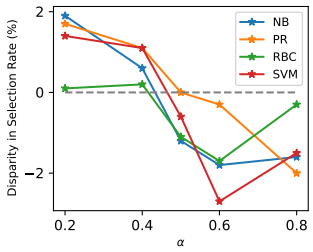


(c) Disparities in Group-wise False Positive Rates.

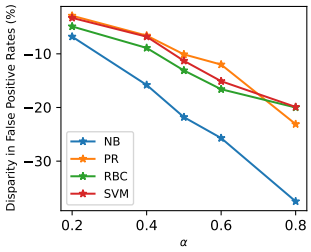
Figure 3.6. Disparities in model performances on a less separable unprivileged group with 2 features.



(a) Disparities in Group-wise Accuracy.



(b) Disparities in Group-wise Selection Rates.



(c) Disparities in Group-wise False Positive Rates.

Figure 3.7. Disparities in model performances on a less separable unprivileged group with 10 features.

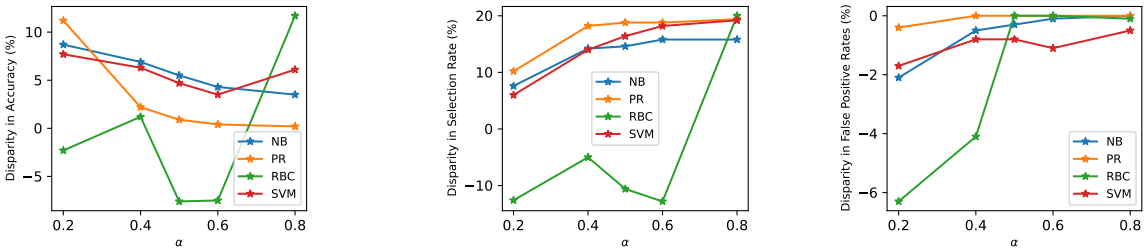
algorithms. Traditional models such as SVM demonstrate similar behavior to Naive Bayes. Notably, we also experimented with decision trees which demonstrated evidence of introduced bias but didn't follow a similar trend with α as the other classifiers. Prejudice Remover (PR) [13] and Reduction based classifier (RBC) [26] appear inherently less biased than traditional models. Both reduced the selection rate disparity as they are specifically designed to do so. However, they still suffer with respect to false positive rates.

The biased outcomes shown in Table 3.3, vs. the unbiased outcomes in Tables 3.1 and 3.2, corroborates our claim that poor feature selection is sufficient to cause outcome bias in otherwise fair and balanced datasets.

With Resource Constraints

The problem is exacerbated when resources are constrained. Since Naive Bayes generates a probability-based score, it is easy to adapt to decisions where the number of positive (or negative) outcomes is limited; we now show results where resources are scarce ($p_{res} = 10\%$) and surplus ($p_{res} = 90\%$). Again, we show this where the true need for the resource varies between $\alpha \leq 0.5$, $\alpha = 0.5$, and $\alpha \geq 0.5$. Table 3.4 and Figures 3.8 and 3.9 show that resource constraints add further detrimental effect on the decision outcomes.

Table 3.4 and Figure 3.8 show that with insufficient resources, privileged individuals are selected 3 times more than unprivileged ones. Moreover, the ratio of SR_u to SR_p increases as α increases. In college admission, this corresponds to selecting far more privileged group members at highly selective prestigious colleges when the number of applicants is far higher than the available positions. Unprivileged intake suffers as the competition for positions increases.



(a) Disparities in Group-wise Accuracy.

(b) Disparities in Group-wise Selection Rates.

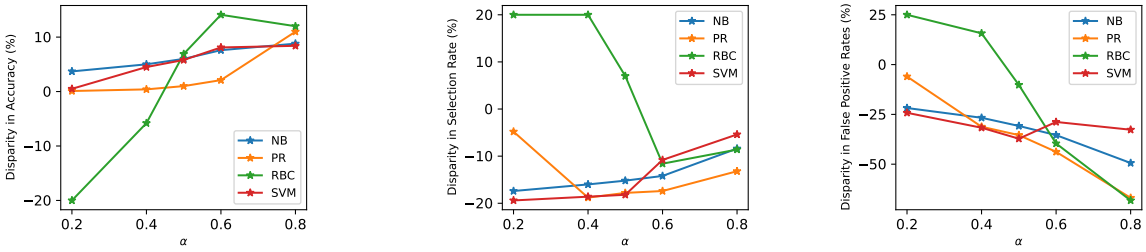
(c) Disparities in Group-wise False Positive Rates.

Figure 3.8. Disparities in model performances on a less separable unprivileged group with 2 features and insufficient resources.

Table 3.4 and Figure 3.8 indicate that if resources are surplus, unprivileged students receive higher acceptance. However, a higher SR_u is accompanied by a high FPR_u. Therefore, the resources are received by less deserving individuals. We can draw parallels between such model behavior and predatory colleges, that lure as many students as possible irrespective of their qualifications. As a result, many less-qualified candidates eventually drop out, having acquired a student loan burden.

Table 3.4. Model performances for a less separable unprivileged group with 2 features and resource constraints.

Classifier	α	Insufficient Resources			Surplus Resources		
		$SP(ACC;\theta)$	$SP(\theta)$	$SP(FPR;\theta)$	$SP(ACC;\theta)$	$SP(\theta)$	$SP(FPR;\theta)$
NB	0.2	8.70	7.60	-2.10	3.70	-17.40	-21.80
	0.4	6.90	14.20	-0.50	5.00	-16.00	-26.70
	0.5	5.50	14.60	-0.30	6.00	-15.20	-30.80
	0.6	4.30	15.80	-0.10	7.60	-14.20	-35.40
	0.8	3.50	15.80	0.00	8.80	-8.40	-49.40
SVM	0.2	7.70	6.00	-1.70	0.50	-19.40	-24.20
	0.4	6.30	14.00	-0.80	4.50	-18.60	-31.70
	0.5	4.70	16.40	-0.80	5.80	-18.20	-37.20
	0.6	3.50	18.20	-1.10	8.10	-10.80	-28.80
	0.8	6.10	19.20	-0.50	8.40	-5.40	-32.70
PR	0.2	11.20	10.20	-0.40	0.10	-4.80	-6.00
	0.4	2.20	18.20	0.00	0.40	-18.80	-31.20
	0.5	0.90	18.80	0.00	1.00	-17.80	-35.40
	0.6	0.40	18.80	0.00	2.10	-17.40	-43.80
	0.8	0.20	19.40	0.00	11.00	-13.20	-66.90
RBC	0.2	-2.30	-12.60	-6.30	-20.00	20.00	25.00
	0.4	1.20	-5.00	-4.10	-5.80	20.00	15.70
	0.5	-7.60	-10.60	0.00	6.90	7.00	-10.20
	0.6	-7.50	-12.80	0.00	14.10	-11.60	-39.60
	0.8	11.70	20.00	-0.10	12.00	-8.60	-68.20



(a) Disparities in Group-wise Accuracy.

(b) Disparities in Group-wise Selection Rates.

(c) Disparities in Group-wise False Positive Rates.

Figure 3.9. Disparities in model performances on a less separable unprivileged group with 2 features and surplus resources.

Table 3.5. Model performances on COMPAS original and de-biased

Dataset	FPR_c	FNR_c	FPR_a	FNR_a
COMPAS original	62.58	14.74	26.01	42.73
COMPAS de-biased	62.05	18.42	37.79	31.12

While other models are not directly suited to resource-constrained predictions, we adapted them and obtained similar results when group-wise classifier accuracy is different.

3.2.4 Fairly Sampled Balanced COMPAS Dataset

We also evaluate using the real-world COMPAS dataset [51]. The dataset is partitioned into independent training (80%) and test (20%) sets. The privileged (also the minority) in the COMPAS dataset is the *Caucasian* group, and the unprivileged is the *African-American* group. The favored class is *no recidivism*. Approximately 61% of the Caucasians are in the favored class, while 48% of the African Americans are in the favored class. To balance the base positive rates between the two groups, we use a de-biasing algorithm [11] that generates synthetic data to augment the unprivileged but favored group. This method generates synthetic data from the *African-American* instances that have favored labels in the training data, using the Synthetic Minority Oversampling Technique (SMOTE) algorithm [52]. We

Table 3.6. Model performances for COMPAS SFBD de-biased by inflating privileged unfavored class

α	AC_p	AC_u	SR_p	SR_u	FPR_p	FPR_u
0.25	73.3	74.0	21.1	10.9	15.5	06.5
0.50	61.8	61.6	62.4	41.6	52.1	30.3
0.75	75.9	72.2	89.0	81.9	76.5	68.7

Table 3.7. Model performances for COMPAS SFBD de-biased by inflating unprivileged favored class

α	AC_p	AC_u	SR_p	SR_u	FPR_p	FPR_u
0.25	74.6	72.5	16.2	05.4	12.0	06.0
0.50	61.2	61.2	75.2	51.1	67.6	34.0
0.75	74.7	73.6	91.7	78.3	83.3	68.7

first randomly select an unprivileged (*African American*) instance $(\mathbf{x}, u, y^+)$, then select k random unprivileged neighbors of $\mathbf{x}$, from the favored class. We randomly choose a neighbor $\mathbf{x}'$ and generate a synthetic data point by choosing a point between $\mathbf{x}$ and $\mathbf{x}'$ in the feature space. This process is repeated until the groups have similar outcomes in the training set.

Table 3.5 shows naive Bayesian prediction rates for Caucasian and non-Caucasians before and after de-biasing. Oversampling the favored class in African Americans increases their false positive rate (FPR_a) significantly.

We perform a 70% train and 30% test split on each ($\alpha \in \{0.25, 0.5, 0.75\}$) resampled de-biased COMPAS dataset. Since the preprocessed COMPAS dataset [53] contains only binary variables, we use a Bernoulli distribution for the attributes. Although the de-biased dataset has little disparity in group-wise optimal model accuracy, we still observe significant disparity in group-wise selection rates. This can be ascribed to both inherent bias and data bias. Consistent with the previous sections, the selection rate grows more favorable to the unprivileged group as α increases.

3.3 Systemic Bias: Why We Expect These Outcomes

It could be argued that this is purely a hypothetical issue, that in an otherwise fair world, we would not see accuracy differences between optimal models for privileged and unprivileged groups. We suggest that this is instead a common and likely form of systemic bias. ML systems are generally designed by the privileged group, and the features considered are the ones that seem natural to that group. The (privileged) developers are unaware of or do not consider features that are more appropriate for unprivileged groups, leading to inherently more accurate systems for the privileged group.

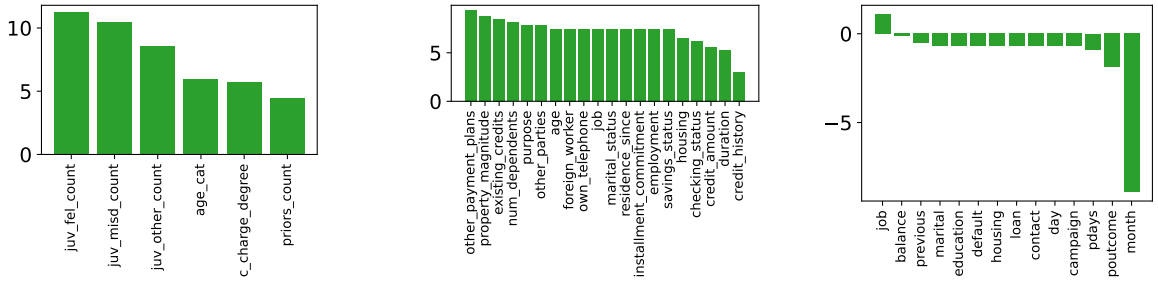
We can see two well-known examples of this from the domain of medical research. One is with heart attack symptoms. Western medicine has long taught that chest pain or discomfort is an early sign of a heart attack. However, women are more likely than men to experience other symptoms, particularly shortness of breath, nausea/vomiting, and back or jaw pain [54]. This is an example where the privileged group (men, who long dominated medical research) used features that were effective for their group, leaving the unprivileged group less likely to properly self-diagnose heart attack (lower accuracy.)

A similar situation, although not as tied to feature selection, arose in cancer research. Cancer research long focused on lung cancer, which at the time disproportionately impacted males, and at a younger age than many other cancers (average age of diagnosis was 66 in 1975-1999 [55]). Public outcry over this gender disparity led to increased investment in breast cancer, which became overfunded relative to other cancers in terms of years of life lost [56]. While not directly a machine learning issue, we see that cancer investment followed what impacted the privileged group (initially males in senior leadership; followed by public outcry from a large voting block). They exemplify situations where features used are obvious to (and work well for) the privileged, eventually harming the unprivileged.

In this regard, we analyzed feature sets of three popular datasets, i.e., COMPAS, German Credit [51], and Bank Marketing [57]. The accuracy obtained while predicting the class label using a single feature f is defined as the predictive power of the feature PP_f. We train two group-wise single-feature models and determine the group-wise predictive powers of f, PP_f^p, and PP_f^u. For better estimates, we performed 10-fold cross-validation and picked the average

Table 3.8. Predictive powers of each feature in COMPAS dataset.

Feature	Avg. PP_p	Avg. PP_u	PPD
juv_fel_count	61.25	50.02	11.23
juv_misd_count	62.15	51.68	10.47
juv_other_count	61.82	53.26	8.56
age_cat	60.91	54.96	5.95
c_charge_degree	60.91	55.18	5.73
priors_count	64.91	60.41	4.50



(a) PPDs in COMPAS dataset.

(b) PPDs in German dataset.

(c) PPDs in Bank dataset.

Figure 3.10. Predictive Power Difference (PPD) of each feature in COMPAS, German, and Bank dataset.

PP_f^s. The predictive power difference of feature f is $PPD_f = PP_f^p - PP_f^u$. The maximum absolute values of PPD_f in the three datasets are 11.2%, 9.4%, and 8.9%, respectively. It suggests that disproportionately predictive features are commonplace in real-world machine learning datasets. Table 3.8 and Figure 3.10a show the PPDs for features of the COMPAS dataset. Similarly, the PPD with respect to each feature in the German and Bank datasets can be found in Figures 3.10b and 3.10c respectively.

4. INTRODUCED BIAS THROUGH MISSING VALUES

Missing value refers to the situation where one or more features of a sample are missing. A sample with missing values is called an incomplete sample and a complete sample is otherwise. Missing value is a common phenomenon due to the uncertain and noisy data collection mechanisms. For example, if the data is collected through questionnaires, participants are often not obligated to provide all the information. Moreover, some features inherently contain missing values. For example, *middle name* is an attribute that is likely to be missing for a large portion of the US population. Similarly, a dataset that tracks the addresses of individuals will have missing values for homeless entities. Finally, features can be mutually exclusive or conditionally optional. A realistic example of the former is foreign individuals often are not eligible to have a social security number (SSN). Likewise, college admission applicants may be required to have at least one TOEFL and IELTS score if not both. Such requirements lead to conditionally optional attributes with missing values (i.e., if a TOEFL score exists IELTS score can be missing).

According to [58], missing values can follow three primary patterns, namely, missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Missing completely at random occurs when missingness is independent of observed and unobserved data. When missingness depends on observed data it is referred to as missing at random. Finally, MNAR is defined as the type of missingness that also depends on unobserved data.

Traditional ML expects complete training samples, as do state-of-the-art fairness enhancing interventions (e.g., [6, 13–15, 24, 26].) Unfortunately, missing data is common; about half of the datasets in the prominent UCI dataset repository contain incomplete samples with one or more missing values [33]. Instead of reducing the sample size by ignoring incomplete samples altogether, data engineers have developed several techniques to interpolate the missing values using individual and population level statistics; commonly known as imputations [59]. Recently several studies investigated the fairness implications of missing value handling mechanisms. Among them Martinez et al. [33] argued in favor of imputations instead of dropping since incomplete cases included many fair decision samples. However,

[60] and [61] suggested that imputation can both amplify or attenuate existing label bias. Machine learning fairness is also being studied through the lens of data management which frequently includes missing value repairing [62, 63]. In short, the impact of missing values on model fairness was studied under the influence of existing label bias. Previous studies on the interactions of fairness and missing values [33, 60–62, 64] failed to investigate whether missing value disparity *itself* is sufficient to introduce bias.

In this chapter, we show that missing value disparity is sufficient to cause bias in model behavior. We start with a constructive proof of missing value disparity induced bias involving simple mechanisms, showing why missing value handling leads to outcome (prediction) bias. We prove by construction that for a Bayes-optimal classifier, missing value disparity handled by dropping or mean imputation is sufficient to induce outcome bias when the group-wise optimal models are different. The use of a Bayes-optimal classifier implies that the findings are also applicable to accuracy-optimal machine learning approaches. We further show that disparity in missing value rates exacerbates biased outcomes given imbalanced group sizes. Empirical validation of induced bias from missing data in more complex situations, including real-world data that may contain several sources of bias, more complex missing value imputation methods (k-NN imputation [34], MICE [35], matrix completion [36]), and both traditional (SVM, Neural Network) and state-of-the-art fairness-aware classifiers is also reported.

The constructive proof formulates an instance of missing value disparity induced bias in model outcome to prove that missing value disparity alone can induce bias in model outcome. Experimental results on a synthetic fair and balanced dataset further validate the theoretical analysis. Even in cases where the changes in prediction probabilities were too low to alter the predicted labels, we observed that missing value disparity drives the classifiers to disproportionately promote the relative rankings of one group and demote the other. Although we mainly focus on missing values in the training data, we found that induced bias became even more prominent when the test data included missing values. This is because missing values in test samples inject disparity in the feature distributions of the test samples; closely tying our results with the ones in [65] and [66].

Having shown that missing value disparity is sufficient to induce bias, we further investigate the extent of this induced bias when coupled with other sources of bias. We hypothesize that missing data disparity will exacerbate existing outcome bias in situations where there are multiple sources of bias. We validate this with experiments on real-world datasets which commonly contain selection bias and label bias among other types of disparities. We observed missing value disparity induced bias in model behavior across different types of classifiers, multiple causal MAR mechanisms for missing values, and several imputation mechanisms. We conclude that induced bias through missing value disparity is a common occurrence and threatens the fairness of machine learning models. It calls for attention towards re-thinking data collection and cleaning algorithms.

4.1 Fairness with Missing Data

In this section, we discuss common types of missingness, state-of-the-art imputation techniques, and related works that study fairness implications under the lens of missing value imputations.

4.1.1 Types of Missingness

Missing values are pervasive in real-world datasets. About 45% of datasets in the UCI repository contain missing values [33]. Partial completion of questionnaires, missing by design, and item non-response are key reasons[33]. Fatigue or technical issues cause partial completion. Similarly, features such as middle name, current address, SSN, higher education, etc., contain missing values by design. Finally, participants avoid responding to questions regarding criminal offenses, medical history, etc., due to privacy and legal concerns.

Rather than tracing the cause of missingness, a more statistical categorization is proposed in [67]. We extend this formulation to incorporate between-group disparities. Here, we consider a binary classification task with "privileged" (p) and "unprivileged" (u) group distinction. Let $\mathcal{D} = \{\mathbf{X}, \mathbf{S}, \mathbf{Y}\}$ is a dataset with N samples, features $\mathbf{X} = [\mathbf{x_1}, \ldots \mathbf{x_N}]^T \in \mathbb{R}^{N \times d}$, class labels $\mathbf{Y} \in \{-, +\}^N$ and sensitive attribute $\mathbf{S} \in \{p, u\}^N$. For simplicity, we limit our discussion to a single binary class label and a single sensitive attribute. However, our analysis

can be easily extended to multiple sensitive attributes with the help of similar group-indicator functions $\mathcal{G}$ as defined in [24, 25] which map the intersections of different sensitive sub-groups to their levels of privilege. Without loss of generality, we assume that $y = +$ is the favorable (or positive) class label. The model, θ predicts $\hat{\mathbf{Y}} \in \{-, +\}^N$. Let $\mathbf{R} = [\mathbf{r}_1, \ldots, \mathbf{r}_n]^T \in \mathbb{1}^{N \times d}$ be an indicator matrix where each element $r_{\mathrm{i,j}}$ indicates whether the value of $x_{\mathrm{i,j}}$ is missing. Here, $\mathbf{r}_{\mathrm{i}} = \bar{0}$ and $\mathbf{R}_{*,\mathrm{j}} = \bar{0}^T$ indicates the i^{th} sample and the j^{th} feature are a complete case and a complete feature respectively. Finally, $\mathbf{x}, \mathbf{r}, s, y, \hat{y}$, are random variables indicating the feature vector, missing value indicator vector, sensitive attribute, class label, and predicted label. Here, $s = p$ and $s = u$ indicate the privileged and the unprivileged group. For a simpler notation, we use $s = p$ and p interchangeably. Similarly, instead of $s = u$, we write u. With these definitions, missing values can be categorized as,

Missing Completely at Random (MCAR)

In MCAR, missingness is independent of observed and unobserved data. That means, $\mathbf{R} \not\perp\!\!\!\perp \mathcal{D}$.

Missing at Random (MAR)

If missingness depends on observed data, i.e., $\mathbf{R} \perp\!\!\!\perp \mathcal{D}_{r_{\mathrm{i,j}}=0}$, then it is called missing at random. MAR is of particular interest in our study since missing value disparity can be defined as $\mathbf{r} \perp\!\!\!\perp s$, i.e., the rate of missingness is dependent on group membership.

Missing Not at Random (MNAR)

Let $\mathcal{D}_{(0)}, \mathcal{D}_{(1)} \subset \mathcal{D}$ are the sets of observed and unobserved data respectively. Then, MNAR is formally defined as, $\mathbf{R} \perp\!\!\!\perp \mathcal{D}_{(0)} \cup \mathcal{D}_{(1)} = \mathcal{D}$. For example, individuals with prior arrests may prefer to omit this information. In contrast, people with no priors probably will be eager to report it. That means the missing value itself controls the likelihood of its missingness.

While previous studies have empirically shown fairness is impacted by missing data imputation in real-world data [60, 62], they do not look at how and why missing data and

imputation impact fairness. In our study, missing value disparity is defined as differences in the rates of missing values based on observed sensitive attributes (MAR), a known real-world phenomenon. We demonstrate that outcome bias appears or increases specifically from missing data handling methods on MAR missing values.

4.1.2 Missing Value Handling Mechanisms

Missing value handling is typically a pre-processing step. We discuss several of these techniques below.

Deletion

If the dataset is sufficiently large or incomplete cases are rare, missing values can be handled by simply discarding the incomplete cases from the training set. Case deletion leads to a smaller number of samples and a skewed dataset as shown in [67]. Pairwise deletion uses a correlation matrix to remove missing values appearing in pairwise analysis. Instead of dropping samples, uncorrelated features with many missing values could also be dropped, thus preserving the samples.

Hot-deck imputation

This strategy matches the incomplete cases with other complete cases based on one or more key variables and picks a donor pool from the matched cases [68]. The missing value is replaced with the aggregate value of one or more members of the donor pool.

Simple imputation

Simple imputation replaces the missing values with a feature-wise statistic such as mean or median for numeric attributes and mode for categorical features. Although intuitive, it often leads to high error imputations.

Iterative Imputations

Iterative imputations gradually improve their predictions of missing values by training a model on the currently available samples to predict the missing values at each iteration until convergence. They often use simple imputation as the initial estimate of missing values. Among the popular iterative methods MICE [35, 69] and MissForest [70] use linear regression and random forest respectively.

Matrix Completion

Missing value imputation can be modeled as matrix completion of the feature matrix X. The most prominent such algorithm, *Softimpute*, computes a low-rank approximation of X with matching observed entries [36].

ML Based Imputation

Imputation can also be done with predicted values from ML models such as k-NN [34], SVM [71], expectation-maximization [72], etc.

Our theoretical analysis covers dropping and mean-imputation approaches to handling missing values. The experimental study includes techniques too complex for analytical treatment, specifically k-NN imputation, MICE, and Softimpute.

