Understanding Disparities in Post Hoc Machine Learning Explanation

Vishwali Mhasawade
vishwalim@nyu.edu
New York University
New York, USA

Salman Rahman
salman@nyu.edu
New York University
New York, USA

Zoe Haskell-Craig
zjh235@nyu.edu
New York University
New York, USA

Rumi Chunara
rumi.chunara@nyu.edu
New York University
New York, USA

ABSTRACT

Previous work has highlighted that existing post-hoc explanation methods exhibit disparities in explanation fidelity (across "race" and "gender" as sensitive attributes), and while a large body of work focuses on mitigating these issues at the explanation metric level, the role of the data generating process and black box model in relation to explanation disparities remains largely unexplored. Accordingly, through both simulations as well as experiments on a real-world dataset, we specifically assess challenges to explanation disparities that originate from properties of the data: limited sample size, covariate shift, concept shift, omitted variable bias, and challenges based on model properties: inclusion of the sensitive attribute and appropriate functional form. Through controlled simulation analyses, our study demonstrates that increased covariate shift, concept shift, and omission of covariates increase explanation disparities, with the effect pronounced higher for neural network models that are better able to capture the underlying functional form in comparison to linear models. We also observe consistent findings regarding the effect of concept shift and omitted variable bias on explanation disparities in the Adult income dataset. Overall, results indicate that disparities in model explanations can also depend on data and model properties. Based on this systematic investigation, we provide recommendations for the design of explanation methods that mitigate undesirable disparities.

CCS CONCEPTS

• **Computing methodologies → Machine learning**; **Causal reasoning and diagnostics**; • **General and reference** → *Evaluation.*

KEYWORDS

fairness, explainability, post hoc explanation methods

ACM Reference Format:
Vishwali Mhasawade, Salman Rahman, Zoe Haskell-Craig, and Rumi Chunara. 2024. Understanding Disparities in Post Hoc Machine Learning Explanation. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3630106.3659043

1 INTRODUCTION

Machine learning models are increasingly being proposed for and utilized in many societal areas such as healthcare, law, education, and policy-making [3, 21, 56, 58, 64, 68]. Particularly in their role as predictive tools, these models often are utilized with a 'black box' nature. This characteristic can obscure the understanding of the underlying mechanisms driving their predictions [19, 49]. The lack of transparency raises concerns about the reliability of these models in situations where safety is a critical factor [4, 11, 25]. For example, in the context of healthcare, a proposed application of machine learning models may be to determine patient treatment plans. However, as these algorithms may sometimes lead to biased predictions for disadvantaged groups, clear insights into the factors influencing the machine learning model decisions are needed [13].

To address the lack of transparency in machine learning models, the field has seen a development towards Explainable AI (XAI), which focuses on creating methods that can explain the workings of these black box models [10, 20, 35]. Among the approaches in Explainable AI, the development of simpler models that emulate the black box models' behaviors has widespread adoption in the field [10]. This approach, known as post hoc explanation, involves developing a local model that provides explanations for individual predictions. Such explanation models are proposed for use as standalone tools, providing global explanations that shed light on the overall behavior and patterns within the black box model [57], or to explain individual predictions, offering insights into the decision-making process of individual instance [37, 46].

Post hoc explanation methods are broadly classified into four categories: counterfactual [61], rule-based [47], perturbation-based [37, 45, 46, 53], and gradient-based [51, 54]. Counterfactual explanations are computationally expensive [29] due to the demanding nature of searching for counterfactual instances in high-dimensional feature spaces. Additionally, some counterfactual suggestions may not be feasible in real-world contexts, as the changes they propose might be impractical to achieve [32]. Rule-based methods,

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3659043

on the other hand, can sometimes generate complex and hard-to-understand rules, particularly with high-dimensional data [47]. Further, finding the most effective rule can also be computationally intensive, especially for complex models. Gradient-based methods have their limitations too; they are sensitive to noise in the input space [65], ineffective in detecting spurious correlations [2], are commonly applied to unstructured data like images, and sometimes produce visually similar explanations for different classes [1]. Though each type of method has limitations, our focus is on perturbation-based post hoc explanation methods, especially LIME (Local Interpretable Model-agnostic Explanations), due to its widespread adoption for tabular data [5] and previous use to highlight disparities in post hoc explanation methods [6, 17].