4.1.3 Related Works

Recently, Zhang et al. [64] estimated fairness on complete cases from an incomplete data domain. [60] empirically showed that imputation with an imbalance in missingness exacerbates the bias in the model outcome. Finally, [62] reported that auto-cleaning, performed to alleviate the poor quality of training data (including both label bias and missing values), also led to worse fairness outcomes. Although the interaction of missing value and label bias received attention lately [60, 62, 64, 73], none investigated bias induced from missing value disparity. As a result, in this work, we focus on missing value disparity induced bias in model outcomes.

4.2 Formal Analysis

We prove that a classifier optimizing for accuracy is expected to produce biased model outcomes given disproportionate group-wise rates of missing values. We assume that the optimal models for the privileged and unprivileged groups are different, and a system with "fairness by un-awareness", i.e., the sensitive attributes are not used by the model. We focus on missing at random (MAR), i.e., the probability of missing data depends on the observed sensitive attribute, with considerations of missing not at random (MNAR) when appropriate. We consider the following MAR causal mechanisms of missing value disparity:

- Disparity in complete case rates, i.e.,

$$\mathbb{P}(\mathbf{r} = \bar{0} \mid p) \neq \mathbb{P}(\mathbf{r} = \bar{0} \mid u)$$

- Disparity in the likelihood of missing feature of a sample,

$$\mathbb{P}(r_{\mathrm{j}} \mid p) \neq \mathbb{P}(r_{\mathrm{j}} \mid u) \text{ where } \mathrm{j} \in \{1, 2, \ldots, n\}$$

We start with a *fair dataset* (FD) which guarantees equality of group-wise base rates, and correctness of the ground truth decision labels ys, i.e., no individuals were discriminated against based on their sensitive attribute. We denote the group-wise positive rate as $\mathbb{P}(y = +) = \mathbb{P}(y = + \mid s) = \alpha$. Our goal is to show that missing data disparity alone is sufficient to cause statistical parity unfairness, even in the absence of other sources of unfairness. Therefore, the equal group-wise base-rate assumption is necessary to eliminate confounding factors in our analysis. Without such equality, statistical parity unfairness could come from many other sources. Let $\mathbb{P}(p) = \beta_p$ and $\mathbb{P}(u) = \beta_u$ be the ratio of privileged and unprivileged samples in the dataset respectively. Clearly, $\beta_p + \beta_u = 1$. Note that, for group-wise balanced datasets, $\beta_p = \beta_u = 0.5$.

Let $\mathbf{x} \mid s, y \sim \mathcal{N}(\boldsymbol{\mu}^{sy}, \boldsymbol{\Sigma}) = \mathcal{N}^{sy}$ where $\boldsymbol{\mu}^{sy} \in \mathbb{R}^d$ and $\Sigma = diag(\boldsymbol{\sigma})$. Without loss of generality, we assume $\boldsymbol{\mu}^{s+} > \boldsymbol{\mu}^{s-}$, $\boldsymbol{\mu}^{s+} - \boldsymbol{\mu}^{s-} = \delta_y \mathbf{e}$ and $\boldsymbol{\mu}^{py} - \boldsymbol{\mu}^{uy} = \delta_s \mathbf{e}$ for all s where δ_y and δ_s are constants and $\mathbf{e}$ is a vector of 1s. We avoid the extreme case where the

distributions between groups are completely different by restricting $\boldsymbol{\mu}^{p-} \leq \boldsymbol{\mu}^{u} \leq \boldsymbol{\mu}^{p+}$ and $\boldsymbol{\mu}^{u-} \leq \boldsymbol{\mu}^{p} \leq \boldsymbol{\mu}^{u+}$. Let $\mathbf{M}$ be the set of indices of incomplete features, i.e., $\forall m \in \mathbf{M}$; $\mathbf{R}_{:,m} \neq \bar{0}^T$. The probability of complete and incomplete cases in group s are as follows,

$$\mathbb{P}(r_m = 0 \mid s) = \gamma_s \quad \text{and} \quad \mathbb{P}(r_m = 1 \mid s) = 1 - \gamma_s$$

We define the missing value disparity as $\gamma_p \neq \gamma_u$. Without loss of generality, the rest of the formal analysis assumes $\gamma_p > \gamma_u$. If the likelihood of a missing feature value is independent of other features, i.e., $\mathbb{P}(r_{\mathrm{j} \notin \mathbf{M}} = 0) = 1$ and for any $m_1, m_2 \in \mathbf{M}$, $r_{m_1} \not\perp\!\!\!\perp r_{m_2}$, then it turns into an instance of complete case disparity as well.

In the absence of label and selection bias, group-wise disparity in the rates of missing values is the only source of bias in the dataset. We prove that applying missing value handling mechanisms on such disparity can induce bias in model outcomes. Specifically, we show that missing value disparity alone prejudices the trained model towards one of the groups which leads to disparity in the model outcome.

We consider the Gaussian Bayesian classifier, θ, as it is the theoretically optimal classifier for normally distributed features. By demonstrating that a Bayes-optimal (maximum accuracy) model exhibits unfairness, we show that *any* method that achieves maximal accuracy (and thus the same outcome) would be expected to be similarly unfair. Moreover, the conditional independence assumption of the Naive Bayes classifier (NBC) also holds under our assumptions, simplifying the analysis of Bayes-optimal to the uni-variate means/variances of Naive Bayes. We expect that if missing data induces unfairness in simple cases, then missing data induced unfairness will still show up in more complex cases (conditional dependence among variables, other sources of unfairness, etc.) Section 4.3 empirically validates that missing data disparity induces unfair statistical parity for other types of classifiers including SVM, Neural Networks (NN), and state-of-the-art fairness-enhancing interventions, and for data where conditional independence and other simplifying assumptions do not hold.

The distribution learned by a naive Bayesian classifier on the complete domain $\mathcal{D}$ is $\mathbf{x} \mid y \sim \mathcal{N}(\boldsymbol{\mu}_\theta^y, \Sigma_\theta^y) = \mathcal{N}_\theta^y$ where $\boldsymbol{\mu}_\theta^y$ and $\Sigma_\theta^y = diag(\boldsymbol{\sigma}_\theta^y)$ are as follows,

$$\boldsymbol{\mu}_\theta^y = \beta_p \boldsymbol{\mu}^{py} + \beta_p \boldsymbol{\mu}^{uy} \tag{4.1}$$

$$\boldsymbol{\sigma}_\theta^y = \sqrt{\boldsymbol{\sigma}^2 + \beta_p \beta_u (\boldsymbol{\mu}^{py} - \boldsymbol{\mu}^{uy})^2} \tag{4.2}$$

The KL-divergence between two vectors of independent uni-variate Gaussian distribution is as follows.

$$\begin{aligned}
KL(P \mid\mid Q) &= \int_{\mathbf{x}} \mathbb{P}(\mathbf{x}) \log \frac{\mathbb{P}(\mathbf{x})}{\mathbb{Q}(\mathbf{x})} d\mathbf{x} \\
&= \int_{\mathbf{x}} \prod_{x_\mathrm{i}} \mathbb{P}(x_\mathrm{i}) \log \left(\frac{\prod_{x_\mathrm{i}} \mathbb{P}(x_\mathrm{i})}{\prod_{x_\mathrm{i}} \mathbb{Q}(x_\mathrm{i})} \right) d\mathbf{x} \qquad (x_\mathrm{i}\text{s are independent of each other.}) \\
&= \int_{\mathbf{x}} \left(\prod_{x_\mathrm{i}} \mathbb{P}(x_\mathrm{i}) \right) \left(\sum_{x_\mathrm{i}} \log \left(\frac{\mathbb{P}(x_\mathrm{i})}{\mathbb{Q}(x_\mathrm{i})} \right) \right) d\mathbf{x} \\
&= \sum_{x_\mathrm{i}} \int_{\mathbf{x}} \left(\prod_{x_\mathrm{j}} \mathbb{P}(x_\mathrm{j}) \right) \left(\log \left(\frac{\mathbb{P}(x_\mathrm{i})}{\mathbb{Q}(x_\mathrm{i})} \right) \right) d\mathbf{x} \\
&= \sum_{x_\mathrm{i}} \left(\int_{x_\mathrm{i}} \mathbb{P}(x_\mathrm{i}) \log \left(\frac{\mathbb{P}(x_\mathrm{i})}{\mathbb{Q}(x_\mathrm{i})} \right) dx_\mathrm{i} \right) \quad (x_\mathrm{i}\text{s are independent and } \textstyle\int \mathbb{P}(x) dx = 1.) \\
\therefore KL(P \mid\mid Q) &= \sum_{x_\mathrm{i}} KL(\mathbb{P}(x_\mathrm{i}) \mid\mid \mathbb{Q}(x_\mathrm{i}))) \qquad (4.3)
\end{aligned}$$

We know that the KL-divergence between two uni-variate Gaussians $\mathbb{P}(x_\mathrm{i}) = \mathcal{N}(\mu_1, \sigma_1)$ and $\mathbb{Q}(x_\mathrm{i}) = \mathcal{N}(\mu_2, \sigma_2)$ is,

$$KL(\mathbb{P}(x_\mathrm{i}) \mid\mid \mathbb{Q}(x_\mathrm{i})) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2} \tag{4.4}$$

Using (4.1), (4.2), (4.3) and (4.4), we derive that in the complete case domain,

$$KL(\mathcal{N}^{py} \mid\mid \mathcal{N}_\theta^y) - KL(\mathcal{N}^{uy} \mid\mid \mathcal{N}_\theta^y) \propto \beta_u^2 - \beta_p^2 \tag{4.5}$$

Equation (4.5) shows that the learned class-wise distributions $\mathcal{N}_\theta^y$, are equidistant from the group-wise distributions $\mathcal{N}^{sy}$ when $\beta_p = \beta_u$ (group-wise balanced). If $\beta_p \neq \beta_u$, the

classifier learns a distribution closer to the group with the higher number of samples. This is aligned with the reported fairness issues with imbalanced datasets in [11].

4.2.1 Dropping

We start with the simple case of handling missing data by dropping the incomplete cases from the training dataset. Since new samples are not introduced, the group-wise complete case feature distributions remain intact. A naive Bayesian classifier learns the class-wise means and variances from the reduced dataset. The learned mean is the weighted sum of the group-wise means and the variance is the pooled variance of the sub-groups. Formally, the class-wise mean and variance learned by an NBC from the reduced dataset are as follows,

$$\boldsymbol{\mu}_\theta^y = \frac{\gamma_p\beta_p\mu^{py} + \gamma_u\beta_u\mu^{uy}}{\gamma_p\beta_p + \gamma_u\beta_u} \tag{4.6}$$

$$\begin{aligned} \boldsymbol{\sigma}_\theta^y &= \sqrt{\boldsymbol{\sigma}^2 + \frac{(\gamma_p\beta_p\gamma_u\beta_u)(\boldsymbol{\mu}^{py} - \boldsymbol{\mu}^{uy})^2}{(\beta_p\gamma_p + \beta_u\gamma_u)^2}} \\ \therefore \boldsymbol{\sigma}_\theta^y &= \sqrt{\boldsymbol{\sigma}^2 + \frac{(\gamma_p\beta_p\gamma_u\beta_u)(\delta_s\mathbf{e})^2}{(\beta_p\gamma_p + \beta_u\gamma_u)^2}} \end{aligned} \tag{4.7}$$

Dropping the missing values modifies (4.5) as follows,

$$KL(\mathcal{N}^{py} \,||\, \mathcal{N}_\theta^y) - KL(\mathcal{N}^{uy} \,||\, \mathcal{N}_\theta^y) \propto \gamma_u^2\beta_u^2 - \gamma_p^2\beta_p^2 \tag{4.8}$$

Equation (4.8) indicates that upon dropping the incomplete samples, the learned distributions move closer to the group which has higher $\gamma_s\beta_s$. When the dataset is balanced ($\beta_p = \beta_u$), the disparity in learned distribution only depends on γ_s. According to our assumption $\gamma_p > \gamma_s$, the classifier distributions will be skewed towards the privileged group. Since the balanced dataset avoids all sources of bias except the missing value disparity, we conclude that the disparity in learned distribution is introduced by missing value disparity.

4.2.2 Imputation

Imputation allows us to turn the incomplete samples into complete ones. For tractability of analysis, we show how simple mean imputation results in unfair outcomes. Section 4.3 empirically demonstrates missing value disparity induced bias from other imputation algorithms. Let $\mathcal{D}' = \{\mathbf{X}', \mathbf{S}, \mathbf{Y}\}$ indicate the imputed dataset where each $\mathbf{x}'_{\text{i}}$ is either $\mathbf{x}_{\text{i}}$ itself or the imputed version of incomplete $\mathbf{x}_{\text{i}}$. Imputation preserves the feature distributions except for feature x_m where $m \in \mathbf{M}$. Simple mean imputation replaces the missing values of the m^{th} feature with the complete sample mean,

$$\mu_m = \mathbb{E}[x_m \mid \mathbf{r} = \bar{0}] = \sum_s \frac{\gamma_s \beta_s}{\gamma_p \beta_p + \gamma_u \beta_u} \mu_m^s$$

Here, $\mu_m^s = \sum_y \mathbb{P}(y \mid s) \mu_m^{sy}$. After imputation the group-wise conditional means changes as follows,

$$\mu_m^{sy\prime} = \gamma_s \mu_m^{sy} + (1 - \gamma_s)\mu_m \tag{4.9}$$

Using (4.9), we obtain the following group-wise means:

$$\mu_m^{s\prime} = \gamma_s \mu_m^{s} + (1 - \gamma_s)\mu_m$$

Mean imputation alters the group-wise conditional variances as well. The group-wise conditional variance is the pooled variance between the complete sample variance σ_m^2 and the variance among the imputed samples (zero).

$$\sigma_m^{sy\prime} = \sqrt{\gamma_s \sigma_m^2 + \gamma_s(1 - \gamma_s)(\mu_m^{sy} - \mu_m)^2} \tag{4.10}$$

The naive Bayesian classifier learns the distributions $\mathbf{x} \mid y \sim \mathcal{N}(\boldsymbol{\mu}_\theta^y, diag(\boldsymbol{\sigma}_\theta^y)) = \mathcal{N}_\theta^y$. The means and variances of the learned distribution $\mathcal{N}_\theta^y$ are identical to (4.1) and (4.2) except at the $m^{th}(\in \mathbf{M})$ index. The learned mean $\mu_{m;\theta}^{y\prime}$ is the weighted sum of the imputed group-wise means.

$$\mu^{y}_{m;\theta} = \mu^{y\prime}_{m} = \sum_{s} \mathbb{P}(s) * \mu^{sy\prime}_{m}$$

$$\therefore \mu^{y}_{m;\theta} = \sum_{s} \gamma_s \beta_s \mu^{sy}_{m} + \sum_{s} \beta_s (1-\gamma_s) \mu_m \tag{4.11}$$

Similarly, the learned class-wise variance is the pooled variance between the imputed group-wise variances.

$$\sigma^{y}_{m;\theta} = \sigma^{y\prime}_{m} = \sqrt{\sum_{s} \beta_s (\sigma^{sy\prime}_{m})^2 + \beta_p \beta_u (\mu^{py\prime}_{m} - \mu^{uy\prime}_{m})^2} \tag{4.12}$$

Previously, (4.3) established that the KL-divergence of multivariate independent distributions is the summation of independent uni-variate KL-divergences. Therefore, we compare the uni-variate KL-divergences.

$$KL(\mathcal{N}^{py}_{\mathrm{i}} \,||\, \mathcal{N}^{y}_{\mathrm{i};\theta}) - KL(\mathcal{N}^{uy}_{\mathrm{i}} \,||\, \mathcal{N}^{y}_{\mathrm{i};\theta}) = \quad \frac{(\mu^{py}_{\mathrm{i}} - \mu^{y}_{\mathrm{i};\theta})^2 - (\mu^{uy}_{\mathrm{i}} - \mu^{y}_{\mathrm{i};\theta})^2}{2(\sigma^{y}_{\mathrm{i};\theta})^2} \tag{4.13}$$

By construction, for $\mathrm{i} \notin \mathbf{M}$, $\mu^{py}_{\mathrm{i}} - \mu^{y}_{\mathrm{i};\theta} = \mu^{uy}_{\mathrm{i}} - \mu^{y}_{\mathrm{i};\theta}$. Therefore, when $\mathrm{i} \notin \mathbf{M}$,

$$KL(\mathcal{N}^{py}_{\mathrm{i}} \,||\, \mathcal{N}^{y}_{\mathrm{i};\theta}) = KL(\mathcal{N}^{uy}_{\mathrm{i}} \,||\, \mathcal{N}^{y}_{\mathrm{i};\theta}) \tag{4.14}$$

Hence, we now concentrate on the KL-divergences between $\mathcal{N}^{sy}_{m}$ and $\mathcal{N}^{y}_{m;\theta}$ where $m \in \mathbf{M}$.

Case 1: Balanced Training Set, $\beta_p = \beta_u = 0.5$

When $\delta_s > 0$, by construction $\mu_m < \mu^{u+}_{m} < \mu^{p+}_{m}$. Imputation moves (reduces) $\mu^{u+\prime}_{m}$, $\mu^{p+\prime}_{m}$ and their average $\mu^{+}_{m;\theta}$ closer to μ_m. Therefore, $\mu^{p+}_{m} - \mu^{+}_{m;\theta} > \mu^{u+}_{m} - \mu^{+}_{m;\theta}$. Consequently, (4.13) indicates that when $\delta_s > 0$, for any $m \in \mathbf{M}$,

$$KL(\mathcal{N}^{p+}_{m} \,||\, \mathcal{N}^{+}_{m;\theta}) > KL(\mathcal{N}^{u+}_{m} \,||\, \mathcal{N}^{+}_{m;\theta}) \tag{4.15}$$

Likewise, imputation increases $\mu_m^{u-\prime}$, $\mu_m^{p-\prime}$ and their average $\mu_{m;\theta}^{-}$ towards μ_m. As a result, $\mu_m^{p-} - \mu_{m;\theta}^{-} < \mu_m^{u-} - \mu_{m;\theta}^{-}$. Therefore, the following inequality holds,

$$KL(\mathcal{N}_m^{p-} \,||\, \mathcal{N}_{m;\theta}^{-}) < KL(\mathcal{N}_m^{u-} \,||\, \mathcal{N}_{m;\theta}^{-}) \tag{4.16}$$

Using (4.3), (4.14), (4.15) and (4.16) we summarize that,

$$\begin{aligned} KL(\mathcal{N}^{p+} \,||\, \mathcal{N}_\theta^{+}) &> KL(\mathcal{N}^{u+} \,||\, \mathcal{N}_\theta^{+}) \\ KL(\mathcal{N}^{p-} \,||\, \mathcal{N}_\theta^{-}) &< KL(\mathcal{N}^{u-} \,||\, \mathcal{N}_\theta^{-}) \end{aligned} \tag{4.17}$$

In contrast, $\delta_s < 0$ changes the directions of the inequalities in (4.17).

$$\begin{aligned} KL(\mathcal{N}^{p+} \,||\, \mathcal{N}_\theta^{+}) &< KL(\mathcal{N}^{u+} \,||\, \mathcal{N}_\theta^{+}) \\ KL(\mathcal{N}^{p-} \,||\, \mathcal{N}_\theta^{-}) &> KL(\mathcal{N}^{u-} \,||\, \mathcal{N}_\theta^{-}) \end{aligned} \tag{4.18}$$

Equations (4.17) and (4.18) indicate that when $\delta_s < 0$, after imputation, NBC more accurately replicates the distribution of the unprivileged positive group and the privileged negative group. It implies that the unprivileged positive group will receive more accurate positive predictions than the privileged positive group and the opposite trend in the negative group. Thus the classifier ends up inherently disfavoring the privileged. The opposite trend is expected in the negative samples. For $\delta_s > 0$, the disparity is inverted from the unprivileged towards the privileged. This completes our constructive proof of training unfair models on fair and balanced training data free from disparities other than missing value disparity.

Case 2: Imbalanced Training Set, $\beta_p \neq \beta_u$

For imbalanced datasets, we expand (4.13) as follows,

$$KL(\mathcal{N}_m^{py} \,||\, \mathcal{N}_{m;\theta}^{y}) - KL(\mathcal{N}_m^{uy} \,||\, \mathcal{N}_{m;\theta}^{y}) = \frac{\delta_s \left[(1-2\gamma_p\beta_p)(\mu_m^{py} - \mu_m) + (1-2\gamma_u\beta_u)(\mu_m^{uy} - \mu_m)\right]}{2(\sigma_{m;\theta}^{y})^2} \tag{4.19}$$

When $\delta_s > 0$, from (4.19) $KL(\mathcal{N}_m^{p+} \,||\, \mathcal{N}_{m;\theta}^{+}) < KL(\mathcal{N}_m^{u+} \,||\, \mathcal{N}_{m;\theta}^{+})$ holds when the product $\gamma_p\beta_p > 0.5$. When both $\gamma_p\beta_p \leq 0.5$ and $\gamma_u\beta_u \leq 0.5$, $KL(\mathcal{N}_m^{p+} \,||\, \mathcal{N}_{m;\theta}^{+}) > KL(\mathcal{N}_m^{u+} \,||\, \mathcal{N}_{m;\theta}^{+})$.

It implies that with fixed β_p and β_u, the rates of missing values can control which group distribution is closely replicated by the Bayesian classifier.

Throughout the above discussion, we assumed that the missing values only appeared in the training data. When the group-wise distributions are identical, $\delta_s = 0$, classifier performance will be similar in both groups. However, if test samples contained missing values, (4.10) shows that missing value disparity instigates feature disparity; the phenomenon shown to introduce bias in [65].

We frame our argument using KL-divergence of the original distributions and the learned distributions. After dropping or imputation, the change in the classifier results in a change in prediction probabilities. The prediction rates are modified only when one or more of the changed prediction probabilities crosses the prediction threshold (typically 0.5). On the other hand, even without a change in the predictions, it changes the overall ranking of the individuals in the population. In resource-constrained situations where favorable decisions are distributed among top-ranked candidates, such as selective college admission, alternate rankings still induce an unfair distribution even if the classification (e.g., meets minimum standards to succeed in college) does not change.

Equations (4.17) and (4.18) show that mean-value imputation with missing value disparity unfairly skews the model toward one group. We avoid complex imputations for the sake of tractability of the theoretical analysis. In Section 4.3, we empirically show induced bias from complex methods, such as k-NN [34] and MICE [35] (summarized in Table 4.1) and classifiers beyond NBC (summarized in Table 4.7 and 4.8).

4.3 Experimental Results

We show empirical evidence of induced bias due to missing value disparity. Our experiment involves both balanced synthetic and several imbalanced real-world datasets, three distinct MAR mechanisms for missing data, five missing value handling mechanisms, and five different types of classifiers including traditional classifiers (SVM, Neural Networks) and two state-of-the-art fairness-aware ones. The real-world datasets are publicly available and convey more complex conditions than the ones discussed in the formal analysis. The source code

of our experiments is available at https://github.com/rakinhaider/MissingValueDisparity. To maintain the reproducibility of the results, we use pre-fixed random seeds in our experiments (further details are found in the README of the above repository).

4.3.1 Synthetic Fair Balanced Dataset

Induced bias studies unfair model outcomes caused by ML components other than historical bias typically appearing in the form of label bias or majority-minority groups. As real-world datasets rarely conform to fairness guarantees, analysis of induced bias from missing value disparity is often non-conclusive on these historically biased data. Instead, synthetic data enable us to maintain the absence of majority/minority groups, equality of the group-wise base rates, and the correctness/fairness of the ground truth labels. It guarantees that missing value disparity is the only form of disparity among the samples. Thus, we can validate that even if datasets with strong fairness guarantees existed, unfair outcomes could still appear due to missing value disparity. Since all confounding factors, such as label bias, are eliminated, we conclude that the unfair experimental results are due to missing value disparity. Note that, we do investigate real-world datasets (where outcome bias also potentially comes from other factors) in Section 4.3.5.