Specifically, recent studies have revealed disparities in the fidelity of post hoc explanation methods, i.e., how accurately the post hoc explanation methods replicate the nature of the black box model, particularly when analyzed across data from different 'gender' and 'race' groups [6, 17]. To address this disparity in fidelity, Dai et al. [18] proposed a fairness-preserving approach for LIME, which includes a fairness constraint in the LIME objective function to ensure that the fairness properties of the black box model are also reflected in LIME. This approach builds upon previous studies that enhanced fairness in machine learning methods through similar constraints [28, 66] but did not extensively discuss how the fairness constraint can improve the fidelity of LIME explanations. Concurrently, Balagopalan et al. [6] developed a robust LIME explanation model using the 'Just train twice' methodology [36]. However, the fidelity improvement with this enhancement was demonstrated only in certain cases; Adult, LSAC, and MIMIC datasets and only for neural network methods [6].

A key point, however, is that fairness issues in machine learning models (i.e., prediction models) are a multifaceted problem that can manifest at the level of the data, the black box models, or their interpretation (e.g., via explanation methods) [7, 24]. While efforts to-date in explanation methods have predominantly focused on improving the post hoc methods [6, 18], this overlooks the role of data and black box models in generating unfair explanations. Given the evidence that sample size, covariate shift, concept shift, and omitted variables can affect model prediction accuracy and lead to disparities in black box model performance, we investigate how these characteristics of the data and the model development process affect explanation disparities. Indeed, sample size imbalance has been linked to bias in prediction [30] and calibration models [48]. Moreover, limited samples of a certain subgroup are known to affect model performance and the ability to generalize for the specific subgroup [14, 42]. A related source of algorithmic unfairness is disparately missing data across subgroups [38, 63], which can result in both an imbalance and a sample that is not representative of the true distribution of the target population, resulting in a distribution shift for some subgroups [43]. Specifically, covariate shift is known to affect black box model performance across subgroups disparately [52]. Concept shift in the outcome, that is, where the conditional distribution of the outcome given the covariates varies across subgroups, also may have an effect on the quality of black box prediction and can lead to disparities [42]. Lastly, omitting variables that have a direct effect on the outcome that is not completely mediated by other covariates may also lead to disparities in black box predictions across subgroups [39]. However, there is little evidence of how characteristics of limited sample size, covariate shift, concept shift, and omitted variable bias will influence the quality of the explanation methods with respect to the test distribution. Considering that the four above-mentioned characteristics have the potential to introduce disparities in black box model predictions, it is pertinent to investigate if and how these disparities in black box model predictions can lead to explanation disparities. This is motivated by the inherent nature of the explanation methods by which they are expected to mimic the nature of the black box model. Consider the LIME explanation method where the explanation produced by LIME at a local point x is obtained by the following generic formula: $\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$, where f is our real function (the black box model), g is a linear surrogate function we use to approximate f in the proximity of x, π_x defines the locality, and Ω represents the model complexity. Data and modeling characteristics that affect black box model performance may also have similar effects on the explanation quality as well. In sum, we explore the following data issues known to affect black box model performance but relatively unexplored in the case of model explanation disparities, 1) limited samples, 2) covariate shift, 3) concept shift, and 4) omitted variable bias.

The remainder of this paper is structured as follows: in Section 2, we discuss related work on post hoc explanation methods and the fidelity of these approaches. In Section 3, we outline the objectives and motivate the data-generating process for each objective of our research. In Section 4, we describe the explanation quality metrics and experimental setup for the synthetic and real-world data experiments. In Section 5, we present the results, and finally we discuss implications of these findings as well as synthesized recommendations based on them in Section 6.

2 RELATED WORK

Challenges in the Use of the Popular Post Hoc Explanation Method LIME. In the realm of Explainable AI (XAI), a distinction exists between models designed for inherent interpretability [9, 12, 31], such as decision trees [33] and rule lists [62, 67], and those employing post hoc explanation methods [46]. Given their higher accuracy, complex models like deep neural networks are frequently preferred in real-world settings, necessitating the use of post hoc methods to elucidate their prediction. Among post hoc explanation techniques, the Local Interpretable Model-agnostic Explanations (LIME) method stands out as a widely used method, particularly for explaining black box models applied to tabular data [5]. LIME is also valued for its model-agnostic nature, its capacity to provide local explanations, and its relative simplicity [46].

However, several challenges with the use of LIME have been noted. This method relies on perturbations, which introduce computational demands, particularly in models with numerous features [46]. Additionally, the fidelity of LIME's explanations is sensitive to adjustments in hyperparameters, including the number of perturbed samples, kernel width, and regularization parameters. Recent research efforts have acknowledged LIME's limitations and proposed improvements. For example, S-LIME introduces frameworks for generating more stable and consistent explanations across various perturbations [70]. Despite several issues associated with LIME, systematic studies exploring performance degradation, especially concerning disadvantaged groups, are lacking. Our research aims to be the pioneering effort in investigating why post hoc explanation methods like LIME exhibit disparate explanations across different subgroups.

Disparities In Post Hoc Explanation Methods and Efforts to Mitigate. Recent research has delved into the exploration of race and gender-based disparities in a range of post hoc explanation methods, including LIME, SHAP, SmoothGrad, IntGrad, VanillaGrad, and Maple [17, 18]. These studies have utilized several datasets, such as German Credit, Student Performance, Adult, and COMPAS. Further, Balagopalan et al. [6] conducted a study revealing explanation disparities among race and gender in both local explanation methods (LIME and SHAP) and global methods (Generalized Additive Model (GAM) and sparse decision tree (Tree)). Their research used Adult, LSAC, MIMIC, and Recidivism datasets covering four critical domains: finance, college admissions, healthcare, and the justice system. In terms of efforts to improve the fairness of explanation methods, Balagopalan et al. [6] demonstrated that balanced samples between the subgroups and robust training for local and global explanation methods can improve the fidelity gap, which refers to how well an explanation model approximates the behavior of a black box model [6]. For local explanation methods, the authors trained a fairer explanation model using the Just Train Twice (JTT) methodology [36]. Although improvements were noted for neural network models applied to Adult, LSAC, and MIMIC datasets, the authors did not see improvements in explanation fairness for the Recidivism dataset. As the properties of these datasets were not investigated, it is not yet clear why improvements were seen in some datasets but not others or under what conditions these methods may improve explainability. Balagopalan et al. [6] also observed that fidelity gaps depend on the representation of the data; they train black box models with features that have no mutual information with respect to the sensitive attribute and observe that fidelity gaps decrease. Although this provides insight into one specific property of data, how much information about the sensitive attribute is available in the data representation, the authors suggest further investigation about other data properties, which forms the focus of this work. In parallel, Dai et al. [17] proposed a method to generate fairness-preserving explanations by adding a penalty term to the LIME objective function [18], an approach similar to creating fair machine learning algorithms [66].

While existing studies effectively highlight disparities and propose fair post hoc explanation methods, they predominantly concentrate on the outputs of the explanation methods. The role of data and black-box models in these disparities are not carefully examined though both the data used and the nature of the black-box models can be significant sources of disparity [7].