We generate a synthetic training dataset $\mathcal{D}$ with 10000 samples each having two features x_1 and x_2, sensitive attribute s, and label y. Each synthetic dataset maintains balance in the number of samples from both sensitive groups, i.e., $\mathbb{P}(u) = \mathbb{P}(p) = 0.5$. The sample generation process also ensures the base rates $\mathbb{P}(y = + \mid p) = \mathbb{P}(y = + \mid u) = \alpha$. In our experiment, we use $\alpha = 0.5$. To generate a sample, we first randomly assign s and y. We follow the generative model discussed in Section 4.2 to sample x_{i}s. More specifically, the x_{i}s are sampled from $x_{\mathrm{i}} \sim \mathcal{N}(\mu_{\mathrm{i}}^{sy}, \sigma)$ where $\mu_{\mathrm{i}}^{p+} = 10$, $\delta_y = 10$, $\delta_g \in \{-3, 0, 3\}$ and $\sigma = 5$. We use the same process to generate the synthetic test set.

The process of introducing artificial missing values in the dataset is called amputation. We induce artificial random missing values in the dataset by first randomly sampling (i, j) pairs and setting $r_{\mathrm{i,j}} = 1$. Then, using $\mathbf{R}$, we drop the value of the j^{th} feature from the i^{th} sample, if $r_{\mathrm{i,j}} = 1$. Let, $\mathbf{R} \bigodot \mathcal{D}$ is the amputated dataset with incomplete samples. Upon

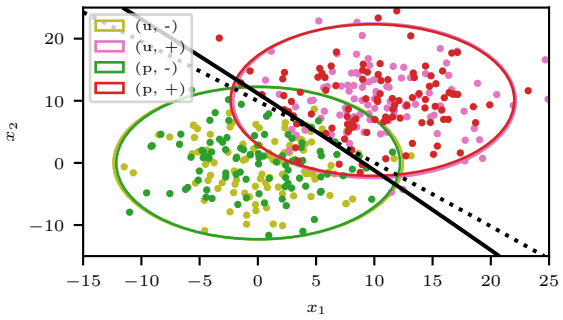


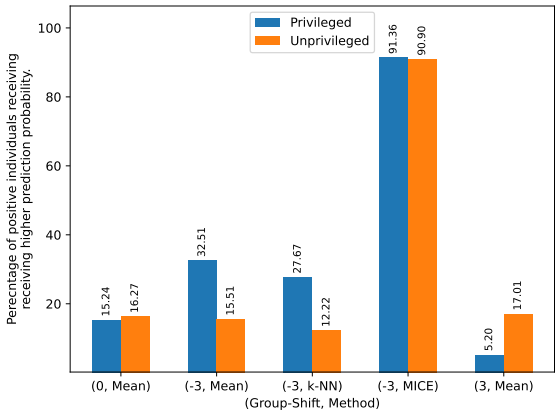
Figure 4.1. Scatter plot of identically distributed privileged and unprivileged samples (i.e., $\delta_s = 0$) with the solid and dotted line indicating θ and θ' boundaries. The ellipses indicate the group-wise two-variate normal distributions.

dropping or imputation on $\mathbf{R} \odot \mathcal{D}$, we obtain the complete dataset $\mathcal{D}'$. This method allows access to both $\mathcal{D}$ and $\mathcal{D}'$, enabling comparison with the baseline models trained on $\mathcal{D}$. In the following experiment, we assume that the missing values only appear on x_2. Since j(= 2) is fixed, the amputation mechanism only samples indices i from the privileged and the unprivileged group according to probability $1 - \gamma_p$ and $1 - \gamma_u$ respectively. By default, we use $\gamma_p = 0.9$ and $\gamma_u = 0.6$. The amputated dataset is later repaired using either dropping the incomplete samples or applying imputation techniques.

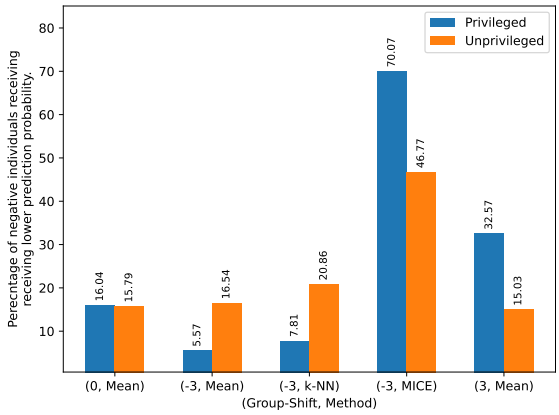
4.3.2 Induced Bias with Independent Features

We begin our discussion with mutually independent features (although correlated with the outcome). The impact on accuracy in this case is small, but the changes in confidence of the predictions are notable. To demonstrate this, we look at the changes in the probability of an outcome as well as ranking with respect to the probability of a positive outcome. Let θ be the Naive Bayes classifier (NBC) trained on the original complete samples $\mathcal{D}$. We refer to this model as the *baseline* model. We further assume θ' is the NBC trained on $\mathcal{D}'$ which was obtained after dropping or imputing the amputated dataset $\mathbf{R} \odot \mathcal{D}$. Define the positive prediction probability from classifier t as $t(\mathbf{x}) = \mathbb{P}(\hat{y} = + \mid \mathbf{x}; t)$. We rank the individuals by $\theta(\mathbf{x})$ and $\theta'(\mathbf{x})$ to obtain $rank(\mathbf{x}; \theta)$ and $rank(\mathbf{x}; \theta')$ respectively. Table 4.1 reports the

group-wise percentages of individuals who received higher or lower $\theta'(\mathbf{x})$ than $\theta(\mathbf{x})$ for each (s, y) pair. Similarly, we report the changes in $rank(\mathbf{x}; t)$ from $\theta(\mathbf{x})$ to $\theta'(\mathbf{x})$. We report the means and standard deviations over 10 random initiations of the synthetic training data and corresponding missing values in Table 4.1 and in Figure 4.2. Intuitively, higher $t(\mathbf{x})$ is favorable for a positive sample and unfavorable for the negative ones. In addition, an improved ranking is advantageous since it increases the likelihood of resource allocation. For example, lower-ranked students are more likely to be admitted.



(a) Percentage of positive samples receiving a higher prediction probability after imputation.



(b) Percentage of negative samples receiving a lower prediction probability after imputation.

Figure 4.2. Changes in prediction probabilities in the positive and negative samples.

Identical Group-wise Distributions

We first consider identically distributed privileged and unprivileged groups, i.e., the feature distributions are independent of the sensitive attribute s. Formally, $x_{\text{i}} \not\perp\!\!\!\perp s$ or $\delta_s = 0$. Figure 4.1 shows 100 samples in each (s, y) group, their normal distribution ellipses, and the joint classifier boundaries. Table 4.1 shows that, with $\delta_s = 0$, the portion of individuals receiving lower or higher $\theta'(\mathbf{x})$ than $\theta(\mathbf{x})$ are similar between $(s, +)$ pairs and $(s, -)$ pairs. On average, the imputation mechanism equally disfavors the positive samples and equally favors the negative ones irrespective of s. The results support the theoretical claim that with $\delta_s = 0$ and mean imputation, $KL(\mathcal{N}^{sy} \,||\, \mathcal{N}^{y}_{\theta'})$ will be equidistant for each s. The overlapping

Table 4.1. Group-wise summaries of changes in positive prediction probabilities and relative rankings of the individuals due to imputation mechanism. Changes in positive prediction probability and relative ranking are denoted by $\delta_{\mathbb{P}}$ and δ_{rn} respectively. The percentages of samples who received higher ($\delta_{\mathbb{P}} > 0$) or lower ($\delta_{\mathbb{P}} > 0$) positive prediction probabilities and the average change ($\Delta_{\mathbb{P}} = mean(\delta_{\mathbb{P}})$) in positive prediction probability after imputation are reported. Similarly, the percentage of samples ranked higher ($\delta_{rn} > 0$) or lower ($\delta_{rn} > 0$) and the average change in rank ($\Delta_{rn} = mean(\delta_{rn})$) in each group after imputation is reported in the latter columns.

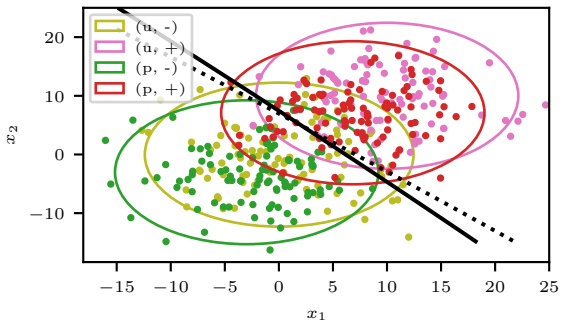
δ_s	Method	(s, y)	Positive Prediction Probabilities			Rankings		
			$\delta_{\mathbb{P}} < 0$	$\delta_{\mathbb{P}} > 0$	$\Delta_{\mathbb{P}}$	$\delta_{rn} < 0$	$\delta_{rn} > 0$	Δ_{rn}
0	Mean	(u, -)	15.79	84.21	8.4e-03	49.57	50.25	5.85
		(u, +)	83.73	16.27	-7.5e-03	49.50	50.29	1.38
		(p, -)	16.04	83.96	8.1e-03	49.21	50.61	5.26
		(p, +)	84.76	15.24	-8.8e-03	51.01	48.84	-12.49
-3 (< 0)	Mean	(u, -)	16.54	83.46	1.7e-02	49.71	50.20	4.65
		(u, +)	84.49	15.51	-2.4e-03	49.90	49.77	-1.32
		(p, -)	5.57	94.43	1.1e-02	48.68	50.95	8.30
		(p, +)	67.49	32.51	2.6e-04	51.45	48.42	-11.63
-3 (< 0)	k-NN	(u, -)	20.86	79.14	1.5e-02	47.52	52.39	11.79
		(u, +)	87.78	12.22	-8.6e-03	48.83	50.94	-10.92
		(p, -)	7.81	92.19	1.1e-02	46.15	53.49	14.12
		(p, +)	72.33	27.67	-9.6e-03	49.46	50.45	-14.99
-3 (< 0)	MICE	(u, -)	46.77	53.23	2.6e-03	41.71	57.69	-0.77
		(u, +)	9.10	90.90	2.2e-03	35.22	62.89	-0.74
		(p, -)	70.07	29.93	5.7e-04	43.70	55.53	0.63
		(p, +)	8.64	91.36	4.8e-03	34.46	64.41	0.88
3 (> 0)	Mean	(u, -)	15.03	84.97	2.8e-03	49.08	50.64	7.15
		(u, +)	82.99	17.01	-1.7e-02	49.14	50.69	3.32
		(p, -)	32.57	67.43	-1.1e-03	49.90	50.00	0.92
		(p, +)	94.80	5.20	-1.2e-02	50.94	48.84	-11.39

sample distributions in Figure 4.1 confirm that any change in the model has a similar effect on both s.

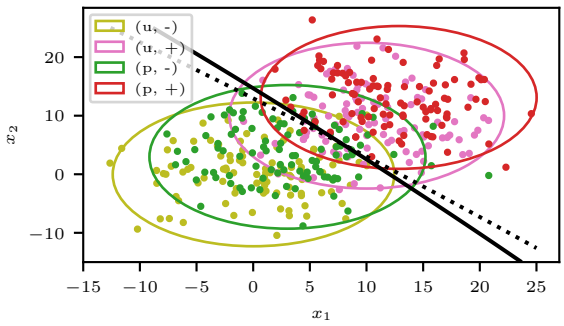
Non-identical Group-wise Distribution

Realistically, the feature distributions of privileged and unprivileged groups are often different. Figure 4.3a and 4.3b show privileged and unprivileged samples where they are non-identically distributed ($\delta_s < 0$ and $\delta_s > 0$ respectively). Table 4.1 contains the statistics of changes in $\theta'(\mathbf{x})$ from $\theta(\mathbf{x})$ with $\delta_s < 0$ after mean, k-NN and MICE imputations. We observe that mean imputation increases the $t(\mathbf{x})$ for 32.5% of individuals in $(p, +)$ as compared to 15.5% in $(u, +)$. That means, more than twice as many privileged individuals received improved $\theta'(x)$ as unprivileged. A steeper average downward trend is found in the ranking of the individuals in $(p, +)$ (-11.63) than in $(u, +)$ (-1.32). Since a lower ranking is advantageous, the privileged group appears to be unfairly favored by θ'. In addition, a larger portion of $(u, -)$ had $\delta_{\mathbb{P}} < 0$ than the $(p, -)$. Furthermore, $(p, -)$ encounters a steeper increase in average ranking than $(u, -)$. This shows that the ability to discern between positive and negative samples in the unprivileged group has decreased, decreasing the likelihood of correct/appropriate decisions.

Table 4.1 also shows unfair shifts in $t(\mathbf{x})$ and $rank(\mathbf{x}, t)$ for k-NN and MICE imputations. Similar to mean imputation, k-NN imputation unfairly advantages groups $(u, -)$ and $(p, +)$, but to a greater extent. On the other hand, MICE imputation favors both positive and negative privileged samples by moving the prediction probability closer to the ground truths. We avoid fair ranking metrics such as normalized discounted metrics [74] since they concentrate on group-wise top-k recommendation statistics and do not capture the smaller and more frequent individual changes in rankings studied here.



(a) Non-identically distributed samples with $\delta_s > 0$.



(b) Non-identically distributed samples with $\delta_s > 0$.

Figure 4.3. Scatter plot of non-identically distributed privileged and unprivileged samples (i.e., $\delta_s \neq 0$) with the solid and dotted line indicating θ and θ' boundaries. The ellipses indicate the group-wise two-variate normal distributions.

4.3.3 Induced Bias with Correlated Features

We now investigate a case where the features x_1 and x_2 are correlated and sampled following the conditional distributions,

$$x_1 \mid s, y \sim \mathcal{N}(\mu^{sy}_{x_1}, \sigma_{x_1}) \quad \text{and} \quad x_2 = a_1 x_1 + a_2 y \tag{4.20}$$

Here, $\mu^{u+}_{x_1} = 10$, $\mu^{u-}_{x_1} = 0$, $\mu^{py}_{x_1} - \mu^{uy}_{x_1} = \delta_s$ where $\delta_s \in \{-3, 0, 3\}$. For this section, we use $a_1 = 1$ and $a_2 = 4$. Being a linear combination of x_1 and y, x_2 also follows the normal distribution, $x_2 \mid s, y \sim \mathcal{N}\left(\mu^{sy}_{x_1} + 4y, \sigma^2_{x_1}\right)$. Clearly $x_2 \perp\!\!\!\perp x_1, y$. However, imputation disproportionately weakens the relation between x_2 and y due to missing value disparity and consequently results in false or unfavorable predictions.

Using (4.20) and 10 different random initiations, we randomly generate 10 synthetic fair datasets $\mathcal{D}$ with missing value disparity similar to the ones described in Section 4.3.1. The missing values are handled by either dropping them or imputing them with mean imputation, MICE [35], k-NN imputation [34], and Softimpute [36]. We train Naive Bayesian classifiers

Table 4.2. Disparities in group-wise accuracies, base prediction rates, and false positive rates of NBC on SFBD with correlated features, missing value disparity, and different values of δ_s.

δ_s	Method	$SP(ACC;\theta)$	$SP(\theta)$	$SP(FPR;\theta)$
0	Baseline	0.36 (0.02)	-0.44 (0.04)	-0.80 (0.04)
	Drop	0.36 (0.03)	-0.44 (0.05)	-0.79 (0.07)
	Mean	0.15 (0.01)	-0.27 (0.03)	-0.42 (0.02)
	MICE	0.25 (0.06)	-0.37 (0.03)	-0.61 (0.06)
	k-NN	0.15 (0.03)	-0.32 (0.04)	-0.46 (0.04)
	Softimpute	0.28 (0.03)	-0.34 (0.02)	-0.62 (0.03)
-3	Baseline	0.10 (0.12)	-8.06 (0.03)	-8.16 (0.12)
	Drop	1.17 (0.10)	-8.08 (0.04)	-9.26 (0.11)
	Mean	0.70 (0.09)	-8.76 (0.05)	-9.46 (0.13)
	MICE	0.06 (0.12)	-8.12 (0.03)	-8.18 (0.13)
	k-NN	0.08 (0.14)	-9.05 (0.07)	-9.13 (0.15)
	Softimpute	3.07 (0.06)	-8.78 (0.09)	-11.86 (0.13)

on the imputed datasets and report the mean $SP(\theta)$, $SP(ACC;\theta)$ and $SP(FPR;\theta)$ (defined in (2.4), (2.7) and (2.8) respectively) and their standard deviations in Table 4.2.

Identical Group-wise Distribution

Table 4.2 reports lower disparity in group-wise accuracies, selection rates, and false positive rates from NBC after each imputation mechanism when the groups are identically distributed. We observe that imputation mechanisms often reduce the group-wise disparities when the underlying group-wise distributions are the same. Since we do not induce missingness in the test samples, they remain unchanged and follow identical distributions. Therefore, any model is expected to perform similarly on both p and u. This further corroborates that the synthetic data is free from label bias and does not demonstrate unfair model predictions even without any fairness interventions.

Non-identical Group-wise Distributions

Table 4.2 shows the induced disparity in model outcome due to missing value disparity when the group-wise distributions are different between privileged and unprivileged groups, i.e., $\delta_s \neq 0$. The increase in disparity in accuracy and false positive rate is most prevalent in dropping and Softimpute imputation whereas k-NN imputation shows the highest increase in positive prediction rate disparity. We further experiment by varying the level of group-wise disparity. We fix the rate of missing values in the privileged group to $\gamma_p = 0.1$ and vary the rate of missingness in the unprivileged group between 0.1 to 0.6. Figure 4.4 shows that with the increase in group-wise missing value disparity, the models exhibit an increased disparity in false positive rates between privileged and unprivileged groups. We attribute the increased false positive rate disparity to missing value disparity since all other disparities are absent in the SFBD.

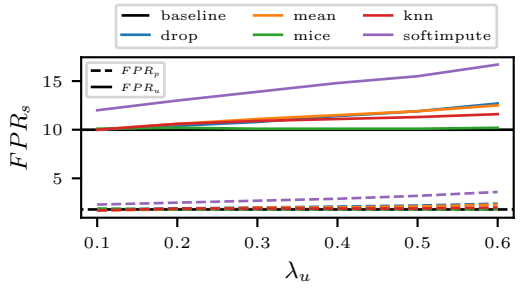


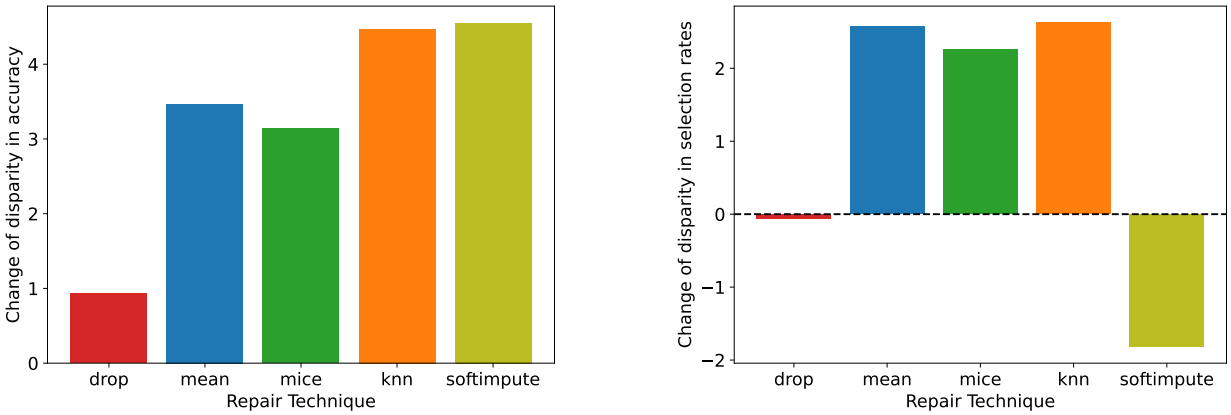
Figure 4.4. Variations in group-wise FPR_s with $1 - \gamma_u$.

4.3.4 Missing Values in Test Data

Real-world scenarios may not guarantee the completeness of the test samples. Therefore, pre-processing imputation pipelines are often needed for test samples as well. Dropping incomplete samples does not make sense for test data, so we concentrate on imputation. In this experiment, we follow the synthetic data generation strategy described in 4.3.1 except for introducing missing value disparity in test samples as well. In Table 4.3 and Figure 4.5 observed increased induced bias when the test samples contained missing values. Test case imputation increases false positive rate disparity to as much as 15.26% for k-NN imputation

Table 4.3. Disparities in group-wise accuracies, base prediction rates, and false positive rates of NBC classifiers on SFBD with correlated features, missing value disparity, and incomplete test samples.

Method	$SP(ACC;\theta)$	$SP(\theta)$	$SP(FPR;\theta)$
Baseline	0.10 (0.12)	-8.06 (0.03)	-8.16 (0.12)
Drop	1.03 (0.08)	-8.00 (0.06)	-9.03 (0.11)
Mean	3.57 (0.11)	-10.63 (0.08)	-14.20 (0.18)
MICE	3.24 (0.10)	-10.32 (0.05)	-13.56 (0.10)
k-NN	4.57 (0.29)	-10.69 (0.30)	-15.26 (0.52)
Softimpute	4.65 (0.03)	-6.24 (0.13)	-10.89 (0.13)



(a) Changes in disparity in accuracy.

(b) Changes in disparity in selection rates.

Figure 4.5. Changes in prediction rate disparities after repairing the amputated synthetic dataset with different imputation mechanisms. Here the test samples were allowed to include missing values and $\delta_s < 0$.

from a baseline difference of 8.16%. Similarly, the difference in group-wise accuracy was increased from 0.1% to 4.65% by Softimpute imputation. It means imputation converts feature x_2 into a poor predictor for the unprivileged group than for the privileged. Such a phenomenon, called feature disparity, is a proven source of induced bias [65].

4.3.5 Real-world Datasets

We investigate the extent of missing value disparity on benchmark datasets studied in FairML literature. Table 4.4 shows 6 real-world datasets, their number of samples, and the

list of sensitive attributes among the features (the bold ones are used in our experiments). The `COMPAS` dataset is used to predict *recidivism* (unfavorable) vs *no-recidivism* (favorable) among a population of 2103 Caucasians (privileged) and 3175 non-Caucasians (unprivileged). Following the recommendations in [75], `COMPAS` is pre-processed to binarize its features. Predicting high-vs-low *incomes* of individuals is the goal of the `Adult` dataset. Recent work [76] suggested that `Adult` covers only an extremely privileged end of the US population. Its use in fairness studies accompanies certain limitations and is discouraged. Among the alternative datasets, derived from the US census by the authors of [76], we study `FolkIncome` which is used to predict the high (favorable) or low (unfavorable) income of individuals while the target of `FolkCoverage` is to predict whether an individual has private (unfavorable) or public (favorable) medical coverage. The `PIMA` dataset is tasked with predicting diabetes among 768 female patients using a range of numeric attributes such as # of pregnancies, blood pressure, etc. Here, 118 patients who are 45 years or above in age are considered unprivileged. Another healthcare-related dataset is `Heart`. Its goal is to predict the presence or absence of cardiovascular disease among individuals. Compared to the other datasets, `PIMA` and `Heart` are more suitable for mean imputation since their features are typically numeric. The dataset is pre-processed to convert the sensitive attribute and the target attribute into binary features. We amputate the complete datasets and later repair them by either dropping or imputation. We compare outcome fairness with and without the amputation-imputation processing. In this way, any increase in outcome disparity can be attributed to missing value disparity since it is the only source of disparity that is added through amputation.