3 DATA GENERATING PROCESS AND OBJECTIVES

3.1 Data Generating Process

Here, we describe the data-generating process (DGP) for assessing the reasons for disparities in model explanations in line with previous work using simple causal graphs for systematic fairness assessments [39, 52, 55]. We refer to the outcome as Y, a binary variable that takes a value of 0 or 1. We consider a sensitive attribute A, such as race or gender, for which we represent the disadvantaged group as $A = 0$ and the advantaged group as $A = 1$. A is associated with the independent covariate L. Two attributes, C, and L have a direct effect on the outcome Y, where L mediates a part of the effect of C on Y. C has a direct effect on Y ($C \rightarrow Y$) and an indirect effect through L ($C \rightarrow L \rightarrow Y$). The relationship between these variables is represented by the causal directed acyclic graph (DAG) in Figure 1(a). In our causal graph, the covariates and the sensitive attribute affect the outcome either through other covariates (i.e., ($A \rightarrow L \rightarrow Y$)) or directly (i.e., $C \rightarrow Y$).

We consider two setups; in the first, presented in Figure 1(a), A has an effect on Y only through L, and in the second, in Figure 1(b) A affects the relationship between L and Y. The second setting allows us to assess non-linear complex functional forms between A and Y. We use these DGPs to represent the underlying relationship between the variables in the general population, from which we will draw samples to form training and testing datasets.

3.2 Objective 1: effect of sample size of disadvantaged group data used for training

In order to investigate the effect of sample size imbalance on disparities, we consider a scenario where we vary the proportion of the sample size of the disadvantaged group ($A = 0$) from 5% to 50% of the total training sample and accordingly vary the proportion of the advantaged group ($A = 1$) from 95% to 50% of the total training sample. To isolate this effect from non-random sampling of data [27], we assume that there is no distribution shift in the predictors L and C between the training and test distributions. It is important to note that the probability of the outcome Y, given the covariates L and C, $P(Y = y \mid L, C)$, remains independent of the sample size of $A = 0$ since $Y \perp\!\!\!\perp A \mid L, C$ and we vary the proportion of both advantaged and disadvantaged group to ensure that the training sample is a perfectly random sample of the population distribution.

The DGP for this objective is represented in Figure 1(a). We assume that, in the general population, attribute A is generated by a binomial probability with $A \in \{0, 1\} \sim \text{Binomial}(1, 0.5)$, C follows a normal distribution $C \sim \mathcal{N}(0, 1)$, and $P(L)$ is dependent on both A and C such that $L \sim \mathcal{N}(0, 0.5) + 0.7 \times A + 0.3 \times C$. These parameters were chosen to allow for differences in the distribution of L across A such that $P(L \mid C) \neq P(L \mid C, A)$. We assume that covariates L and C have a direct effect of a similar magnitude on the outcome, Y, which follows a binomial distribution $Y \sim \text{Binomial}(1, Y_p)$ with probability of as $Y_p : P(Y = 1 \mid L, C) = \begin{cases} 0.1 & \text{if } i < 0 \\ 0.9 & \text{if } i \geq 1 \end{cases}, i = 0.5 \times C - 1.5 \times L + 0.5$.

We also consider both cases, when the black-box machine learning model includes information on the sensitive attribute A during training and when it does not [40, 60], to assess the fairness properties of explanation metrics when the population Bayes-optimal model is not subgroup Bayes-optimal [44].

3.3 Objective 2: effect of covariate shift in disadvantaged group data between training and test distributions

Here, we explore the effect of a covariate shift, where the training distribution of covariate L for $A = 0$ is not representative of the test distribution (or the population distribution). That is, $P_{\text{train}}(L \mid A = 0) \neq P_{\text{test}}(L \mid A = 0)$, however the conditional probabilities are consistent; $P_{\text{train}}(Y \mid L, A = 0) = P_{\text{test}}(Y \mid L, A = 0)$. We generate a covariate shift in L for $A = 0$ by sampling for the training distribution depending on both the sensitive attribute A and the covariate L, such that there is missing data for group $A = 0$ for all observations with L below a threshold t. In this way, we vary the overlap in the range of L between the test set and the training set from 100% to 20%. It should be noted that the probability distribution of Y given $A = 0$ is not the same in the training and test sets; $P_{\text{train}}(Y \mid A = 0) \neq P_{\text{test}}(Y \mid A = 0)$. As the overlap between the training and testing sets is reduced, the model has less information in the training set to learn about the association between the variables in the general population for $A = 0$. As such, we hypothesize that less overlap in the training and test distributions for the disadvantaged group will lead to greater disparities.