We partition each dataset into independent train (80%) and test (20%) sets of complete samples. We generate a missing value indicator matrix $\mathbf{R}$ corresponding to the training data to introduce missing values. Random missing matrix $\mathbf{R}$ can be generated by using sampling strategies that can sample (i, j) index pairs to set $r_{\mathrm{i,j}} = 1$. Instead of the single-column strategy used in Section 4.3.1, here we experiment with three different sampling strategies. We call these (i, j) pair sampling strategies *amputation strategies*. Each amputation strategy ensures that the probability of incomplete samples in group s is proportional to γ_s. The amputation strategies are discussed below.

Table 4.4. Benchmark datasets, number of samples and features in each of them, and their corresponding sensitive attributes. Bold sensitive attributes are used in our experiments.

Dataset	Source	Number of Samples	Sensitive Attributes
COMPAS [3]	Law	5278	**sex**, race
Adult [77]	Census	48842	**race**, **sex**
FolkIncome [76]	Census	196,604	**AGEP**, **RAC1P**, **SEX**
German [38]	Finance	1000	**age**, **sex**
PIMA [78]	Healthcare	768	**Age**
Heart [79]	Healthcare	70,000	**age**, **gender**

Table 4.5. Group-wise summaries of changes in positive prediction probabilities and relative rankings of the individuals due to imputation mechanism on real-world benchmark datasets. The column definitions are same as discussed in Table 4.1.

Dataset	(s, y)	Probabilities			Ranks		
		$\delta_{\mathbb{P}} < 0$	$\delta_{\mathbb{P}} > 0$	$\Delta_{\mathbb{P}}$	$\delta_{rn} < 0$	$\delta_{rn} > 0$	Δ_{rn}
COMPAS	(u, -)	48.25	51.75	-1.8e-04	2.56	0.55	-0.46
	(u, +)	43.30	56.70	4.1e-04	4.08	2.25	-0.36
	(p, -)	43.42	56.58	7.0e-04	2.82	2.55	0.16
	(p, +)	33.52	66.48	1.0e-03	2.17	5.97	0.98
FolkIncome	(u, -)	72.74	27.26	-1.4e-03	43.37	54.60	-5.95
	(u, +)	73.24	26.76	-1.5e-03	40.09	59.25	-14.82
	(p, -)	66.76	33.24	-5.1e-04	27.39	71.39	20.11
	(p, +)	76.92	23.08	-1.5e-03	40.45	58.86	-13.33
German	(u, -)	33.85	66.15	-4.3e-04	22.31	22.31	-0.36
	(u, +)	37.31	62.69	1.3e-03	23.08	32.69	0.11
	(p, -)	30.85	69.15	1.0e-03	19.79	30.64	-0.03
	(p, +)	37.28	62.72	9.7e-04	24.56	28.42	0.03
PIMA	(u, -)	20.91	79.09	7.3e-03	29.09	35.45	0.05
	(u, +)	28.33	71.67	4.0e-03	28.33	29.17	0.08
	(p, -)	22.60	77.40	4.3e-03	30.52	28.75	-0.03
	(p, +)	30.00	70.00	6.1e-03	28.00	23.43	0.04
Heart	(u, -)	15.04	84.94	5.2e-03	33.45	64.40	0.40
	(u, +)	14.55	85.40	6.0e-03	35.66	60.87	-0.70
	(p, -)	16.71	83.08	4.5e-03	32.66	65.19	4.60
	(p, +)	16.37	83.63	5.7e-03	34.37	62.42	-6.89

- **Strategy 1:** Randomly sample row indices i with probability γ_s and randomly sample feature index j with probability $\frac{1}{d}$. Here, each sample can have at most 1 missing value.
- **Strategy 2:** The missing values only appear on the most correlated feature (constant j) with the class label. We randomly sample row indices i with probability γ_s.
- **Strategy 3:** The row indices i are sampled with γ_s and each feature is randomly picked with probability 0.5 (allowing multiple missing values per sample).

The amputation strategies are analogous to real-world situations. For example, college applicants may report either GRE scores or IELTS scores but not both. *Strategy 1* mimics such a scenario where the participants may randomly choose to hide one of the criteria. In `COMPAS` recidivism prediction, defendants could be willing to hide prior criminal records while being indifferent towards other information. This influences the design of *strategy 2* which induces missing values only on the most correlated feature. Finally, *strategy 3* replicates the overall poor quality of data from selected demographics.

Since *strategy 1* uniformly distributes the missing values over all the features, imputation is expected to have less effect on the feature distributions than with the other two strategies. Therefore, we use it to introduce missing values and study the changes in positive prediction probabilities and relative rankings due to imputation. We experiment with 10 random train-test splits of each dataset and their corresponding random R sampled using strategy 1. We report the means and standard deviations of the statistics of changes in predictions over these random splits in Table 4.5. Recall that, a higher positive prediction rate and lower ranking increases the likelihood of resource allocation to the related individual. In addition, the allocation of resources to the negative samples is also an undesired outcome. In COMPAS recidivism prediction, mean imputation disfavors the privileged by increasing the rankings of 5.97% of $(p, +)$ individuals as opposed to 2.25% of $(u, +)$. Similarly, compared to 2.56% of $(u, -)$ a higher portion of 2.82% of $(p, -)$ receives reduced ranking risking a higher false incarceration of the privileged. Unfair advantages are also observed when positive prediction probability is compared.

In `FolkIncome`, 71% of the samples in $(p, -)$ received an increased ranking after imputation compared to only 54% of samples in $(u, -)$. It implies that mean imputation favors

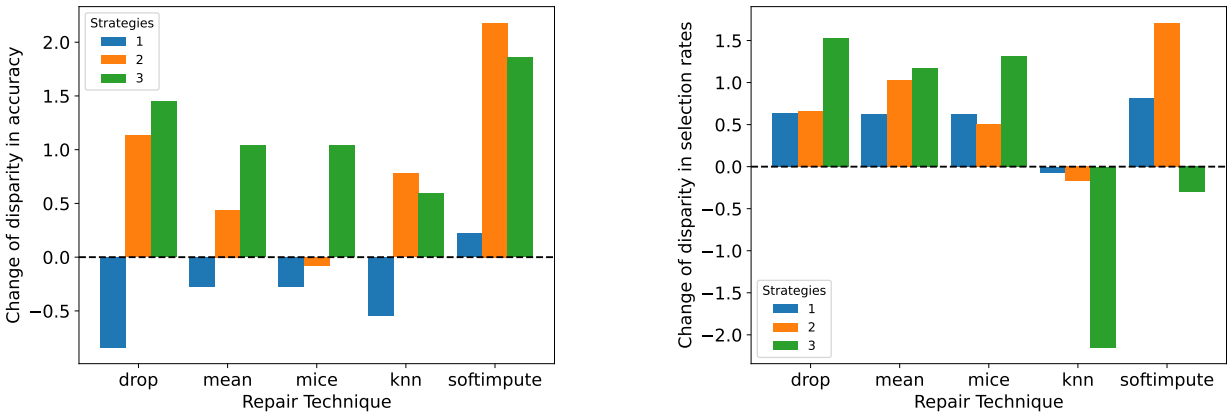
the negative individual in the privileged group by protecting them from undesired resource allocation. On the other hand, the group-wise mean changes of the rankings among the negative samples are close to each other. However, approximately 3% more individuals from the unprivileged positive samples received higher positive prediction probability than their privileged counterparts. An increased positive prediction probability among the positive samples indicates a higher likelihood of favorable decisions in the unprivileged group. Similar to the `COMPAS` dataset, a big portion of individual rankings remained unchanged after imputation. However, we observed increased rankings of 30% and 22% of samples in the negative privileged and unprivileged groups respectively. It suggests that imputation results in more unfair treatment towards the privileged group. Thus, in both `FolkIncome` and `German` datasets, we observe that imputation mechanisms introduce unfair shifts in the positive prediction probabilities and the relative ranking of the individuals.

Table 4.5 also reports that mean imputation disproportionately disfavors the $(u, +)$ in PIMA dataset by increasing the rankings of 6% more individuals than $(p, +)$ However, the changes in $t(\mathbf{x})$ are relatively close. To summarize, mean imputation improves the positive prediction probability for a slightly higher ($\sim$1%) portion of $(p, +)$ but it results in a much larger improvement in their rankings. Finally, compared to the other datasets, `Heart` shows the most balanced rates of changes among the privileged and unprivileged groups after imputation.

Next, we study whether the pattern of missing value occurrence impacts the bias introduced by the missing value handling techniques. In this regard, we compare the increase in disparities after dropping and imputation across the three amputation strategies. Table 4.6 and Figure 4.6 report the disparities after dropping or imputing missing values introduced using each of the three strategies. We observe that strategies 2 and 3 result in higher accuracy and false positive rate disparities compared to strategy 1. Among the imputation mechanisms, Softimpute introduces the highest accuracy and false positive disparity in conjunction with strategy 2. Strategy 2 removes the most predictive features from the unprivileged group at a higher rate than the privileged. Therefore, the prediction behavior worsens compared to the baseline and strategy 1. Strategy 3 induces a higher number of missing values than

Table 4.6. Disparities in group-wise accuracies, base prediction rates, and false positive rates of NBC classifiers on `PIMA` with missing value disparity induced using different strategies.

Strategies	Method	$SP(ACC;\theta)$	$SP(\theta)$	$SP(FPR;\theta)$
-	Baseline	7.72 (10.27)	-21.09 (11.25)	-15.96 (13.31)
1	Drop	6.88 (11.29)	-21.73 (13.26)	-16.21 (12.58)
	Mean	7.45 (11.50)	-21.71 (11.87)	-16.40 (13.91)
	MICE	7.45 (11.50)	-21.71 (11.87)	-16.40 (13.91)
	k-NN	7.18 (11.15)	-21.02 (10.15)	-15.42 (13.21)
	Softimpute	7.94 (11.70)	-21.91 (12.39)	-17.53 (16.15)
2	Drop	8.85 (10.41)	-21.75 (12.19)	-18.00 (15.53)
	Mean	8.16 (11.01)	-22.12 (13.33)	-18.51 (15.77)
	MICE	7.64 (10.44)	-21.60 (12.69)	-16.93 (13.12)
	k-NN	8.50 (12.51)	-20.92 (12.95)	-17.00 (17.28)
	Softimpute	9.90 (11.08)	-22.80 (13.19)	-20.74 (15.74)
3	Drop	9.17 (10.09)	-22.62 (13.23)	-19.16 (17.23)
	Mean	8.76 (11.00)	-22.26 (12.61)	-18.86 (15.45)
	MICE	8.76 (11.02)	-22.41 (12.75)	-18.96 (15.74)
	k-NN	8.31 (10.60)	-18.94 (11.63)	-13.93 (13.09)
	Softimpute	9.58 (8.09)	-20.79 (12.95)	-18.34 (15.53)



(a) Changes in disparity in accuracy.

(b) Changes in disparity in selection rates.

Figure 4.6. Changes in prediction rate disparities after repairing the amputated PIMA dataset with different imputation mechanisms where amputation was performed using several different amputation strategies.

both strategy 1 and 2. Although strategy 3 seems to hurt classifier performance more than strategy 1, strategies 2 and 3 show similar effects.

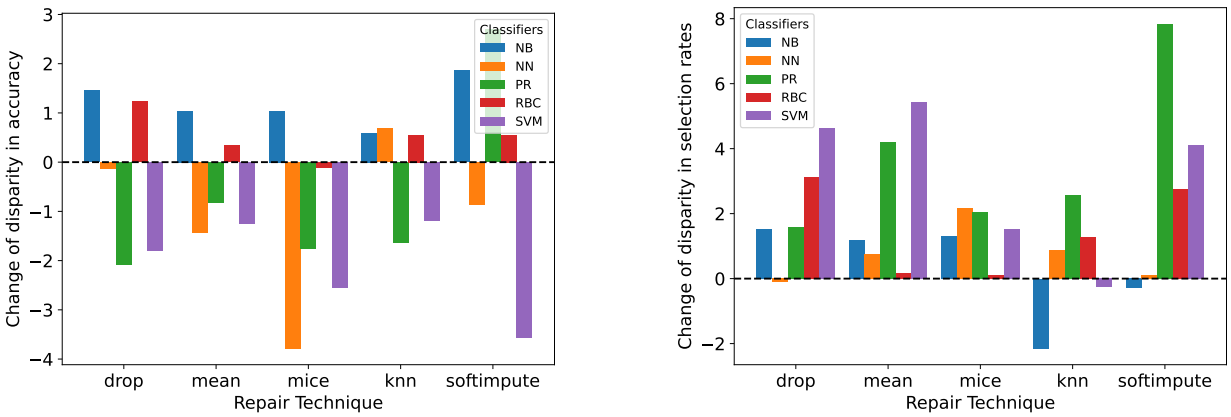
Finally, we study the impact of missing value disparity on state-of-the-art fairness interventions. We report the classifier performances on `FolkIncome` (Table 4.7) and `PIMA` (Table 4.8 and Figure 4.7) as representatives of large and small datasets with the highest and lowest amount of samples respectively among others. We apply amputation *strategy 3* since it introduces missing values at a higher rate. We report the performance of NBC, SVM, a shallow $(5 \times 5 \times 2)$ layer neural network (NN), and state-of-the-art fair classifiers such as prejudice remover (PR) [13] and reduction-based classifier (RBC) [26]. On `FolkIncome` dataset, both fairness unaware NBC and fairness aware classifiers show increased accuracy and selection rate disparities after most of the imputation mechanisms. The Softimpute imputation technique magnified the disparities introduced by the NB classifier most. The state-of-the-art classifiers induced the highest false positive rate disparities after Softimpute and k-NN imputation respectively. On `PIMA`, NBC performs worse after dropping mechanisms. The fairness-aware classifiers demonstrated the highest disparities after Softimpute and dropping mechanisms. Among the five imputations, SVM displays the lowest average absolute disparity in accuracy compared to the other classifiers on both datasets. However, it increases the prediction rate disparity or false positive disparity at a higher rate. To summarize, missing value disparity induced bias is observed from different types of classifiers.

Table 4.7. Disparities in group-wise accuracies, base prediction rates and false positive rates of different classifiers on `FolkIncome` with missing value disparity induced using strategy 3.

θ	Method	$SP(ACC;\theta)$	$SP(\theta)$	$SP(FPR;\theta)$
NBC	Baseline	3.30 (0.41)	-17.29 (0.60)	-12.25 (0.81)
	Drop	3.49 (0.48)	-16.81 (0.64)	-11.54 (0.93)
	Mean	3.11 (0.41)	-16.88 (0.60)	-11.63 (0.82)
	MICE	3.27 (0.44)	-16.92 (0.60)	-11.69 (0.87)
	k-NN	3.71 (0.56)	-18.36 (1.13)	-13.64 (1.65)
	Softimpute	5.76 (0.41)	-23.82 (0.53)	-20.76 (0.71)
SVM	Baseline	3.72 (9.26)	-0.32 (4.08)	-0.42 (1.97)
	Drop	1.90 (10.72)	-0.31 (3.10)	0.38 (1.48)
	Mean	3.66 (7.47)	-0.62 (4.40)	-0.20 (1.62)
	MICE	4.77 (7.49)	-2.55 (3.40)	-0.53 (1.60)
	k-NN	-1.19 (8.18)	1.97 (3.24)	0.74 (2.06)
	Softimpute	-1.04 (8.65)	1.20 (3.01)	0.56 (1.80)
NN	Baseline	11.42 (4.53)	-5.43 (3.26)	-2.15 (2.07)
	Drop	11.50 (4.37)	-4.00 (1.74)	-0.83 (1.05)
	Mean	10.98 (4.39)	-3.77 (1.30)	-0.53 (1.44)
	MICE	10.23 (3.97)	-5.26 (2.12)	-1.70 (1.66)
	k-NN	11.77 (4.75)	-3.36 (1.30)	-0.34 (1.28)
	Softimpute	13.79 (5.54)	-2.05 (2.05)	-0.12 (0.87)
PR	Baseline	5.23 (0.42)	-14.93 (1.01)	-9.35 (1.13)
	Drop	5.48 (0.39)	-15.15 (0.47)	-9.48 (0.63)
	Mean	5.18 (0.39)	-15.42 (0.66)	-10.05 (0.82)
	MICE	5.21 (0.38)	-15.30 (0.47)	-9.76 (0.65)
	k-NN	5.57 (0.41)	-17.36 (3.66)	-11.84 (3.96)
	Softimpute	5.36 (0.34)	-15.67 (0.41)	-9.78 (0.65)
RBC	Baseline	4.41 (0.52)	-15.94 (1.27)	-11.07 (1.98)
	Drop	4.49 (0.57)	-16.21 (1.41)	-11.73 (2.20)
	Mean	4.13 (0.46)	-15.61 (1.03)	-10.57 (1.48)
	MICE	4.45 (0.39)	-16.19 (1.33)	-11.63 (1.94)
	k-NN	4.32 (0.43)	-16.46 (1.28)	-11.63 (1.94)
	Softimpute	4.56 (0.50)	-17.65 (0.49)	-13.27 (0.78)

Table 4.8. Disparities in group-wise accuracies, base prediction rates and false positive rates of different classifiers on `PIMA` with missing value disparity induced using strategy 3.

θ	Method	$SP(ACC;\theta)$	$SP(\theta)$	$SP(FPR;\theta)$
NBC	Baseline	7.72 (10.27)	-21.09 (11.25)	-15.96 (13.31)
	Drop	9.17 (10.09)	-22.62 (13.23)	-19.16 (17.23)
	Mean	8.76 (11.00)	-22.26 (12.61)	-18.86 (15.45)
	MICE	8.76 (11.02)	-22.41 (12.75)	-18.96 (15.74)
	k-NN	8.31 (10.60)	-18.94 (11.63)	-13.93 (13.09)
	Softimpute	9.58 (8.09)	-20.79 (12.95)	-18.34 (15.53)
SVM	Baseline	-3.65 (9.63)	0.75 (10.06)	-4.41 (10.68)
	Drop	1.85 (10.72)	-5.39 (12.06)	-11.71 (13.54)
	Mean	2.41 (9.68)	-6.19 (12.89)	-8.09 (20.67)
	MICE	1.10 (8.86)	-2.27 (9.36)	-5.92 (11.70)
	k-NN	2.46 (13.43)	0.49 (16.36)	-6.69 (22.68)
	Softimpute	-0.09 (11.44)	4.86 (17.99)	-0.40 (25.40)
NN	Baseline	15.59 (13.91)	-0.46 (5.06)	-1.40 (4.59)
	Drop	15.46 (13.81)	0.36 (8.26)	-0.01 (13.78)
	Mean	14.16 (8.44)	-1.20 (3.20)	-0.51 (1.80)
	MICE	11.80 (14.86)	-2.62 (8.02)	0.57 (5.55)
	k-NN	16.28 (10.21)	1.34 (4.28)	0.22 (4.95)
	Softimpute	14.73 (10.89)	-0.56 (4.88)	0.10 (5.24)
PR	Baseline	11.50 (11.64)	-22.51 (11.90)	-20.22 (13.23)
	Drop	9.41 (11.23)	-24.08 (8.73)	-19.81 (11.12)
	Mean	10.68 (11.65)	-26.70 (12.01)	-23.17 (12.30)
	MICE	9.75 (11.55)	-24.57 (9.97)	-20.39 (11.87)
	k-NN	9.87 (12.83)	-25.09 (9.61)	-20.93 (13.66)
	Softimpute	14.20 (10.04)	-30.35 (14.23)	-30.69 (15.99)
RBC	Baseline	14.54 (8.55)	-26.59 (12.84)	-30.60 (15.59)
	Drop	15.77 (10.34)	-29.70 (9.44)	-34.49 (13.77)
	Mean	14.88 (11.31)	-26.77 (11.24)	-30.70 (14.93)
	MICE	14.42 (10.28)	-26.69 (10.48)	-30.18 (14.39)
	k-NN	15.09 (10.32)	-27.88 (11.68)	-31.42 (14.64)
	Softimpute	15.09 (11.58)	-29.34 (11.17)	-33.64 (15.49)



(a) Changes in disparity in accuracy.

(b) Changes in disparity in selection rates.

Figure 4.7. Changes in prediction rate disparities after repairing the amputated PIMA dataset with different imputation mechanisms and several different classifiers.

5. INTRODUCED BIAS THROUGH DISPARITY IN NOISE TOLERANCE

Machine learning model performance varies largely due to noisy samples in the dataset. Real-world datasets are collected from both reliable and unreliable resources. Unreliable sources such as social media, crowd-sourcing, etc., can result in noisy labels and annotations. On the other hand, tailor-made datasets are often published for specific fields of study such as medical diagnosis [80], readmission [81], etc. However, they are not noise-free due to uncertainty in measurements, human errors, and misclassification (i.e., misdiagnosis). The prevalence of noisy samples is a threat to model performance. Traditional machine learning algorithms were designed to avoid noise in data to achieve high generalization accuracy. However, complex deep learning mechanisms are more prone to overfit noise in the training data and are also known to be susceptible to small perturbations in the input. Therefore, the disparity in group-wise rates of noise or the tolerance of noise among the sensitive groups can lead to disproportionate behavior from deep neural networks.

Real-world datasets can contain both noisy features and noisy ground-truths. Label noise can result from crowd-sourced annotation by uninitiated annotators, and even from experts' mistakes in challenging tasks. For example, restaurant ratings collected from Yelp [82] vary significantly among the participants since they are untrained critics of the service. Similarly, illusive medical diagnosis can cause specialist doctors to misdiagnose [80]. On the other hand, noisy samples can appear naturally due to uncertainty in measurements, or be intentionally introduced by adversaries to fool the machine learning model. Such intentionally perturbed noisy samples are called adversarial samples. Since DNNs typically learn a discontinuous input-output mapping, Szegedy et al. [83] showed that it is possible to efficiently generate adversarial samples with small additive noise, ϵ, to the input. Subsequent studies proposed numerous methods, commonly known as evasion attacks, to generate adversarial samples [84–88].

The performance of a machine learning model against noisy samples is called its robustness. In deep learning, a more popular notion is adversarial robustness which indicates the performance of a classifier against adversarially perturbed samples generated by an adver-

sarial attack. To achieve robustness against adversarial attacks, several defense mechanisms have been proposed in the literature. The adversarial training mechanism proposed by [85] is regarded as one of the most successful defenses against adversarial perturbations. They proposed feeding the adversarial samples into the training process. This technique is also the foundation of several other defenses [84, 89, 90]. Several adversarial training proposed robustness-aware regularizations [91, 92]. Among them TRADES [91] achieves best performance by balancing the trade-off between standard and robust accuracy. Other defense mechanisms highlights the use of external data [93–95] (often generated [96, 97]), special architecture [98], etc., to improve adversarial robustness.

In this chapter, we show that prominent adversarial training methods tend to exacerbate the disparities in the model predictions. In an attempt to improve the robustness of a machine learning model, adversarial training ends up sacrificing model fairness. We hypothesize that bias amplification from adversarial training is due to the disparity in noise tolerance among the protected groups. That means, one group is more noise-tolerant than the others and the adversarial training mechanisms improve the performance of the noise-tolerant group at a higher rate than the vulnerable groups. To address this issue we propose a re-weighted objective function that draws more attention towards the group with lower noise tolerance or more vulnerable to adversarial attack. Our proposed methods show that re-weighting improves the fairness of state-of-the-art adversarial training techniques while sacrificing minimal adversarial robustness.

Furthermore, we analyze the impact of label noise on the fairness of deep neural networks. Unlike noisy input features, which make it challenging to maintain prediction accuracy, noisy ground-truths make the training task more challenging to achieve high generalization performance from the model. Recently it was reported that the over-parameterized neural networks are expressive enough to learn highly accurate models on noisy ground-truths. In particular, [99] showed that deep neural networks demonstrate a *double-descent* phenomenon in test error with model complexity. That means that as the neural network model becomes sufficiently large and is trained sufficiently long, it starts to improve generalization performance. Although increasing model complexity beyond the interpolation region improves the overall accuracy of the joint model, it is still unclear whether it also improves the fairness of the

classifiers. As a result, in this chapter, we further analyze the relations between disparity in group-wise prediction rates and model complexity under different levels of label noises.