The DGP for this objective is also represented in Figure 1(a). Similar to objective 1, the data-generating process for each variable is as follows: $A \in \{0, 1\} \sim \text{Binomial}(1, 0.5)$, $C \sim \mathcal{N}(0, 1)$, $L = \mathcal{N}(0, 0.5) + 0.7 \times M + 0.3 \times C$, $Y_p = P(Y = 1) = \begin{cases} 0.1 & \text{if } i < 0 \\ 0.9 & \text{if } i \geq 1 \end{cases}$, $i = 0.5 \times C - 1.5 \times L + 0.5$, $Y \sim \text{Binomial}(1, Y_p)$. Note that for both objectives 1 and 2, we evaluate model performance on a test set representative of the population. That is, the proportion of $A = 0$ in the test set is 50%, and the full distribution of the values for L is represented. As such, $P_{\text{train}}(A) \neq P_{\text{test}}(A)$. Again, we consider both cases when A is and is not included in the covariates to which the black-box model has access.

3.4 Objective 3: effect of concept shift

Here we examine concept shift, where $P(Y \mid L, A = 0) \neq P(Y \mid L, A = 1)$ [41]. That is the relationship between Y and L changes depending on A. Here, we vary the magnitude of concept shift by increasing the effect of the sensitive attribute on the distribution of the outcome, that is, by varying the degree of difference between the distribution of the outcome across groups that results in $P(Y) \neq P(Y|A)$.

The DGP for this objective is represented in figure 1(b). We generate the distribution of variables A and C in the general population following the same procedure outlined in objectives 1 and 2. In order to capture the impact of A on $P(Y|L)$, we augment the direct effect of A and L such that $L \sim \mathcal{N}(0, 0.1) + 0.7 \times A + 0.3 \times C$. The concept shift is produced by specifying that the probability Y_p depends on $i = 0.5 \times C + -1 \times L + 1.5 \times A \times L + \beta \times (1 - A) \times L - 0.2$, through the step function $Y_p = \begin{cases} 0.1 & \text{if } i < 0 \\ 0.9 & \text{if } i \geq 0 \end{cases}$. Note that the relationship between L and Y_p is determined by A and β, where β is the strength of the concept shift. We consider $\beta = 1.5$ as 'low' concept shift, $\beta = 0.5$ a 'moderate' concept shift, and $\beta = -0.5$ a 'high' concept shift. The coefficients on C, L, and the intercept term were chosen in order to ensure an equal distribution of $Y = 1$ and $Y = 0$ in the training sample to ensure class balance. As before, we consider the impact of including or not including information on A while training the black box model.

3.5 Objective 4: effect of the magnitude of direct effect of the omitted covariate

Finally, we test the impact of omitting a covariate, C, that has a direct effect on the outcome in the black box model on explanation disparities. Following the data-generating process represented in Figure 1(a), the distribution of A and C are generated as before, $L \sim \mathcal{N}(0, 0.5) + 0.3 \times A + 0.3 \times C$, and as before, $Y \sim \text{Binomial}(1, Y_p)$. We vary the direct effect of the attribute, C, on the outcome that is not mediated by other covariates, such that $Y_p \sim \alpha \times C + L - 0.2$ where $\alpha \in \{0, 0.5, 1, 1.5\}$. We assess the disparities resulting from not including the variables C in model training (or test).

4 METHODS

4.1 Notation

From the underlying population we draw a dataset $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$ comprising n observations. Each observation consists of a d-dimensional feature vector $x_i \in \mathbb{R}^d$ for the i^{th} data point in $\mathcal{D}$, along with a corresponding binary class label $y_i \in \{0, 1\}$. We consider a machine learning model $f : \mathbb{R} \to \{0, 1\}$, such as logistic regression (LR) or a neural network (NN), trained on samples from $\mathcal{D}$. For a given instance x_i and machine learning model f, a local explanation method is denoted as $E : (x_i, f) \to \psi \in \mathbb{R}^d$, where ψ represents the output vector of feature importance. The local model, E, is designed to mimic the behavior of f in the vicinity of x_i [17, 69].