5.1 Fairness Against Adversaries

In this section, we introduce the types of attacks and their threat models. We discuss established evasion attacks and proposed defenses against them. We also define the robust-fairness metrics used in our analysis. Finally, we conclude this section by discussing the

5.1.1 Adversarial Attacks

Notable attacks against ML models are evasion attacks, poisoning attacks, extraction attacks, and inference attacks. Poisoning attacks involve altering the training data to degrade the model performance. The goal of a model extraction attack is to extract or deduce the parameters of a black-box target model. A model attribute inference attack is a privacy attack that focuses on inferring individuals' sensitive attributes from a machine learning model. Finally, in evasion attacks, an adversary perturbs the input example to force misclassification from the model. Let, the dataset $\mathcal{D} = \{\mathbf{x}, y, \mathbf{s}\}_{i=1}^{N}$ where $\mathbf{x} \in \mathbb{R}^d$ (or $\in [0,1]^{m \times n}$ in case of grayscale images) is an input example, $y \in \{1, \ldots, k\}$ is the target attribute, and each $s \in \mathbf{s}$ is a sensitive attribute such as race, gender, etc. $f : \mathbb{R}^d \rightarrow \{1, \ldots, k\}$ is the classifier. We assume "fairness-through-unawareness", i.e., the classifier is unaware of the sensitive attributes of the samples. The parameterized version of the classifier function can be written as $f(x; \theta)$ where θ indicates the classifier parameters. Let the prediction from the classifier be $\hat{y} = f(x; \theta)$. The objective of an evasion attack is as follows,

$$\text{Find} \quad x_{adv} \quad \text{s.t.} \quad f(x_{adv}) \neq y \quad \text{and} \quad || \, x_{adv} - x \, || \leq \epsilon$$

Here, $|| \cdot ||$ indicates the norm. Since a majority of adversarial attacks are proposed on image samples, in the remainder of this work we assume $x, x_{adv} \in [0,1]^{m \times n}$.

Evasion attacks can be both white-box attacks [84–86, 100] and black-box [87] attacks. In addition, in the literature, both targeted and untargeted attacks are proposed. Untargeted

attacks simply focus on forcing a misclassification whereas targeted attacks alter the input samples in a way such that it is misclassified as a pre-specified incorrect class. We introduce several well-known evasion attacks below.

L-BFGS attack

Szegedy *et al.* [83] formulates the process as finding the smallest additive perturbation, δ that causes a misclassification as,

$$\begin{aligned} &\text{minimize} \quad || \ \delta \ || \\ \text{s.t.} \quad &f(x+\delta) = y' \neq y \qquad \text{and} \quad x_{adv} \in [0,1]^{m \times n} \end{aligned} \tag{5.1}$$

The box constraint in Equation 5.1 makes it difficult to find an exact solution. Instead, they propose an approximation of δ using box-constrained L-BFGS. Let $l(x, y)$ indicate the loss function. The following problem is solved to produce an adversarial sample.

$$\text{minimize} \quad c \ || \ \delta \ || + l(x+\delta, y') \quad \text{s.t.} \quad x + \delta \in [0,1]^{m \times n}$$

To find the adversarial sample closest to x, the smallest value of $c > 0$ is picked that forces $f(x+\delta) = y' \neq y$. Line search is used to determine the optimal value of c.

Gradient Descent Attack

Biggio *et al.* [101] proposed gradient descent optimization of the following objective function to find the adversarial sample in a binary classification setting where $y \in \{0, 1\}$. Here, the predicted label $\hat{y} = f(x) = \mathbb{1}(g(x) > 0)$.

$$x^0_{adv} = \arg\min_x \left(g(x) - \lambda \mathbb{P}(x \mid \hat{y} = 1)\right) \ \text{s.t.} \ \ || \ x^0_{adv} - x \ || < \epsilon \tag{5.2}$$

Here, x^0_{adv} indicates adversarial samples where $\hat{y} = 0$ but $y = 1$. The density function in Equation 5.2 is added to ensure that x_{adv} is within the support of $\mathbb{P}(x)$. During optimization,

the density function is approximated by the kernel density estimator (KDE). Finally, after each descent step x_{adv} is clipped within a ϵ distance from x.

Fast Gradient Sign Method (FGSM)

The fast gradient sign method (FGSM) is one of the most prominent evasion attacks due to its efficiency in generating adversarial samples. According to [84], adversarial samples can be computed by perturbing x as follows,

$$x_{adv} = x + \epsilon sign(\nabla_x l(x, y)) \tag{5.3}$$

The perturbation moves the sample in the direction of the gradient which increases the loss. After perturbation, the pixels of adversarial images are clipped back to the $[0, 1]$ range. Equation 5.3 indicates that it is an l_∞-bounded adversary, i.e., $|| \ x_{adv} - x \ ||_\infty \leq \epsilon$. A small perturbation may appear benign but due to the high dimensionality of the neural networks its impact is amplified and forces misclassification. FGSM is popular as the fastest adversarial sample generation technique.

Projected Gradient Descent (PGD)

Although FGSM is a fast adversarial attack, the adversarial sample may not be the strongest one. In this regard, Madry *et al.* [85] proposed an iterative noise addition mechanism that improves the adversarial sample at each iteration. At each step, perturbation is performed following 5.3. Then, the perturbed sample is projected back to the $x + \mathcal{S}$ where $\mathcal{S}$ is the set of allowed perturbations. For example, for an l_∞ bounded adversary $\mathcal{S}$ is an ϵ ball around x. Formally,

$$x^{t+1} = \prod_{x+\mathcal{S}} \Big(x^t + \alpha sign(\nabla_x l(x, y) \Big) \tag{5.4}$$

Basic Iterative Method (BIM)

Basic iterative method (BIM) or iterative fast gradient sign method (IFGSM) [100] is a multi-step adversarial technique similar to PGD. The l_∞ bounded BIM adversary, clamps the perturbed values within ϵ distance from x. Moreover, PGD allows starting the iteration from any random point within $x + \mathcal{S}$ whereas BIM starts with perturbation x.

Carlini and Wagner (CW) L_2 and L_∞ attack

Carlini and Wagner [86] proposes three adversarial attacks with L_2, L_0 and L_∞ adversaries. To satisfy the box-constraint in Equation 5.1, they parameterize the adversarial sample as,

$$x_{adv} = \frac{1}{2}(\tanh(w) + 1) \tag{5.5}$$

The L_2 adversary chooses a $y' \neq y$ and searches for w such that it solves the following minimization.

$$\begin{aligned} \text{minimize} \quad & || \frac{1}{2}(\tanh(w) + 1) - x ||_2 + cZ\left(\frac{1}{2}(\tanh(w) + 1)\right) \\ \text{where} \quad & Z(x) = \max(\max(\{g(x)_{\mathrm{i}} : \mathrm{i} \neq y'\}) - g(x)_{y'}, -\kappa) \end{aligned} \tag{5.6}$$

Here, $g(x)$ indicates the logits of the classifier. The untargeted L_2 adversary instead defines $Z(x)$ as, $Z(x) = \max(\max(g(x)_y - \{g(x)_{\mathrm{i}} : \mathrm{i} \neq y\}), -\kappa)$. Similarly, the L_∞ adversary solves Equation 5.6 with the L_∞ norm instead of L_2 with a few tweaks to prevent oscillation of gradient descent.

Others

Several other attacks have been proposed in literature such as variants of FGSM [90, 102–105] and PGD [87], DeepFool [88], Universal Attack Perturbation (UAP) [106], etc. Recently, an ensemble of parameter-free attacks, combinedly known as *AutoAttack* [87], is proposed as a benchmark of robustness evaluation in *RobustBench* [87]. The ensemble includes two untargeted attacks (APGD and Square) and two targeted attacks (APGDT and FAB^T) For

the sake of brevity, we omit discussion on other attack mechanisms and refer to a survey of adversarial attacks [107].

5.1.2 Defenses Against Adversarial Attacks

Although many defenses against adversarial samples have been proposed in recent years, using more powerful and adapted attacks most of them could be broken. Typical adversarial training methods feed adversarial samples back to the model to improve robust accuracy. The adversarial training proposed by Madry *et al.* [85] is widely accepted as the most effective defense mechanism. We refer to this adversarial training process as *Madry*. They defined the adversarial training objective as,

$$\underset{\theta}{\text{minimize}} \quad E_{(x,y)\sim\mathcal{D}}\left[\max_{\delta\in\mathcal{S}} l(x+\delta, y; \theta)\right] \tag{5.7}$$

It indicates that the goal of adversarial training is to achieve the lowest expected loss over the adversarial samples. A similar formulation was proposed by Goodfellow *et al.* [84] where the joint minimization objective was defined as follows for $\alpha \in [0, 1]$,

$$\underset{\theta}{\text{minimize}} \quad \alpha \underset{(x,y)\sim\mathcal{D}}{E} l(x, y; \theta) + (1-\alpha) \underset{(x,y)\sim\mathcal{D}}{E} l(x+\delta, y; \theta) \tag{5.8}$$

Here, $x + \delta$ is the adversarial sample generated using FGSM attack. We refer to this adversarial training process as *Goodfellow*. Note that, when $\alpha = 0$, *Goodfellow* is essentially *Madry* against FGSM attack which was later studied in [90] as “Free" adversarial training. Training against single-step attacks offers a faster adversarial training mechanism. Using FGSM instead of PGD offers a faster adversarial training alternative. However, subsequent studies found that *Madry* with FGSM achieves robustness by overfitting against single-step attacks but fails against multi-step attacks such as PGD. As a result, variations of *Madry* were proposed in [90, 105]. Wong *et al.* [90] suggested that *Madry* with random initialization and FGSM is as effective *Madry* with multi-step attacks such as PGD. Other studies proposed improving the diversity of adversarial samples to reduce overfitting [105].

Adversarial regularization techniques are also proposed to improve model robustness. *TRADES* [91] the most notable adversarial regularization technique. In this work, the authors studied the trade-off between robustness and accuracy. They decomposed the robust error, i.e., error against adversarial samples, into two parts. They are the misclassification error on the clean samples and the boundary error. Based on the boundary error a regularization term is proposed which shifts the model boundary away from the data points.

Both adversarial training and adversarial regularization methods are highly dependent on the produced adversarial samples. Therefore, they might be susceptible to improved attacks that exploit the weaknesses of the defense mechanisms. As a result, researchers have proposed defense mechanisms with theoretical guarantees of robustness [108, 109]. However, certified defense mechanisms are accompanied by higher computation costs. In addition, these mechanisms are yet to achieve high performance on large datasets. As a result, we omit their details. Finally, adversarial training can be interpreted as augmenting the dataset with adversarial samples. As a result, several recent works introduced unlabelled additional samples to improve adversarial robustness [93, 94]. Since the availability of additional samples may be limited, artificially generated samples are also used to improve robustness [96].

5.1.3 Fair-Robustness Metrics

In this section, we discuss the metrics used in assessing the model's fairness concerning its group-wise robustness against adversarial samples.

Robust Statistical Parity

We study the statistical parity of the prediction rates of the robust classifiers. That means we extend the idea of statistical parity in the rates of robust predictions $\hat{y}_{adv} = f(x_{adv})$. we define the robust statistical parity as,

$$SP_{rob}(\mathcal{D}) = \mid \mathbb{P}(\hat{y}_{adv} = 1 \mid p) - \mathbb{P}(\hat{y}_{adv} = 1 \mid u) \mid \tag{5.9}$$

Other fairness notions in FairML literature require equality of different performance statistics, such as equalized accuracy, equalized odds (equality of group-wise false positive and false negative rates), etc. Following the formulation in Equation 5.9, we can similarly define robust equalized accuracy and robust equalized false positive rates as follows,

$$SP_{rob}(\mathcal{D}, acc) =\mid \mathbb{P}(\hat{y}_{adv} = y \mid p) - \mathbb{P}(\hat{y}_{adv} = y \mid u) \mid \tag{5.10}$$

$$SP_{rob}(\mathcal{D}, fpr) =\mid \mathbb{P}(\hat{y}_{adv} = 1 \mid y = 0, p) - \mathbb{P}(\hat{y}_{adv} = 1 \mid y = 0, u) \mid \tag{5.11}$$

Robustness Bias

Nanda *et al.* [37] defined robustness bias as one group lying significantly closer to the decision boundary than the others. That means, given a noise threshold $\tau > 0$, one subgroup is more vulnerable to the attacks than the others. If $d_f(\mathbf{x})$ is the minimal distance of $\mathbf{x}$ from the decision boundary of f, given a binary sensitive attribute s, the robustness bias,

$$RB(s, \tau) = \quad \mid \mathbb{P}(d_f(\mathbf{x}) > \tau \mid p, y = \hat{y}) - \mathbb{P}(d_f(\mathbf{x}) > \tau \mid u, y = \hat{y}) \mid \tag{5.12}$$

This definition can be extended to non-binary sensitive attributes by replacing $s = p$ and $s = u$ by $s = s'$ and $s \neq s'$ for $s' \in \text{supp}(s)$. Equation 5.12 is highly dependent on the choice of τ. As an alternative, [37] proposes the following metric for any sensitive group $s' \in \text{supp}(s)$.

$$\hat{I}_{s'}(\tau) = \frac{\mid (x, y) : d_f(x) > \tau, y = \hat{y}, s = s' \mid}{\mid (x, y) : s = s' \mid} \tag{5.13}$$

$$\sigma(s') = \frac{(AUC)(\hat{I}_{s'}) - (AUC)(\hat{I}_{s \neq s'})}{(AUC)(\hat{I}_{s \neq s'})} \tag{5.14}$$

Since it is challenging to determine $d_f(x)$ for high-dimensional neural network boundaries, [37] proposed an upper bound for Equation 5.13 using x_{adv} samples produced by attacks such as CW.

$$\hat{I}^{CF}_{s'}(\tau) = \frac{\mid (x,y) : \tau \leq \mid\mid x - x_{adv} \mid\mid, y = \hat{y}, s = s' \mid}{\mid (x,y) : s = s' \mid} \tag{5.15}$$

5.1.4 Related Works

Recently, the interaction between fairness and robustness has received attention in the literature. For example, Roh et al. [29, 110, 111] defined robustness as accuracy against noisy samples. The authors of [111] used a sanitized validation set to obtain fairer models. Instead of an additional validation set, [29] modeled the problem as a bi-level optimization where the outer optimization balances a fairness constraint and the inner level optimizes for model accuracy. Finally, [110] proposed a batch-wise re-balancing of noisy samples across the sensitive groups to improve fairness and accuracy against noisy input. Although these studies investigated the interactions between random noisy samples and fairness, they do not consider robustness against tailor-made adversarial samples which are most prone to degrade the model performance. The interactions between adversarial robustness and fairness of ML models have recently gained attention [37, 112–115]. Zeng *et al.* [114] and Mehrabi *et al.* [116] proposed data poisoning attacks against model fairness and defense against such poisoning. Since poisoning attacks modify the training data, it is expected to directly influence the model performance. However, we investigate the more subtle impact of evasion attacks on model fairness. In this regard, [112, 113, 117] reported that adversarial training against evasion attacks led to a class-wise disparity in robust accuracy. In [113], it is suggested that adversarial training emphasizes the easier classes to improve robustness and amplifies class-wise disparity in the robust accuracy. Moreover, stronger adversarial training leads to larger class-wise robustness accuracy disparity [112]. Finally, [37] defined robustness bias as the closeness of each group to the decision boundary. They reported a group-wise robustness accuracy disparity against evasion attacks due to the disparity in proximity to the

Table 5.1. Dataset used in the experiments with a corresponding number of samples, target attributes, and sensitive attributes.

Dataset	# samples	Target	Sensitive Attributes
UTKFace	20K	Age	Race, Gender
RAFDB	30K	Emotions	Race, Gender, Age
FairFace	108K	Age	Race, Gender
Celeba	202K	Attractive, Smiling	Gender

decision boundary of the samples in each group. None of the previous studies investigated the influence on group fairness while aiming for adversarial robustness against evasion attacks.

5.1.5 Experimental Results

We experiment with four facial image datasets; UTKFace[1], RAFDB[2], FairFace[3], and Celeba[4]. Their sizes, the target attributes, and the sensitive attributes are listed in Table 5.1. Their sizes vary from 20K to 200K. Celeba contains only binary class labels whereas the others include multi-class target attributes. Similarly, gender is typically a binary sensitive attribute while race and age are non-binary attributes.

For simplicity and ease of comparison, we analyze the fairness of a binary classifier between privileged and unprivileged groups, i.e., binary sensitive attributes. Our analysis focuses on gender-based disparities where, without loss of generality, we assume *Male* and *Female* as the privileged and the unprivileged group. We map the multi-class target attributes into favorable and unfavorable classes. The *Age* classes in the UTKFace and FairFace dataset that correspond to age 40 or above are considered unfavorable classes and the others as favorable. Finally, we assume the positive emotion labels, i.e., *Surprise* and *Happiness*, in RAFDB are favorable classes.

We examine the adversarial training algorithms introduced in section 5.1.2, *Goodfellow*, *Madry*, and *TRADES*. Both *Goodfellow* and *Madry* feed the adversarial samples back to

[1]↑https://susanqq.github.io/UTKFace/
[2]↑http://www.whdeng.cn/raf/model1.html
[3]↑https://github.com/joojs/fairface
[4]↑https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

the training process. We also experiment with different adversarial attacks to generate adversarial samples during training. We apply each method in conjunction with one of FGSM, BIM, PGD, and CW(l_2) attacks to generate adversarial samples during the training separately. Since AutoAttack significantly increases the training time, following [115], we limit its scope to model evaluation only. In addition to AutoAttack, we analyze the disparities against FGSM, BIM, PGD, and CW as well.

We pick Resnet-18 [118] architecture to analyze robustness disparities. For faster training, we use the parameters pre-trained on ImageNet for initialization. For each dataset, the baseline models are obtained without adversarial training where a Resnet-18 architecture is trained with Adam [119] optimizer for 30 epochs and 10^{-4} learning rate. Using the *Goodfellow* adversarial training strategy, we obtain four separate robust models where the training adversarial sample generation was dictated by FGSM, BIM, PGD, and CW respectively. We refer to these models as *(Goodfellow, attack)* where *attack* is the corresponding attack used in the training process. Similarly, the combination of *Madry* adversarial training with four different adversarial attacks yields four separate robust models each of which is denoted as *(Madry, attack)* respectively. In our experiment, each multi-step adversarial attack is allowed to run for 10 iterations, and the maximum adversarial noise $\epsilon = 8/255$ (except CW). We use the same hyper-parameters as a baseline for training both *Goodfellow* and *Madry*. Since we observed slower convergence from *Madry* than others, a longer training (epochs=50) for *Madry* is chosen. Finally, we apply *TRADES* algorithm to produce a robust classifier where the hyper-parameters are set following [91]. Interested readers are referred to the shared repository https://github.com/rakinhaider/FairnessUnderAttack/ for further experimental details.

Robust Statistical Parity

In this section, we compare the disparities in robust prediction rates of the adversarially trained classifiers with the baseline classifiers against a range of attacks. In our analysis, we use the robust statistical parity metrics defined in Equations 5.10 and 5.11. Table 5.2 (Figure 5.1) and 5.3 report the statistical parities of robust accuracies and false positive rates

Table 5.2. Disparities (absolute differences) in group-wise robust accuracy against each adversarial attack with and without adversarial training.

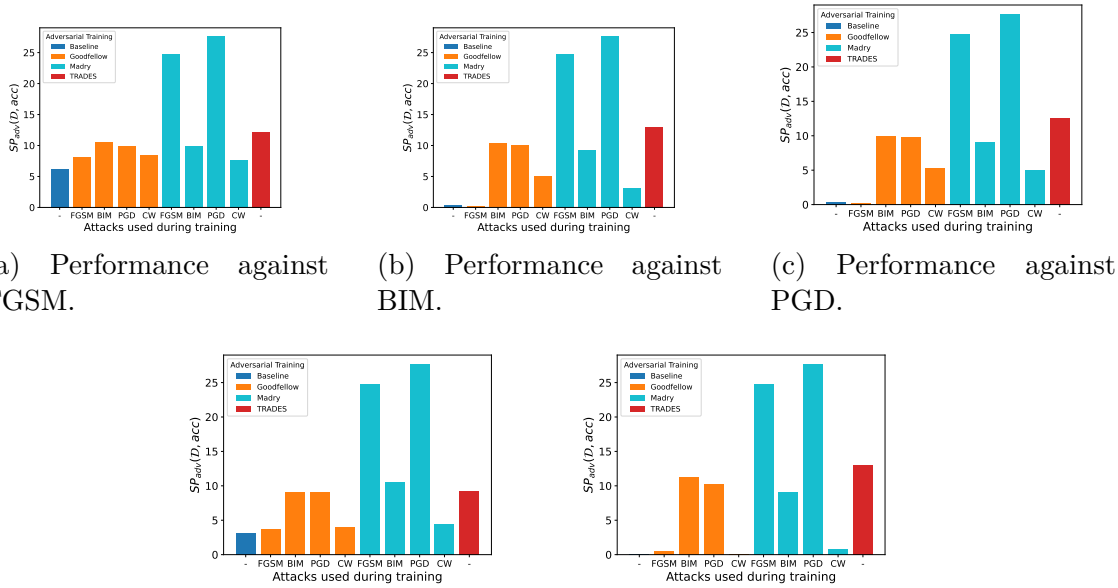
Dataset	Method	$\mid SP_{rob}(ACC;\theta) \mid$				
		FGSM	BIM	PGD	CW	AutoAttack
UTKFace	Baseline	6.2	0.4	0.4	3.1	0.0
	(Goodfellow, FGSM)	8.2	0.2	0.3	3.7	0.5
	(Goodfellow, BIM)	10.6	10.3	9.9	9.1	11.2
	(Goodfellow, PGD)	10.0	10.1	9.8	9.0	10.2
	(Goodfellow, CW)	8.5	5.1	5.2	4.0	0.0
	(Madry, FGSM)	24.7	24.7	24.7	24.7	24.7
	(Madry, BIM)	9.9	9.3	9.1	10.6	9.0
	(Madry, PGD)	27.6	27.6	27.6	27.6	27.6
	(Madry, CW)	7.7	3.2	5.0	4.4	0.8
	TRADES	12.2	12.9	12.6	9.3	13.0
RAFDB	Baseline	5.7	0.0	0.1	1.1	0.0
	(Goodfellow, FGSM)	9.6	0.1	0.0	1.6	0.0
	(Goodfellow, BIM)	10.0	11.6	11.0	10.3	10.8
	(Goodfellow, PGD)	9.4	10.5	10.6	9.3	10.1
	(Goodfellow, CW)	10.0	2.5	3.2	4.4	0.4
	(Madry, FGSM)	9.1	10.6	10.8	11.4	7.8
	(Madry, BIM)	8.9	8.8	8.9	8.5	8.79
	(Madry, PGD)	9.2	9.4	9.4	9.2	9.4
	(Madry, CW)	8.8	3.4	5.2	7.5	1.0
	TRADES	9.8	10.7	10.6	8.4	10.7
FairFace	Baseline	0.8	0.0	0.0	1.3	0.0
	(Goodfellow, FGSM)	3.7	0.0	0.0	0.8	0.0
	(Goodfellow, BIM)	0.8	8.0	8.2	2.9	0.0
	(Goodfellow, PGD)	2.5	1.1	3.0	4.7	0.0
	(Goodfellow, CW)	2.3	0.0	0.1	5.8	0.0
	(Madry, FGSM)	4.3	0.0	0.0	0.1	0.0
	(Madry, BIM)	9.7	9.8	9.7	9.9	10.8
	(Madry, PGD)	9.6	9.6	9.6	9.7	9.8
	(Madry, CW)	2.1	0.1	0.0	3.4	0.0
	TRADES	10.4	7.8	8.2	10.7	1.3
Celeba	Baseline	4.8	0.1	0.1	4.2	0.0
	(Goodfellow, FGSM)	4.0	4.5	0.5	22.5	0.2
	(Goodfellow, BIM)	4.8	5.3	5.3	0.6	5.3
	(Goodfellow, PGD)	3.8	4.0	3.9	1.3	4.0
	(Goodfellow, CW)	11.1	3.2	4.1	1.4	2.8
	(Madry, FGSM)	11.5	18.1	18.0	20.9	12.5
	(Madry, BIM)	12.7	12.8	12.8	6.7	12.9
	(Madry, PGD)	11.0	11.3	11.3	5.7	11.3
	(Madry, CW)	19.5	5.8	7.8	3.5	4.7
	TRADES	3.3	3.2	3.1	2.1	3.2

Table 5.3. Disparities (absolute differences) in group-wise false positive rates against each adversarial attack with and without adversarial training.