4.2 Explanation Quality Metrics

We utilize two metrics to measure the fidelity gap across groups, as introduced by Balagopalan et al. [6]. Fidelity is the degree to which an explanation model precisely reflects the predictions of a black box model. For a black box model f and explanation model E, the fidelity quantifies how closely E approximates the behavior of f. Mathematically, the explanation fidelity for data points $(x_i, y_i)_{i=1}^N$ is calculated as: $\frac{1}{N} \sum_{i=1}^N Q(f(x_i), E(x_i))$, where Q is a performance metric such as accuracy. This measure allows for the evaluation of fairness by illustrating the alignment between the machine-learning model and the explanation model.

Maximum Fidelity Gap from Average. The Maximum Fidelity Gap from the Average measures the largest deviation in fidelity for any group from the average fidelity across all groups. This metric assesses the extent to which the fidelity of an explanation model for disadvantaged groups deviates from the overall average fidelity [6, 16, 36]. The maximum fidelity gap from average, Δ_Q is represented as follows:

$$\Delta_Q = \max_j \left[\frac{1}{N} \sum_{i=1}^{N} Q(f(x_i), E(x_i)) - \frac{1}{N_j} \sum_{i:\delta_j^i = 1} Q(f(x_i), E(x_i)) \right]$$

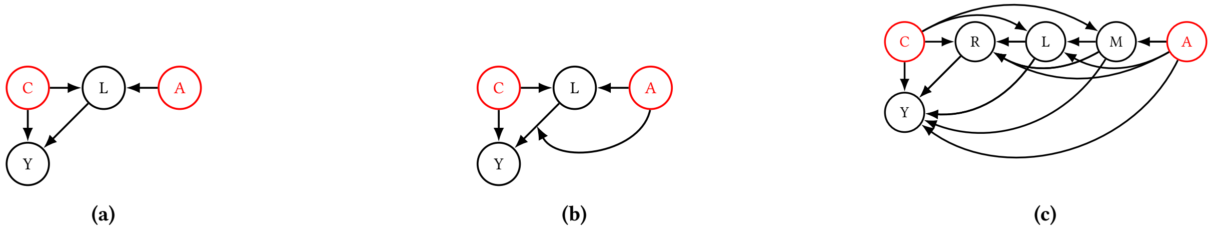


Figure 1: Causal DAG for the synthetic datasets (a, b) and the Adult dataset (c). In (a) is the causal graph describing the data-generating process (DGP) for objectives 1, 2, and 4, (b) is the causal graph for the DGP for objective 3 (concept shift). The concept shift is represented as the arrow showing the effect of A on the relationship between L and Y, such that $P(Y \mid L, A = 0) \neq P(Y \mid L, A = 1)$. In (c) we consider gender as the sensitive attribute of interest. A and M represent gender and marital status, respectively. C is age and nationality, L is the level of education, R corresponds to the working class, occupation, and hours per week, and Y is the income class.

where Q represents performance metric such as Accuracy, N is the total number of data points, $\delta^i_j = 1$ indicates that point x_i belongs to the j-th group, and N_j is the number of data points where $\delta_j = 1$. We specifically focus on the maximum fidelity gap from the average for the 'Accuracy' performance metric following the performance metric used by Balagopalan et al. [6], where we assess the accuracy between the predictions of the black box model $f(:)$ and the explanation method $E(:)$ and denote it as Δ_{Acc}.