Dataset	Method	$\mid SP_{rob}(FPR;\theta) \mid$				
		FGSM	BIM	PGD	CW	AutoAttack
UTKFace	Baseline	13.7	4.6	8.5	10.0	9.6
	(Goodfellow, FGSM)	11.0	8.7	10.7	10.2	2.5
	(Goodfellow, BIM)	8.6	8.3	9.3	8.2	5.6
	(Goodfellow, PGD)	9.8	9.0	9.2	9.2	7.7
	(Goodfellow, CW)	7.5	6.4	7.1	7.2	7.4
	(Madry, FGSM)	21.4	21.4	21.4	21.4	21.4
	(Madry, BIM)	2.4	2.0	2.1	1.9	2.3
	(Madry, PGD)	4.8	4.8	4.8	4.8	4.8
	(Madry, CW)	8.2	4.7	4.9	7.4	6.3
	TRADES	4.6	4.6	4.7	7.1	4.2
RAFDB	Baseline	3.5	0.1	0.5	1.8	0.8
	(Goodfellow, FGSM)	3.4	3.0	1.0	2.9	2.7
	(Goodfellow, BIM)	1.5	1.4	1.2	3.0	1.4
	(Goodfellow, PGD)	2.4	0.8	1.0	3.7	1.7
	(Goodfellow, CW)	2.4	0.9	2.8	0.6	2.4
	(Madry, FGSM)	1.7	1.0	2.3	2.3	1.5
	(Madry, BIM)	1.0	1.0	1.0	1.4	0.9
	(Madry, PGD)	1.9	2.1	2.1	1.7	2.1
	(Madry, CW)	1.5	1.8	0.9	4.8	1.1
	TRADES	3.7	5.6	4.9	5.1	5.5
FairFace	Baseline	13.4	9.8	14.2	14.0	14.0
	(Goodfellow, FGSM)	15.8	7.5	10.1	13.0	8.7
	(Goodfellow, BIM)	4.3	3.9	3.5	9.9	3.6
	(Goodfellow, PGD)	8.5	6.7	7.1	11.1	6.4
	(Goodfellow, CW)	14.6	12.3	13.0	17.0	12.6
	(Madry, FGSM)	18.4	11.4	11.1	16.8	13.1
	(Madry, BIM)	0.0	0.0	0.0	0.0	0.0
	(Madry, PGD)	0.2	0.2	0.2	0.1	0.2
	(Madry, CW)	14.5	12.2	12.3	18.8	12.1
	TRADES	0.0	0.0	0.0	0.1	0.0
Celeba	Baseline	3.0	0.2	0.1	0.1	0.0
	(Goodfellow, FGSM)	35.7	5.3	5.9	27.1	0.3
	(Goodfellow, BIM)	36.1	35.3	35.3	38.5	35.2
	(Goodfellow, PGD)	37.2	36.8	36.9	39.1	36.7
	(Goodfellow, CW)	29.0	5.4	7.3	38.2	5.0
	(Madry, FGSM)	23.1	26.2	26.3	23.9	15.7
	(Madry, BIM)	28.9	28.7	28.7	32.9	28.7
	(Madry, PGD)	29.8	29.5	29.5	32.8	29.4
	(Madry, CW)	33.8	7.5	10.4	38.1	6.0
	TRADES	15.4	15.5	15.5	16.3	15.5

Table 5.4. Disparities (absolute differences) in group-wise robust selection rates against each adversarial attack with and without adversarial training.

Dataset	Method	$\mid SP_{rob}(\theta) \mid$				
		FGSM	BIM	PGD	CW	AA
UTKFace	Baseline	**18.0**	14.5	**14.4**	**19.5**	10.2
	(Goodfellow, FGSM)	15.1	15.0	11.7	19.2	11.6
	(Goodfellow, BIM)	9.1	9.5	9.6	7.0	8.6
	(Goodfellow, PGD)	9.3	9.8	9.9	7.7	9.6
	(Goodfellow, CW)	11.6	**15.3**	13.8	15.2	10.1
	(Madry, FGSM)	11.7	11.7	11.7	11.7	**11.7**
	(Madry, BIM)	0.5	0.3	0.5	0.2	0.3
	(Madry, PGD)	2.7	2.7	2.7	2.7	2.7
	(Madry, CW)	11.3	11.9	11.3	14.9	9.7
	TRADES	6.8	8.1	7.9	6.1	8.3
RAFDB	Baseline	4.3	1.7	2.5	0.1	1.4
	(Goodfellow, FGSM)	4.8	4.1	4.4	1.1	4.7
	(Goodfellow, BIM)	5.6	6.3	**6.2**	**5.7**	**5.5**
	(Goodfellow, PGD)	4.6	6.0	6.0	3.9	4.6
	(Goodfellow, CW)	**7.5**	0.2	2.0	1.8	1.2
	(Madry, FGSM)	3.8	**4.9**	4.1	5.4	2.1
	(Madry, BIM)	0.6	0.5	0.6	0.8	0.2
	(Madry, PGD)	1.5	1.6	1.6	1.3	1.6
	(Madry, CW)	4.4	2.2	0.2	1.1	5.1
	TRADES	3.0	2.8	2.9	1.6	2.9
FairFace	Baseline	7.9	6.7	6.4	8.9	6.1
	(Goodfellow, FGSM)	11.8	7.8	**9.2**	11.3	**10.6**
	(Goodfellow, BIM)	5.6	6.0	5.7	9.2	5.4
	(Goodfellow, PGD)	7.3	6.2	6.5	8.7	7.0
	(Goodfellow, CW)	11.3	**8.2**	7.8	**13.1**	6.5
	(Madry, FGSM)	**11.9**	7.4	7.4	11.2	8.6
	(Madry, BIM)	0.0	0.0	0.0	0.0	0.0
	(Madry, PGD)	0.1	0.1	0.1	0.1	0.2
	(Madry, CW)	10.7	6.3	6.0	13.6	7.3
	TRADES	0.1	0.1	0.0	0.0	0.0
Celeba	Baseline	6.3	16.1	16.1	6.7	16.1
	(Goodfellow, FGSM)	**49.9**	26.1	29.4	5.2	**40.2**
	(Goodfellow, BIM)	37.4	35.8	35.9	46.4	35.7
	(Goodfellow, PGD)	38.3	**36.8**	**36.9**	**47.4**	36.7
	(Goodfellow, CW)	1.0	34.4	32.3	42.3	34.8
	(Madry, FGSM)	23.0	0.6	1.7	5.8	20.4
	(Madry, BIM)	29.2	28.6	28.6	36.1	28.5
	(Madry, PGD)	29.9	29.3	29.4	36.6	29.2
	(Madry, CW)	0.7	33.6	30.9	51.1	34.9
	TRADES	14.1	13.5	13.5	24.5	13.3



(a) Performance against FGSM.

(b) Performance against BIM.

(c) Performance against PGD.

(d) Performance against CW.

(e) Performance against AutoAttack.

Figure 5.1. Disparities in group-wise robust accuracy with and without adversarial training on UTKFace dataset.

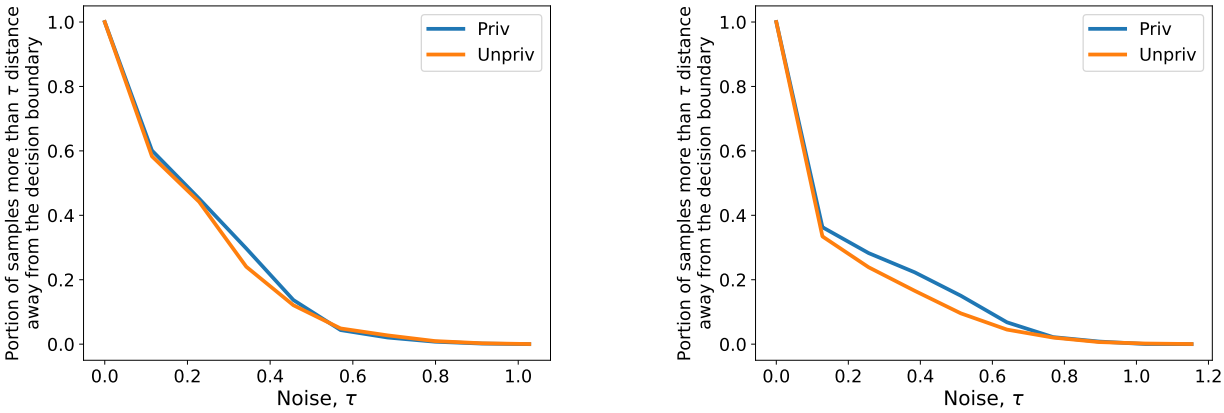
of each classifier against five state-of-the-art attacks. Except *(Goodfellow, FGSM)* classifier, adversarial training exacerbated the disparity in robust accuracy on the UTKFace dataset irrespective of the adversarial attack. The maximum disparity in robust accuracy is observed from *(Madry, PGD)* classifier. It shows a steady 27.6% disparity against all of the attacks. Among the 45 combinations of UTKFace classifiers and adversarial attacks (nine variations of robust classifiers each analyzed against five adversarial attacks), we observed an increased disparity in robust accuracy compared to the baseline in 42 of them. That means the adversarial training on the UTKFace dataset exacerbates the robust accuracy approximately 93% of the time. Furthermore, adversarial training increased the disparity in false positive rates but it is observed less frequently compared to the group-wise robust accuracy disparity. Only 15 of the 45 (33%) combinations of the UTKFace classifiers and adversarial attacks produced higher false positive rate disparity than the baselines. Similarly, except for a single classifier, adversarial training amplifies the disparity in group-wise emotion prediction

accuracy using the RAFDB dataset. Here, *(Goodfellow, BIM)* and *(Madry, FGSM)* against AutoAttack and FGSM respectively demonstrate the maximum disparity in robust accuracy (10.8%) whereas *TRADES* against AutoAttack demonstrates the highest disparity in robust false positive rates. About 95% and 75.5% of the RAFDB robust classifier and attack combinations resulted in an increased robust accuracy disparity and robust false positive rate disparity respectively.

The adversarially trained classifiers on the FairFace dataset often performed poorly against stronger attacks such as AutoAttack and often resulted in very low ($\sim 0\%$) group-wise accuracy. Otherwise, the robust classifiers on both FairFace and Celeba demonstrate bias amplification concerning accuracy and false positive rates. In FairFace and Celeba, the disparity in robust accuracy is increased in 66.7% (30 of 45) and 80% (36 of 45) of the classifier and attack combinations. On larger datasets, we observed that *Goodfellow* algorithm contributes less towards bias amplification but more effective defenses such as *Madry* and *TRADES* tend to increase bias at a higher rate. In addition, robust Celeba classifiers increased false positive disparity significantly higher than the baseline. Finally, we observe little evidence of selection rate disparity (defined in Equation 5.9) amplification by adversarial training. Further details on selection rate disparity can be found in Table 5.4.

Robustness Bias

Nanda *et al.* [37] defines the robustness bias curve as the $\hat{I}_s^{CF}(\tau)$ by τ. A lower robustness bias curve indicates that a larger portion of samples are susceptible to smaller noises. As a result, σ_s, in Equation 5.14, compares the areas under the robustness bias curve (AUC) to determine whether there exists a disparity in group-wise vulnerability to adversarial attacks. We analyze the absolute values of $\sigma(s)$ since we focus more on the increase in disparity instead of its direction. Figures 5.2a and 5.2b show robustness bias curves without and with adversarial training respectively. We observe that the gap between the group-wise robustness bias curves is increased by adversarial training mechanisms. This is made further evident in Table 5.5 where we compare the disparities in group-wise AUCs of adversarial training with the baseline models on each dataset. We observe that the group-wise disparity



(a) Without adversarial training.

(b) With adversarial training *(Madry, CW)*.

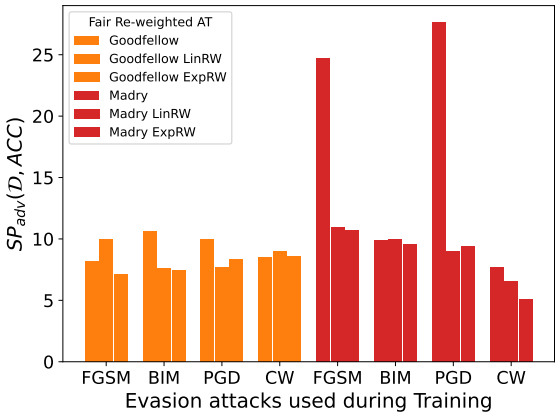
Figure 5.2. Robustness bias curves with and without adversarial training on UTKFace dataset.

Table 5.5. Magnitudes of disparity in area under robustness bias curve ($abs(\sigma(p))$) in each dataset with and without adversarial training.

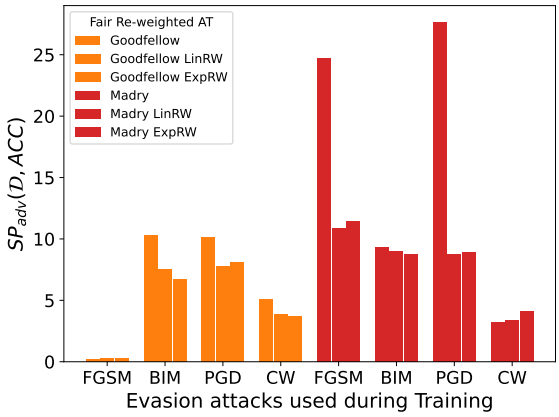
Method	Magnitudes of $\sigma(p)$			
	UTKFace	RAFDB	FairFace	Celeba
Baseline	0.043	0.200	0.031	0.074
(Goodfellow, FGSM)	**0.058**	0.151	0.016	**0.222**
(Goodfellow, PGD)	**0.076**	0.074	**0.058**	0.001
(Goodfellow, CW)	**0.125**	0.153	**0.059**	**0.177**
(Goodfellow, BIM)	**0.069**	0.089	0.014	0.002
(Madry, FGSM)	**0.321**	0.108	**0.104**	**0.209**
(Madry, PGD)	**0.443**	0.009	0.006	0.001
(Madry, CW)	**0.148**	0.067	0.008	**0.075**
(Madry, BIM)	0.027	0.018	**0.037**	0.001
TRADES	**0.084**	0.079	**0.282**	0.024

in the UTKFace dataset is enhanced by every adversarial training method in consideration except *(Madry, BIM)*. Adversarial training enhances this disparity on large-scale datasets about 50% of the time. Notably, RAFDB demonstrates the maximum baseline disparity between group-wise AUCs, and this disparity is greatly alleviated by the adversarial training processes.

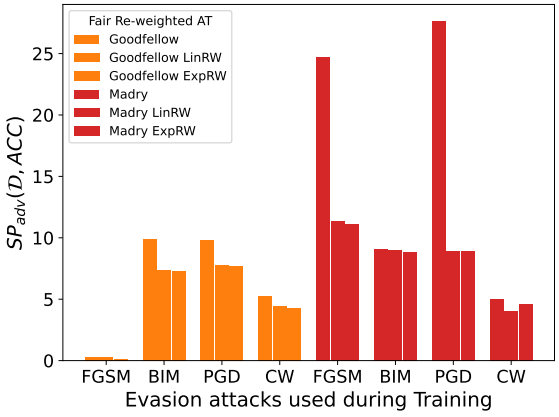
5.1.6 Fair Re-weighted Adversarial Training



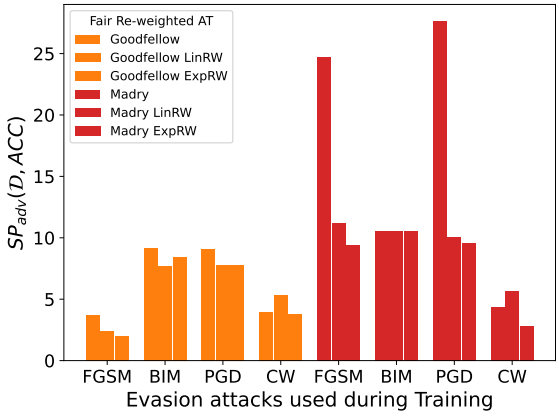
(a) Disparities in robust accuracy against FGSM.



(b) Disparities in robust accuracy against BIM.



(c) Disparities in robust accuracy against PGD.



(d) Disparities in robust accuracy against CW.

Figure 5.3. Disparities in robust accuracy with and without re-weighting (linear and exponential).

Table 5.6. Difference of group-wise robust accuracies against adversarial attack after adversarial training without and with re-weighting.

Method	Weights	$\mid SP_{rob}(ACC;\theta) \mid$				Robust Accuracy			
		Test Attack				Test Attack			
		FGSM	BIM	PGD	CW	FGSM	BIM	PGD	CW
(Goodfellow, FGSM)	-	8.2	0.2	0.3	3.7	51.2	**2.0**	**2.4**	**8.9**
	LinRW	10.0	0.2	0.2	2.4	51.6	1.0	1.3	6.1
	ExpRW	**7.1**	0.2	**0.1**	**1.9**	**52.7**	1.2	1.6	6.6
(Goodfellow, BIM)	-	10.6	10.3	9.9	9.1	53.2	50.4	50.9	**57.6**
	LinRW	7.6	7.5	7.4	**7.7**	53.9	50.5	51.1	57.0
	ExpRW	**7.4**	**6.7**	**7.2**	8.4	**54.2**	**51.3**	**51.5**	57.0
(Goodfellow, PGD)	-	10.0	10.1	9.8	9.0	52.8	49.1	49.5	**57.6**
	LinRW	**7.7**	**7.7**	7.8	7.8	**54.0**	**51.3**	**51.8**	57.2
	ExpRW	8.3	8.0	**7.7**	**7.7**	53.6	50.8	51.3	57.4
(Goodfellow, CW)	-	**8.5**	5.1	5.2	4.0	**42.7**	10.5	12.4	29.9
	LinRW	9.0	3.8	4.4	5.3	42.2	9.1	11.5	**26.7**
	ExpRW	8.6	**3.7**	**4.2**	**3.8**	42.2	**9.0**	**11.1**	**26.7**
(Madry, FGSM)	-	24.7	24.7	24.7	24.7	54.7	**54.7**	**54.7**	**54.7**
	LinRW	10.9	**10.9**	11.3	11.1	**54.2**	28.2	30.7	34.1
	ExpRW	**10.7**	11.4	**11.1**	**9.4**	54.3	34.2	36.3	38.8
(Madry, BIM)	-	9.9	9.3	9.1	10.6	**55.3**	**54.7**	54.7	57.0
	LinRW	9.9	9.0	9.0	**10.5**	55.2	54.4	54.6	57.0
	ExpRW	**9.5**	**8.7**	**8.8**	**10.5**	**55.3**	**54.7**	**54.8**	**57.2**
(Madry, PGD)	-	27.6	27.6	27.6	27.6	**55.6**	**55.6**	**55.6**	55.6
	LinRW	**9.0**	**8.7**	**8.**9	10.0	55.1	54.2	54.2	57.1
	ExpRW	9.4	8.9	**8.9**	**9.5**	55.0	54.3	54.4	**57.2**
(Madry, CW)	-	7.7	**3.2**	5.0	4.4	**44.8**	**13.1**	**14.7**	**32.0**
	LinRW	6.6	3.4	**4.0**	5.6	43.7	11.7	13.8	29.9
	ExpRW	**5.1**	4.1	4.6	**2.8**	43.5	12.0	13.5	28.2

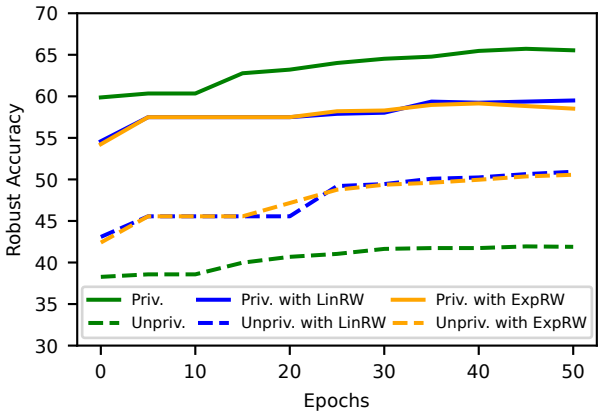


Figure 5.4. Group-wise learning curves of *(Madry, PGD)* classifiers with and without re-weighting.

In this section, we discuss a method to reduce bias amplification by adversarial training while maintaining the robustness of the classifier against attacks. To reduce this group-wise disparity, we propose assigning higher emphasis on the samples that are members of the group that is most prone to misclassification subject to adversarial attacks. In other words, we propose re-weighting the samples at a rate that is proportional to their corresponding group-wise error rates. Here, we use the group-wise average loss as a proxy for group-wise error rates. Let's assume that the group-wise average loss at an iteration is $\mathcal{L}^s_{adv} = E[l(\mathbf{x}_{adv}, y) \mid s]$. Instead of using the average adversarial loss $\mathcal{L}_{adv} = E[l(\mathbf{x}_{adv}, y)]$ during back-propagation, we propose the application of the weighted sum of the group-wise adversarial losses as shown below,

$$\mathcal{L}_{adv} = \sum_s w^s N\mathbb{P}(s)\mathcal{L}^s_{adv} \tag{5.16}$$

At each iteration, we first compute the average group-wise adversarial losses, $\mathcal{L}^s_{adv}$. We compute the weight, w_{i} corresponding to the i^{th} sample, $(\mathbf{x}_{\mathrm{i}}, y_{\mathrm{i}}, s_{\mathrm{i}})$ as,

$$w_{\mathrm{i}} = \frac{\mathcal{L}^{s_{\mathrm{i}}}_{adv}}{N\mathbb{P}(p)\mathcal{L}^p_{adv} + N\mathbb{P}(u)\mathcal{L}^u_{adv}} \tag{5.17}$$

We refer to this technique as Linear Re-weighting (LinRW). To put further emphasis on the group with higher loss, we can exploit the Exponential Re-weighting (ExpRW) technique where each w_{i} is as follows,

$$w_{\mathrm{i}} = \frac{exp(\mathcal{L}_{adv}^{s_{\mathrm{i}}})}{N\mathbb{P}(p)\mathrm{e}xp(\mathcal{L}_{adv}^{p}) + N\mathbb{P}(u)\mathrm{e}xp(\mathcal{L}_{adv}^{u})} \tag{5.18}$$

Note that, both of the re-weighting strategies guarantee that $\sum_{\mathrm{i}}^{N} w_{\mathrm{i}} = 1$. At each iteration, the w_{i}s are considered constants and not updated during the back-propagation step. Figure 5.4 shows the learning curves of the *(Madry, PGD)* with and without the re-weighting. Re-weighting minimizes the gap between the group-wise loss and eventual error rates. Table 5.6 and Figure 5.3 shows the performance of LinRW and ExpRW with *Goodfellow* and *Madry* on UTKFace dataset. We observe that the reweighting techniques reduce the disparity in robust accuracy while maintaining equally high robustness against adversarial samples. Except *(Goodfellow, CW)*, re-weighting achieves a lower disparity than the corresponding adversarial training irrespective of the attack. Notably, the disparities from *(Madry, PGD)* and *(Madry, FGSM)* classifiers were reduced by a factor of two by re-weighting. Re-weighting often improves the robustness of the classifier but sacrifices the robustness significantly when trained against weaker attacks such as *(Goodfellow, FGSM)* and *(Madry, FGSM)*.

5.2 Fairness under Noisy Labels

The classical bias-variance trade-off theory states that models with higher complexity have lower bias and higher variance. Hence, after the model complexity is sufficiently increased it starts overfitting the training data and deteriorates generalization performance. However, large-scale neural networks with millions of parameters, often enough to even fit random labels [118], are known to outperform smaller models in many challenging tasks. To explain the improved generalization of large neural networks it was shown that generalization performance follows a double-descent phenomenon with increasing model complexity. It suggests that although the under-parameterized regime follows the bias-variance trade-off, in the over-parameterized regime, classifier performance improves with model complexity.

In this section, we compare the fairness of the under-parameterized and the over-parameterized regimes. It means we investigate how disparities in group-wise prediction rates change with model complexity.

5.2.1 Deep Double Descent

The deep double descent phenomenon is frequently observed in a variety of tasks, architectures, and optimization techniques. Nakkiran et al. [99] proposed an explanation of deep double descent using *effective model complexity* (EMC). Given a training procedure $\mathcal{T}$ and a set of samples $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ sampled from a distribution $\mathbb{P}$, let the resulting classifier be denoted by $\mathcal{T}(\mathcal{D})$. EMC is defined as the maximum number of samples n required to achieve approximately zero training error using $\mathcal{T}$. Let $Error(\mathcal{T}(\mathcal{D}))$ indicate the mean error of the model $\mathcal{T}(\mathcal{D})$ on training set $\mathcal{D}$. Formally, [99] defined EMC for a classification small error $\epsilon > 0$ as follows,

$$EMC_{\mathbb{P},\epsilon}(\mathcal{T}) = \max\{n : E_{\mathcal{D}\sim\mathcal{P}^n}[Error(\mathcal{T}(\mathcal{D}))] < \epsilon\} \tag{5.19}$$

The definition of EMC divides the training task with training data containing n samples into three regimes.

- **Under-parameterized regime:** If $EMC_{\mathbb{P},\epsilon}(\mathcal{T})$ is sufficiently smaller than n, then increasing EMC decreases test error.
- **Over-parameterized regime** In the over-parameterized refime, $EMC_{\mathbb{P},\epsilon}(\mathcal{T})$ is sufficiently larger than n. In this regime, increasing the EMC also decreases test error.
- **Critically parameterized regime:** When $EMC_{\mathbb{P},\epsilon}(\mathcal{T}) \approx n$, it is called the critical regime. In this regime, increasing the EMC typically hurts the generalization performance.

Equation 5.19 defines the joint effective model complexity of a classification task. However, given a sensitive attribute s the classification could be much more complex for one group than the other. Therefore, we can define the group-wise effective complexity as follows,

$$EMC_{\mathbb{P},\epsilon}(\mathcal{T}, s) = \max\{n : E_{\mathcal{D}_s \sim \mathbb{P}^n|s}[Error(\mathcal{T}(\mathcal{D}_s))] < \epsilon\} \tag{5.20}$$

Here, $\mathcal{D}_s$ is a set of samples from the conditional distribution of $\mathbb{P} \mid s$. The group-wise EMC can be used to measure the disparity in the complexity of the classification task within each group. The group-wise disparity in effective model complexities poses a severe dilemma in model selection since increasing model complexity could simultaneously degrade and improve generalization performance for different groups.

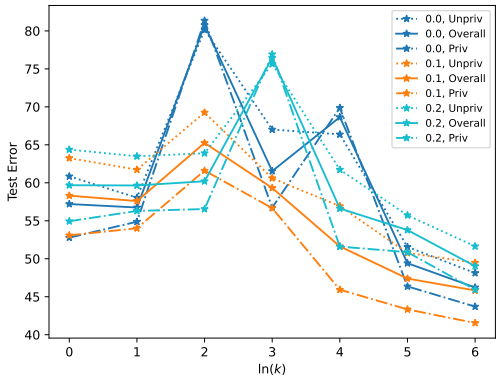
EMC of a given classification task is dependent on a variety of model choices such as the number of training epochs, the number of parameters in the classifier, and the inherent complexity of the classification task. In addition, increased label noise tends to increase EMC as it increases the challenges of the training process. That means, given a classification task and fixed training hyper-parameters, EMC can also be influenced by the rate of label noise. Here, we empirically study the fairness implications concerning varying EMCs. We study the changes in disparities in model predictions with increased model complexity and label noise. Furthermore, we study whether there exists a disparity in effective model complexities in traditional classification tasks.

5.2.2 Experimental Results

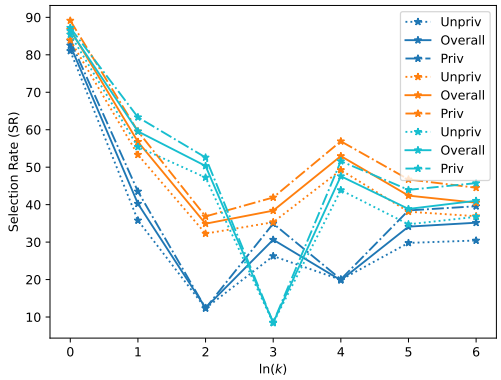
In our experiments, we investigate group-wise model performances on facial image classification tasks. We use RAFDB adm UTKFace datasets in our experiments. The RAFDB dataset is used to train classifiers to identify one of 7 emotions from a facial image. On the other hand, the classification task in the UTKFace dataset is to classify individuals into their corresponding *Age* groups. The age of the individuals in this dataset spans from 0 to 105. For simplicity, the *Age* attribute is quantized into 6 bins of 20-year intervals. Both RAFDB and UTKFace contain *gender* of the corresponding individual in the image. In addition to Male and Female annotations, RAFDB *gender* attribute also includes *unknown* for individuals who were difficult to identify as either Male or Female. In our analysis, we use *gender* as the sensitive attribute where *Male*s are considered as privileged individuals and others as unprivileged. A prediction of young (less than 40 years) is considered a favorable/ positive

prediction in the UTKFace. Similarly, the favorable/positive predictions in the RAFDB dataset are the positive emotions such as *Surprise* and *Happiness*.

To experiment with model complexity, we train Resnet18 architecture by varying the width (number of filters) of the convolution layers as suggested in [99]. That means, the layers widths of the Resnet18 architecture are parameterized by k and set to $[k, 2k, 4k, 8k]$ respectively where $k \in \{1, 2, 4, 8, 16, 32, 64\}$. Each Resnet18 architecture is trained with Adam optimizer with a learning rate of 0.0001 and 200 epochs. We further experiment with different rates of label noise, $p \in \{0.0, 0.1, 0.2\}$. Here, $p = 0.0$ indicates the clean dataset whereas the others indicate that we randomly assign incorrect labels with probability p (only once at the beginning of training; not per epoch).



(a) Group-wise error rates vs model complexity.



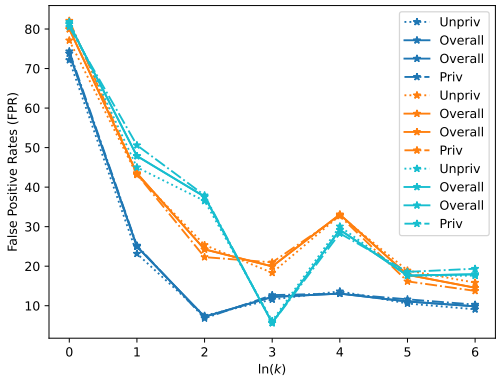
(b) Group-wise selection rates vs model complexity.

Figure 5.5. Error rates and selection rates with varying model complexity.

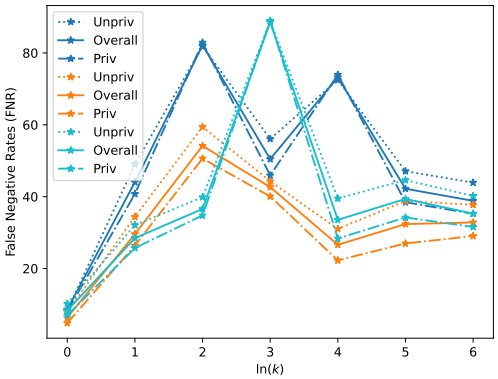
RAFDB

Figures 5.5 and 5.6 show the changes in group-wise model performance with variations in model complexity. Similar to the results in [99], we observe that not only the overall accuracy but also the group-wise accuracies satisfy the double descent phenomenon with increasing model complexity. However, we often observe a gap between group-wise model accuracies in both under-parameterized and over-parameterized regimes. In figure 5.5a we observe that the training process reaches the critical regime near $k = 4$, 4, and 8 for label

noise $p = 0.0$, 0.1, and 0.2 respectively. We observe minimal disparity in group-wise false positive rates across the model complexities in Figure 5.6a. However, except at the critical regime, a relatively higher disparity in group-wise selection rates and false negative rates are observed. Notably, least selection rate disparity near the critical regions. In short, although increasing model complexity beyond the critical regime monotonically improves accuracy, it doesn't show monotonic improvement in group-wise disparities.



(a) Group-wise false positive rates vs model complexity.

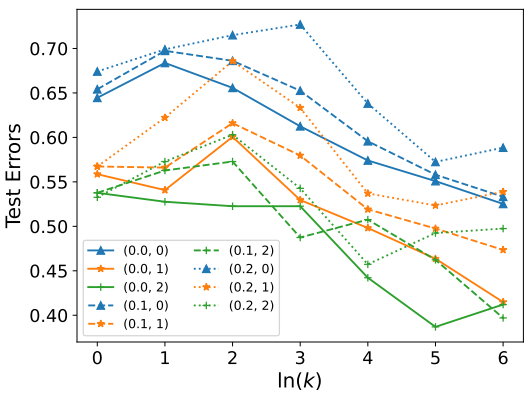


(b) Group-wise false positives rates vs model complexity.

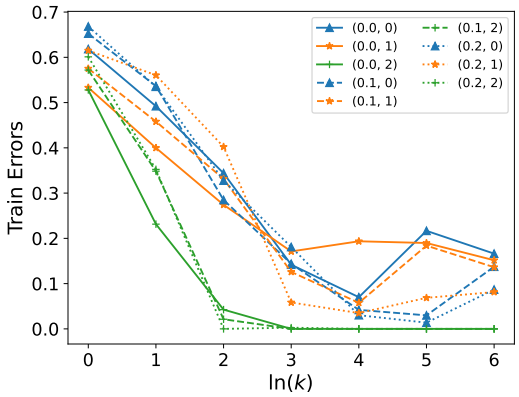
Figure 5.6. False positive and negative rates with varying model complexity.

To identify disparity in group-wise effective model complexities, we train group-wise models from the RAFDB dataset and analyze their performances on the corresponding group of individuals. In Figure 5.7, we denote the *Male*, *Female*, and *Unknown* group by 0, 1, and 2 respectively. In Figure 5.7a, we observe that when $p = 0.0$ the *Male*, *Female* and *Unkown* group reaches the critical regime near $k = 2$, 4, and 8 respectively. That means, inside this region increasing model complexity simultaneously improves the generalization performance of the Male individuals but might degrade performance for the others. We further observe that increasing label noise moves the critical regimes disproportionately among the three groups. For example, for label noise probability $p = 0.2$, the three groups reach the critical regimes near $k = 8$, 4, and 4 respectively. Here, the classifier on *Male* individuals reaches the critical regime later than the other two groups. Finally, analyzing the training errors we find that the classifier on individuals with *unknown* gender, achieves close to zero training

error at a faster rate than the others. It indicates that the effective model complexity of the classification task within the *unknown* gender group could be smaller than the others. As a result, given a training procedure $\mathcal{T}$ any changes to increase the effective model complexity could simultaneously degrade the generalization performance on one and improve the other. These results highlight the importance of proper model complexity selection to obtain the most fair classifier.



(a) Test errors from group-wise models on RAFDB.



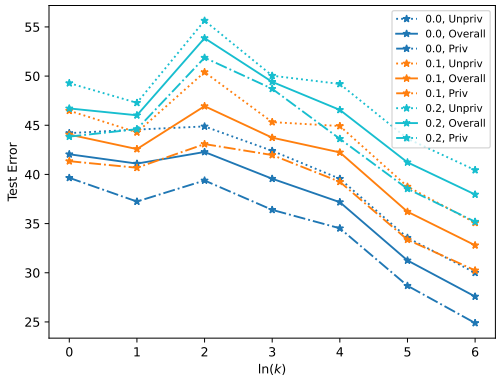
(b) Train errors from group-wise models on RAFDB.

Figure 5.7. Train and test errors from group-wise models on RAFDB.

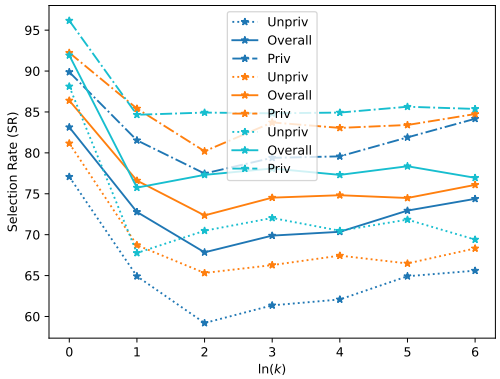
UTKFace

Similar to RAFDB, we observe a double descent phenomenon in the UTKFace classification task in Figure 5.8a. However, the disparities in group-wise performance don't show monotonicity with increasing model complexity.

Finally, we analyze the training and test errors of group-wise models to unveil evidence of group-wise disparity in effective model complexities. Figure 5.10a indicates that for each p both the *Male* (0) and *Female* (1) groups reach the critical regime at a similar model complexity. In contrast, the train errors in Figure 5.10b show that the group-wise classifier on the *Female* individuals tends to attain approximately zero training error faster than the other group-wise model. Therefore, we conclude that in UTKFace the disparity in group-wise EMC is minimal, if any.

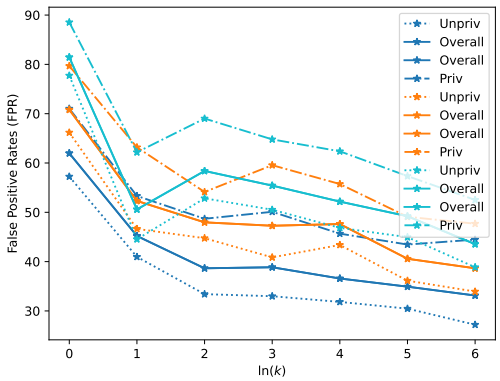


(a) Group-wise Error Rates vs Model Complexity.

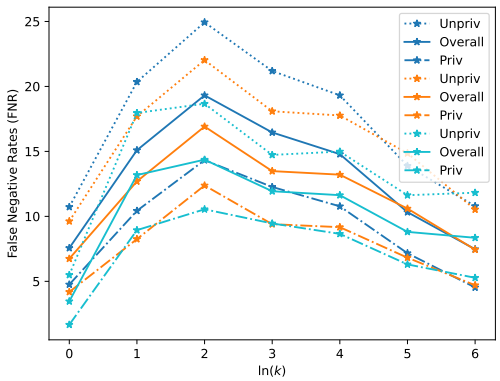


(b) Group-wise Selection Rates vs Model Complexity.

Figure 5.8. Error Rates and Selection Rates with varying model complexity.

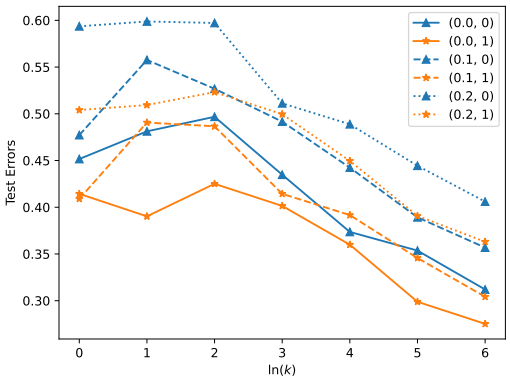


(a) Group-wise False Positive Rates vs Model Complexity.

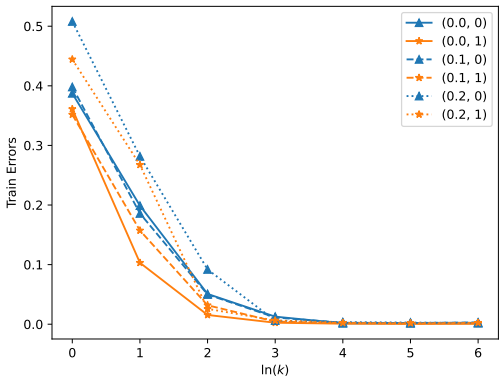


(b) Group-wise False Positives Rates vs Model Complexity.

Figure 5.9. False Positive and Negative rates with varying model complexity.



(a) Test errors from group-wise models on UTKFace.



(b) Train errors from group-wise models.

Figure 5.10. Train and test errors from group-wise models on UTKFace.

6. CONCLUSION

The rampant development of artificial intelligence and machine learning systems in recent years has made autonomous decision-making systems inextricably linked to our daily lives. Decisions made by an autonomous system undergo the same lens of scrutiny as the ones made by a human agent. Therefore, the fairness of machine learning systems has been brought to attention in recent literature. The foundation of the majority of the studies on the fairness of machine learning systems lies in the conventional wisdom that biased ground truths in the training data lead to biased models. The overarching goal of this thesis is to establish that label bias is not a necessary condition of unfairness in machine learning and identify other sources of biases that contribute to introducing bias in model outcomes.

The main contribution of this thesis is identifying several sources of biases in model predictions other than label bias. The studied instances typically arise from often unavoidable development choices during data collection or training. As a result, these introduced biases can be labeled as systemic bias which is defined as setting up the system, policy, and/or institution in a way to inherently support unfair outcomes by disadvantaging one group more than the others. While label bias-induced unfair predictions are simply an end-product of including discriminatory patterns in training, systemic biases are the driving forces that lead to unfairness in machine learning. Identification of such sources of biases unveils limitations of machine learning under current development practices and could lead to model development with theoretical fairness upper bounds. Therefore, we study the scopes of systemic discrimination through disparately effective feature sets, the disparity in group-wise rates of missing values, the disparity in group-wise noise tolerance, and the disparity in group-wise classification complexity. We investigate the conditions where such discrimination can cause unfair outcomes. This thesis lays the groundwork to identify sources of bias, encourage discussions on systemic bias in FairML, and develop benchmarks for pre-deployment audits of autonomous decision-making systems.

First, we demonstrate *model*-induced bias, as opposed to label-induced bias. We show that a Bayes-optimal classifier can be expected to induce biases in the outcome that are otherwise absent in the data. Experimental results validate that if group-wise optimal model

accuracies for demographic groups are different, the joint optimal Bayesian model trained on a fair dataset demonstrates disparate impact. We show that disparity in group-wise accuracy can arise from a disproportionately predictive feature set. It is tempting to address this by using separate models for different groups, but this may violate ethical and legal standards (e.g., U.S. civil rights laws, E.U. GDPR Article 9). A second approach is to optimize for fairness rather than accuracy [26], as in [120] and many more recent works. We suggest that a better approach is to eliminate the underlying disparity, using methods such as participatory design to assess [121] and propose better predictive features for all.

Secondly, we investigate missing value disparity induced bias. We show that, under the influence of missing value disparity, missing value handling techniques shift the model towards group-wise disparities. In particular, we show that missing value disparity is expected to move the class-wise classifier distributions closer to one group than the other. Our theoretical analysis established that mean imputation on disproportionate group-wise rate of missing values, results in optimal Bayesian classifiers that represents one demographic sub-group in each class better than the others. Experiments on synthetic datasets shows that missing value disparity induced bias extends further in other imputation mechanisms as well. Although the aftermath of imputation doesn't always change the prediction for an individual, usually it unfairly promotes or demotes the individual in ranking. Therefore, in resource-constrained scenarios such as college admission prediction disparity in missing value rates in training data leads to bias negatively impacting groups with higher missing data rates. We also report that missing value disparities can further exacerbate existing unfairness in real-world datasets. Our findings suggest developers of high-stake decision-making systems must carefully asses the changes in model behavior due to missing value treatments.

Finally, we show empirical evidence that disparity in group-wise tolerance to noise can be detrimental to model fairness. The experimental results suggest that established adversarial training mechanisms often exacerbate the disparity in robustness between groups of individuals. We observe significant increases in the difference between group-wise robust accuracy and false positive rates introduced by adversarial training on facial image classification tasks. To mitigate this increased disparity, we propose a fair re-weighting technique to improve disparity in adversarially trained robust models. Our proposed method significantly

reduces the disparity introduced by adversarial training against iterative adversarial attacks without sacrificing the robustness of the overall models. Since both robustness and fairness are pressing concerns of autonomous decision-making systems, a conflict between them highlights that model choices to optimize for robustness can sacrifice fairness or introduce biases.

We envision multiple future directions of deeper investigations of introduced bias. For example, we studied disparity among a set of numeric features. However, in computer vision, the input samples can be in the form of images or videos. It is worth identifying the types of visible or subtle group-wise disparities in the inputs that lead to dissimilarities in group-wise learned representations. Similarly, the recent trend of large language models can also be investigated as they are trained on unregulated publicly available web-crawled data. Existing disparities in internet usage among racial sub-groups are likely to be captured in the text corpus. Since text streams are rarely annotated with sensitive attributes, identifying and balancing group-wise representation in the text corpus could be a challenging line of research.

Furthermore, our experiments unveiled that fairness requirements conflict with robustness objectives. It calls for attention to introduced bias from other desired model properties such as privacy, explainability, distributed learning, etc. For example, differential privacy mechanisms often add small privatization noise to obfuscate the contribution of a specific individual to the results of a function. The level of privatization noise could be amplified by downstream processing and have a striking effect on model fairness. Moreover, added constraints such as spatially separated data or computation resources ushered in the development of distributed and federated learning algorithms. Recently, it was reported that distributed algorithms often perform unfairly compared to their centralized alternatives [122]. In federated learning, maintaining model fairness is even more challenging due to the heterogeneity of data and resources among the clients, uneven querying between the server and the clients, and a lack of global group-wise statistics to measure group fairness. Such disconnects between the individual clients can likely exacerbate the existing label biases or introduce biases in resulting models. We encourage a line of research that investigates bias amplification from federated learning methods, unveiling characteristics that lead to unfair

models in a distributed setup, and possibly designing mitigation mechanisms to improve the fairness of federated learning algorithms.

In conclusion, studies on the introduced bias are vital for understanding the limitations of machine learning models in achieving fairness under current learning practices. Since the pre-conditions of introduced bias are often systemic in nature, mitigation techniques should involve a system-wise overhaul. Rather than developing algorithmic mitigations for introduced bias, we focus on developing a thorough understanding of the root causes of unfairness in machine learning. We expect that further studies on introduced bias will enrich computer science ethics standards such as IEEE P7003 [123] and develop a pre-development checklist of sources of bias.

REFERENCES

[1] C. C. Miller, *Can an Algorithm Hire Better Than a Human?* [Online; accessed 05-12-2019], 2015. [Online]. Available: https://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html.

[2] K. Petrasic, B. Saul, J. Greig, M. Bornfreund, and K. Lamberth, *Algorithms and Bias: What Lenders Need to Know*, [Online; accessed 05-12-2019], 2017. [Online]. Available: https://www.whitecase.com/publications/insight/algorithms-and-bias-what-lenders-need-know.

[3] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, *Machine Bias*, [Online; accessed 05-12-2019], 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[4] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin, "Ensuring fairness in machine learning to advance health equity," *Annals of internal medicine*, vol. 169, no. 12, pp. 866–872, 2018.

[5] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[6] F. Kamiran and T. Calders, "Classifying without discriminating," in *Proceedings of the 2nd International Conference on Computer, Control and Communication*, IEEE, Karachi, Pakistan, 2009, pp. 1–6.

[7] D. Pedreschi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ICDM)*, Las Vegas, Nevada, USA: ACM, 2008, pp. 560–568.

[8] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Sydney, NSW, Australia: ACM, 2015, pp. 259–268.

[9] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building classifiers with independency constraints," in *ICDM Workshops 2009, IEEE International Conference on Data Mining Workshops*, Miami, Florida, USA: IEEE Computer Society, 2009, pp. 13–18.

[10] F. Kamiran and T. Calders, "Classification with no discrimination by preferential sampling," in *Proceedings of the 19th Machine Learning Conf. Belgium and The Netherlands*, Leuven, Belgium, 2010, pp. 1–6.

[11] Y. Zhou, M. Kantarcioglu, and C. Clifton, "On improving fairness of AI models with synthetic minority oversampling techniques," in *Proceedings of the 2023 SIAM International Conference on Data Mining, SDM 2023, Minneapolis-St. Paul Twin Cities, MN, USA, April 27-29, 2023*, SIAM, 2023, pp. 874–882. DOI: 10.1137/1.9781611977653.CH98. [Online]. Available: https://doi.org/10.1137/1.9781611977653.ch98.

[12] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination aware decision tree learning," in *Proceeding of the 10th IEEE International Conference on Data Mining (ICDM)*, Sydney, Australia: IEEE Computer Society, 2010, pp. 869–874.

[13] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Proceedings of the Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD*, ser. Lecture Notes in Computer Science, vol. 7524, Bristol, UK: Springer, 2012, pp. 35–50.

[14] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel, "Fairness through awareness," in *Proceedings of the Innovations in Theoretical Computer Science*, Cambridge, MA, USA: ACM, 2012, pp. 214–226.

[15] 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, WWW*, Perth, Australia: ACM, 2017, pp. 1171–1180.

[16] R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proceedings of the 30th International Conference on Machine Learning, ICML*, ser. JMLR Workshop and Conference Proceedings, vol. 28, Atlanta, GA, USA: JMLR.org, 2013, pp. 325–333.

[17] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data Mining and Knowledge Discovery*, vol. 21, no. 2, pp. 277–292, 2010.

[18] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., 2016, pp. 3315–3323. [Online]. Available: https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html.

[19] G. Pleiss, M. Raghavan, F. Wu, J. M. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon *et al.*, Eds., 2017, pp. 5680–5689. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/b8b9c74ac526fffbeb2d39ab038d1cd7-Abstract.html.

[20] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil, "Leveraging labeled and unlabeled data for consistent fair binary classification," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019, pp. 12 739–12 750. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/ba51e6158bcaf80fd0d834950251e693-Abstract.html.

[21] I. Zliobaite, F. Kamiran, and T. Calders, "Handling conditional discrimination," in *Proceedings of the 11th IEEE International Conference on Data Mining, ICDM*, Vancouver, BC, Canada: IEEE Computer Society, 2011, pp. 992–1001.

[22] R. Nabi and I. Shpitser, "Fair inference on outcomes," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI)*, New Orleans, Louisiana, USA: AAAI Press, 2018, pp. 1931–1940.

[23] S. Chiappa and W. S. Isaac, "A causal bayesian networks viewpoint on fairness," in *Proceedings of the Privacy and Identity Management. Fairness, Accountability, and Transparency in the Age of Big Data IFIP*, Vienna, Austria: Springer, 2018, pp. 3–20.

[24] M. J. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, J. G. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, 2018, pp. 2569–2577. [Online]. Available: http://proceedings.mlr.press/v80/kearns18a.html.

[25] M. J. Kearns, S. Neel, A. Roth, and Z. S. Wu, "An empirical study of rich subgroup fairness for machine learning," in *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, ACM, 2019, pp. 100–109. DOI: 10.1145/3287560.3287592. [Online]. Available: https://doi.org/10.1145/3287560.3287592.

[26] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. M. Wallach, "A reductions approach to fair classification," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, ser. Proceedings of Machine Learning Research, vol. 80, PMLR, 2018, pp. 60–69. [Online]. Available: http://proceedings.mlr.press/v80/agarwal18a.html.

[27] A. Agarwal, M. Dudik, and Z. S. Wu, "Fair regression: Quantitative definitions and reduction-based algorithms," in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, ser. Proceedings of Machine Learning Research, vol. 97, PMLR, 2019, pp. 120–129. [Online]. Available: http://proceedings.mlr.press/v97/agarwal19d.html.

[28] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, J. Furman, G. E. Marchant, H. Price, and F. Rossi, Eds., ACM, 2018, pp. 335–340. DOI: 10.1145/3278721.3278779. [Online]. Available: https://doi.org/10.1145/3278721.3278779.

[29] Y. Roh, K. Lee, S. E. Whang, and C. Suh, "Fairbatch: Batch selection for model fairness," in *9th International Conference on Learning Representations, ICLR 2021*, Virtual Event, Austria: OpenReview.net, 2021.

[30] Y. Alufaisan, L. R. Marusich, J. Z. Bakdash, Y. Zhou, and M. Kantarcioglu, "Does explainable artificial intelligence improve human decision-making?" In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, AAAI Press, 2021, pp. 6618–6626. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16819.

[31] B. Vidyanathan. "Systemic racial bias in the criminal justice system is not a myth." (Jun. 29, 2020), [Online]. Available: https://www.thepublicdiscourse.com/2020/06/65585/.

[32] G. F. Alaka Holla. "Is there systemic bias in the classroom?" (Jul. 30, 2021), [Online]. Available: https://blogs.worldbank.org/developmenttalk/there-systemic-bias-classroom.

[33] F. Martinez-Plumed, C. Ferri, D. Nieves, and J. Hernandez-Orallo, "Fairness and missing values," arXiv:1905.12728, 2019. arXiv: 1905.12728.

[34] U. Pujianto, A. P. Wibawa, M. I. Akbar, *et al.*, "K-nearest neighbor (k-nn) based missing data imputation," in *2019 5th International Conference on Science in Information Technology (ICSITech)*, IEEE, 2019, pp. 83–88.

[35] T. E. Raghunathan, J. M. Lepkowski, J. Van Hoewyk, P. Solenberger, *et al.*, "A multivariate technique for multiply imputing missing values using a sequence of regression models," *Survey methodology*, vol. 27, no. 1, pp. 85–96, 2001.

[36] T. Hastie, R. Mazumder, J. D. Lee, and R. Zadeh, "Matrix completion and low-rank SVD via fast alternating least squares," *Journal of Machine Learning Research JMLR*, vol. 16, pp. 3367–3402, 2015.

[37] V. Nanda, S. Dooley, S. Singla, S. Feizi, and J. P. Dickerson, "Fairness through robustness: Investigating robustness disparity in deep learning," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT*, Virtual Event / Toronto, Canada: ACM, 2021, pp. 466–477.

[38] 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, (FAT*)*, Atlanta, GA, USA: ACM, 2019, pp. 329–338.

[39] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Survey*, vol. 54, no. 6, 115:1–115:35, 2021.

[40] S. Mahabadi and A. Vakilian, "Individual fairness for k-clustering," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119, PMLR, 2020, pp. 6586–6596. [Online]. Available: http://proceedings.mlr.press/v119/mahabadi20a.html.

[41] J. Kang, J. He, R. Maciejewski, and H. Tong, "Inform: Individual fairness on graph mining," in *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, ACM, 2020, pp. 379–389. DOI: 10.1145/3394486.3403080. [Online]. Available: https://doi.org/10.1145/3394486.3403080.

[42] F. Petersen, D. Mukherjee, Y. Sun, and M. Yurochkin, "Post-processing for individual fairness," *CoRR*, vol. abs/2110.13796, 2021. arXiv: 2110.13796. [Online]. Available: https://arxiv.org/abs/2110.13796.

[43] T. Davidson, D. Bhattacharya, and I. Weber, *Racial bias in hate speech and abusive language detection datasets*, 2019. arXiv: 1905.12516 [cs.CL].

[44] T. A. Rahman, B. Surma, M. Backes, and Y. Zhang, "Fairwalk: Towards fair graph embedding," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI*, Macao, China: ijcai.org, 2019, pp. 3289–3295.

[45] S. Bourli and E. Pitoura, "Bias in knowledge graph embeddings," in *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, (ASONAM)*, The Hague, Netherlands: IEEE, 2020, pp. 6–10.

[46] A. J. Bose and W. L. Hamilton, "Compositional fairness constraints for graph embeddings," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, Long Beach, Calif., USA: PMLR, 2019, pp. 715–724.

[47] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proceedings of the Conference on Fairness, Accountability and Transparency (FAT)*, New York, NY, USA: PMLR, 2018, pp. 77–91.

[48] A. Das and W. S. Geisler, "A method to integrate and classify normal distributions," *Journal of Vision*, vol. 21, no. 10, pp. 1–1, 2021.

[49] L. C. Andrews, *Special Functions of Mathematics for Engineers*. Spie Press, 1998, vol. 49.

[50] S. I. Wang and C. D. Manning, "Fast dropout training," in *Proceedings of the 30th International Conference on Machine Learning, ICML*, ser. JMLR Workshop and Conference Proceedings, vol. 28, Atlanta, GA, USA: JMLR.org, 2013, pp. 118–126.

[51] D. Dua and C. Graff, *UCI machine learning repository*, http://archive.ics.uci.edu/ml, 2017.

[52] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, Jun. 2002.

[53] F. P. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA, 2017, pp. 3992–4001.

[54] *What are the warning signs of heart attack?* 2015. [Online]. Available: https://www.heart.org/en/health-topics/heart-attack/warning-signs-of-a-heart-attack.

[55] L. Eldgridge, *What is the average age for a lung cancer diagnosis?* https://www.verywellhealth.com/what-is-the-average-age-for-lung-cancer-2249260, Jul. 2020.

[56] A. J. Carter and C. N. Nguyen, "A comparison of cancer burden and research spending reveals discrepancies in the distribution of research funding," *BMC Public Health*, vol. 12, no. 526, Jul. 2012.

[57] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," *Decision Support Systems*, vol. 62, pp. 22–31, 2014.

[58] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2019, vol. 793.

[59] T. Emmanuel, T. M. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," *J. Big Data*, vol. 8, no. 1, p. 140, 2021. DOI: 10.1186/s40537-021-00516-9. [Online]. Available: https://doi.org/10.1186/s40537-021-00516-9.

[60] Y. Zhang and Q. Long, "Fairness in missing data imputation," *CoRR*, vol. abs/2110.12002, 2021. arXiv: 2110.12002. [Online]. Available: https://arxiv.org/abs/2110.12002.

[61] S. Caton, S. Malisetty, and C. Haas, "Impact of imputation strategies on fairness in machine learning," *J. Artif. Intell. Res.*, vol. 74, pp. 1011–1035, 2022.

[62] S. Guha, F. A. Khan, J. Stoyanovich, and S. Schelter, "Automated data cleaning can hurt fairness in machine learning-based decision making," in *39th IEEE International Conference on Data Engineering, ICDE*, 2023, pp. 3747–3754.

[63] M. T. Islam, A. Fariha, A. Meliou, and B. Salimi, "Through the data management lens: Experimental analysis and evaluation of fair classification," in *Proccedings of the SIGMOD '22: International Conference on Management of Data*, 2022, pp. 232–246.

[64] Y. Zhang and Q. Long, "Assessing fairness in the presence of missing data," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, K. Nguyen, P. S. Liang, J. W. Vaughan, and Y. Dauphin, Eds., vol. 34, Curran Associates, Inc., 2021, pp. 16 007–16 019. [Online]. Available: https://proceedings.neurips.cc/paper/2021/file/85dca1d270f7f9aef00c9d372f114482-Paper.pdf.

[65] C. M. R. Haider, C. Clifton, and Y. Zhou, "Unfair ai: It isnt just biased data," in *Proceedings of the IEEE International Conference on Data Mining*, 2022, pp. 957–962.

[66] C. Ashurst, R. Carey, S. Chiappa, and T. Everitt, "Why fair labels can yield unfair predictions: Graphical conditions for introduced unfairness," in *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI*, 2022, pp. 9494–9503.

[67] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2019, vol. 793.

[68] R. R. Andridge and R. J. Little, "A review of hot deck imputation for survey non-response," *International statistical review*, vol. 78, no. 1, pp. 40–64, 2010.

[69] S. Van Buuren, "Multiple imputation of discrete and continuous data by fully conditional specification," *Statistical methods in medical research*, vol. 16, no. 3, pp. 219–242, 2007.

[70] D. J. Stekhoven and P. Buhlmann, "Missforestnon-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012.

[71] H. Feng, C. Guoshun, Y. Cheng, B. Yang, and Y. Chen, "A SVM regression based approach to filling in missing values," in *Knowledge-Based Intelligent Information and Engineering Systems, 9th International Conference, KES 2005, Melbourne, Australia, September 14-16, 2005, Proceedings, Part III*, R. Khosla, R. J. Howlett, and L. C. Jain, Eds., ser. Lecture Notes in Computer Science, vol. 3683, Springer, 2005, pp. 581–587. DOI: 10.1007/11553939_83. [Online]. Available: https://doi.org/10.1007/11553939%5C_83.

[72] O. Delalleau, A. C. Courville, and Y. Bengio, "Efficient EM training of gaussian mixtures with missing data," *CoRR*, vol. abs/1209.0521, 2012. arXiv: 1209.0521. [Online]. Available: http://arxiv.org/abs/1209.0521.

[73] Y. Wang and L. Singh, "Analyzing the impact of missing values and selection bias on fairness," *Int. J. Data Sci. Anal.*, vol. 12, no. 2, pp. 101–119, 2021. DOI: 10.1007/s41060-021-00259-z. [Online]. Available: https://doi.org/10.1007/s41060-021-00259-z.

[74] K. Yang and J. Stoyanovich, "Measuring fairness in ranked outputs," in *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, 2017, 22:1–22:6. DOI: 10.1145/3085504.3085526.

[75] F. P. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Proccedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, 2017, pp. 3992–4001.

[76] F. Ding, M. Hardt, J. Miller, and L. Schmidt, "Retiring adult: New datasets for fair machine learning," in *Proccedings of the Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems*, 2021, pp. 6478–6490.

[77] B. Becker and R. Kohavi, *Adult*, UCI Machine Learning Repository, 1996.

[78] J. W. Smith, J. E. Everhart, W. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in *Proceedings of the annual symposium on computer application in medical care*, 1988, p. 261.

[79] S. Ulianova, *Cardiovascular disease dataset*, https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset/, 2018.

[80] H. A. Haenssle *et al.*, "Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Annals of oncology*, vol. 29, no. 8, pp. 1836–1842, 2018.

[81] G. F. Cooper *et al.*, "An evaluation of machine-learning methods for predicting pneumonia mortality," *Artificial intelligence in medicine*, vol. 9, no. 2, pp. 107–138, 1997.

[82] M. Luca, "Reviews, reputation, and revenue: The case of yelp. com," *Com (March 15, 2016). Harvard Business School NOM Unit Working Paper*, no. 12-016, 2016.

[83] C. Szegedy *et al.*, "Intriguing properties of neural networks," in *Proceedings of the 2nd International Conference on Learning Representations, ICLR*, Banff, AB, Canada: Conference Track Proceedings, 2014.

[84] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA: Conference Track Proceedings, 2015.

[85] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proceedings of the 6th International Conference on Learning Representations, ICLR*, Vancouver, BC, Canada: OpenReview.net, 2018.

[86] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," in *Proceedings of the 2017 IEEE Symposium on Security and Privacy, S&P*, San Jose, CA, USA: IEEE Computer Society, 2017, pp. 39–57.

[87] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *Proceedings of the 37th International Conference on Machine Learning, ICML*, Virtual Event: PMLR, 2020, pp. 2206–2216.

[88] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, Las Vegas, NV, USA: IEEE Computer Society, 2016, pp. 2574–2582.

[89] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *Proceedings of the 5th International Conference on Learning Representations, ICLR*, Toulon, France: OpenReview.net, 2017.

[90] E. Wong, L. Rice, and J. Z. Kolter, "Fast is better than free: Revisiting adversarial training," in *Proceedings of the 8th International Conference on Learning Representations, ICLR*, Addis Ababa, Ethiopia: OpenReview.net, 2020.

[91] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proceedings of the 36th International Conference on Machine Learning, ICML*, Long Beach, California, USA: PMLR, 2019, pp. 7472–7482.

[92] M. Cissé, P. Bojanowski, E. Grave, Y. N. Dauphin, and N. Usunier, "Parseval networks: Improving robustness to adversarial examples," in *Proceedings of the 34th International Conference on Machine Learning, ICML*, Sydney, NSW, Australia: PMLR, 2017, pp. 854–863.

[93] Y. Carmon, A. Raghunathan, L. Schmidt, J. C. Duchi, and P. Liang, "Unlabeled data improves adversarial robustness," in *Proceedings of the Annual Conference on Neural Information Processing Systems, NeurIPS*, Vancouver, BC, Canada, 2019, pp. 11 190–11 201.

[94] A. Najafi, S. Maeda, M. Koyama, and T. Miyato, "Robustness to adversarial perturbations in learning from incomplete data," in *Proceedings of the Annual Conference on Neural Information Processing Systems 2019, NeurIPS*, Vancouver, BC, Canada, 2019, pp. 5542–5552.

[95] J. Alayrac, J. Uesato, P. Huang, A. Fawzi, R. Stanforth, and P. Kohli, "Are labels required for improving adversarial robustness?" In *Proceedings of the Annual Conference on Neural Information Processing Systems, NeurIPS*, Vancouver, BC, Canada, 2019, pp. 12 192–12 202.

[96] S. Gowal, S. Rebuffi, O. Wiles, F. Stimberg, D. A. Calian, and T. A. Mann, "Improving robustness using generated data," in *Proceedings of the Annual Conference on Neural Information Processing Systems, NeurIPS*, Virtual, 2021, pp. 4218–4233.

[97] Z. Wang, T. Pang, C. Du, M. Lin, W. Liu, and S. Yan, "Better diffusion models further improve adversarial training," in *Proceedings of the International Conference on Machine Learning, ICML*, Honolulu, Hawaii, USA: PMLR, 2023, pp. 36 246–36 263.

[98] S. Peng *et al.*, "Robust principles: Architectural design principles for adversarially robust cnns," in *Proceedings of the 34th British Machine Vision Conference, BMVC*, Aberdeen, UK: BMVA Press, 2023, pp. 739–740.

[99] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, "Deep double descent: Where bigger models and more data hurt," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=B1g5sA4twr.

[100] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proceedings of the 5th International Conference on Learning Representations, ICLR*, Toulon, France: OpenReview.net, 2017.

[101] B. Biggio *et al.*, "Evasion attacks against machine learning at test time," in *Proceedings of the Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD*, Prague, Czech Republic: Springer, 2013, pp. 387–402.

[102] C. Xie *et al.*, "Improving transferability of adversarial examples with input diversity," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, Long Beach, CA, USA: Computer Vision Foundation / IEEE, 2019, pp. 2730–2739.

[103] Y. Dong *et al.*, "Boosting adversarial attacks with momentum," in *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, Salt Lake City, UT, USA: Computer Vision Foundation / IEEE Computer Society, 2018, pp. 9185–9193.

[104] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," in *Proceedings of the 8th International Conference on Learning Representations, ICLR*, Addis Ababa, Ethiopia: OpenReview.net, 2020.

[105] F. Tramèr, A. Kurakin, N. Papernot, I. J. Goodfellow, D. Boneh, and P. D. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *Proceedings of the 6th International Conference on Learning Representations, ICLR*, Vancouver, BC, Canada: OpenReview.net, 2018.

[106] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, Honolulu, HI, USA: IEEE Computer Society, 2017, pp. 86–94.

[107] S. H. Silva and P. Najafirad, "Opportunities and challenges in deep learning adversarial robustness: A survey," 2020. arXiv: 2007.00753.

[108] R. Zhai *et al.*, "MACER: attack-free and scalable robust training via maximizing certified radius," in *Proceedings of the 8th International Conference on Learning Representations, ICLR*, Addis Ababa, Ethiopia: OpenReview.net, 2020.

[109] F. Croce and M. Hein, "Provable robustness against all adversarial l_p- perturbations for p >= 1," in *Proceedings of the 8th International Conference on Learning Representations, ICLR*, Addis Ababa, Ethiopia: OpenReview.net, 2020.

[110] Y. Roh, K. Lee, S. Whang, and C. Suh, "Sample selection for fair and robust training," in *Proceedings of the Annual Conference on Neural Information Processing Systems, NeurIPS*, Virtual, 2021, pp. 815–827.

[111] Y. Roh, K. Lee, S. Whang, and C. Suh, "Fr-train: A mutual information-based approach to fair and robust training," in *Proceedings of the 37th International Conference on Machine Learning, ICML*, Virtual Event: PMLR, 2020, pp. 8147–8157.

[112] X. Ma, Z. Wang, and W. Liu, "On the tradeoff between robustness and fairness," in *Proceedings of the Annual Conference on Neural Information Processing Systems, NeurIPS*, New Orleans, LA, USA, 2022.

[113] H. Xu, X. Liu, Y. Li, A. K. Jain, and J. Tang, "To be robust or to be fair: Towards fairness in adversarial training," in *Proceedings of the 38th International Conference on Machine Learning, ICML*, Virtual Event: PMLR, 2021, pp. 11 492–11 501.

[114] H. Zeng, Z. Yue, L. Shang, Y. Zhang, and D. Wang, "On adversarial robustness of demographic fairness in face attribute recognition," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI*, Macao, SAR, China: ijcai.org, 2023, pp. 527–535.

[115] F. Croce *et al.*, "Robustbench: A standardized adversarial robustness benchmark," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021*, Virtual, 2021.

[116] N. Mehrabi, M. Naveed, F. Morstatter, and A. Galstyan, "Exacerbating algorithmic bias through fairness attacks," in *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*, Virtual Event: AAAI Press, 2021, pp. 8930–8938.

[117] P. Benz, C. Zhang, A. Karjauv, and I. S. Kweon, "Robustness may be at odds with fairness: An empirical study on class-wise accuracy," in *Proceedings of the NeurIPS 2020 Workshop on Pre-registration in Machine Learning*, Virtual Event: PMLR, 2020, pp. 325–342.

[118] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, ser. Lecture Notes in Computer Science, vol. 9908, Springer, 2016, pp. 630–645. DOI: 10.1007/978-3-319-46493-0_38. [Online]. Available: https://doi.org/10.1007/978-3-319-46493-0%5C_38.

[119] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA: Conference Track Proceedings, 2015.

[120] K. Mancuhan and C. Clifton, "Combating discrimination using bayesian networks," *Artificial Intelligence and Law*, vol. 22, no. 2, pp. 211–238, Jun. 2014. DOI: 10.1007/s10506-014-9156-4. [Online]. Available: http://dx.doi.org/10.1007/s10506-014-9156-4.

[121] N. Grgic-Hlaca, E. M. Redmiles, K. P. Gummadi, and A. Weller, "Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018*, Lyon, France: ACM, 2018, pp. 903–912. DOI: 10.1145/3178876.3186138. [Online]. Available: https://doi.org/10.1145/3178876.3186138.

[122] H. Zhang, M. Jayaweera, B. Ren, Y. Wang, and S. Soundarajan, "Unfairness in distributed graph frameworks," in *IEEE International Conference on Data Mining, ICDM 2023, Shanghai, China, December 1-4, 2023*, G. Chen, L. Khan, X. Gao, M. Qiu, W. Pedrycz, and X. Wu, Eds., IEEE, 2023, pp. 1529–1534. DOI: 10.1109/ICDM58522.2023.00203. [Online]. Available: https://doi.org/10.1109/ICDM58522.2023.00203.

[123] A. R. Koene, L. Dowthwaite, and S. Seth, "IEEE p7003™ standard for algorithmic bias considerations: Work in progress paper," in *Proceedings of the International Workshop on Software Fairness, FairWare@ICSE 2018, Gothenburg, Sweden, May 29, 2018*, Y. Brun, B. Johnson, and A. Meliou, Eds., ACM, 2018, pp. 38–41. DOI: 10.1145/3194770.3194773. [Online]. Available: https://doi.org/10.1145/3194770.3194773.