Mean Fidelity Gap Amongst Subgroups. This metric illustrates the average difference in fidelity between groups. Within the Mean Fidelity Gap, performance metrics such as AUROC, Residual Error, and Accuracy may be used to quantify disparities between black box model predictions and their explanations. This metric is computed as follows [6]:

$$\Delta_Q^{group} = \frac{2}{G(G-1)} \sum_{p=1}^{G} \sum_{j=p+1}^{G} |Q_p - Q_j|$$

with

$$Q_p = \frac{1}{N_p} \sum_{i:\delta^i_p=1} Q(f(x_i), E(x_i))$$

where G is the total number of groups, Q_p and Q_j are the performance metrics for the p^{th} and j^{th} groups respectively, δ_p denotes the p-th group, and N_j is the number of data-points in δ_p. As described above, the Q metric includes Accuracy, denoted by Δ^{group}_{Acc}. We use Accuracy as the performance metric for both explanation quality metrics to compare how accurate the explanations for the disadvantaged group are in comparison to overall and how accurate the explanations for the disadvantaged group are in comparison to the advantaged group.

4.3 Experimental Setup

Simulation and Real-world Data. The simulated data, as outlined in Section 3, is constructed to reflect a specific causal structure. This dataset comprises 20,000 data points featuring a sensitive attribute A (such as gender), covariates L and C, and a binary outcome Y. Specifically, the covariate L mediates a part of the effect of C on Y. The process for generating this data is detailed in Section 3. We assess the quality of the explanations generated with respect to the sensitive attribute A.

In addition to the simulated data, we employ the widely used Adult dataset for real-world analysis [22, 34]. We chose this dataset due to the importance of the dataset in developing and evaluating post-hoc explainability methods [26, 50, 59], the general popularity of the dataset for fairness analysis [6, 8, 17, 23] and the availability of a well studied causal DAG [15, 39]. This previously used causal graph is shown in Figure 1(c). For the task using the Adult dataset, the goal is to predict whether an individual's income is above or below \$50,000. The data consist of age, working class, education level, marital status, occupation, relationship, race, gender, capital gain and loss, working hours, and nationality variables for 48842 individuals. In Adult, disparities in explanation quality (maximum fidelity gap and mean fidelity gap between groups using accuracy and error) have been found for both logistic regression and neural network black box models with respect to gender [6], but the source of these disparities has not been investigated. Following [39], we consider the variable 'hours-per-week' as one of the predictors with a direct effect on the outcome of interest, income Y represented by L in the synthetic experiment. All other covariates that may be associated with both 'hours-per-week' and income are represented as C in the synthetic experiment, in particular age and nationality. We consider gender as the sensitive attribute of interest ('A' in the synthetic experiment) as has been done in previous studies on algorithmic fairness[6, 39], and considering 'males' as the disadvantaged group since we observe that it is easier to predict the outcome for the advantaged group 'females' than it is to predict the outcome for the disadvantaged group 'males.'

Moreover, for this dataset and task, we detect a statistically significant concept shift ($p \leq 0.1$) of gender (A) on the relationship between 'hours-per-week' (L) and the outcome (Y). The concept shift was tested using logistic regression with an interaction term, such that logit(income) = $\beta_1\times$ sex + $\beta_2\times$'hours-per-week' + $\beta_3\times$sex$\times$ 'hours-per-week'. Furthermore, following the proposed causal graph for the Adult dataset, we use 'nationality' as the omitted variable (C in the synthetic experiment), which has a direct effect on Y along with indirect effects on Y through other covariates. We examine these effects further by performing the following experiments, which match the experiments we performed with synthetic data. We analyze the effect of the sample size of the disadvantaged group (objective 1) by further restricting the proportion of females in the training dataset from 10% to 100% (with a 10% increment). We also perform experiments examining the effect of non-overlap between training and test distribution of 'hours-per-week' (objective 2) by limiting the observations of males in the training dataset to individuals working less than 100, 80, 60, 40, and 20 hours per week and thus introducing covariate shift in 'hours-per-week' for males. We also check the impact of concept shift alone by ensuring the training set is 50% females and 50% male (objective 3). This ensured that we had an equal representation of both groups in the training set. We examine the effect of omitting a covariate that has a direct effect on the outcome (objective 4) by omitting a) gender, b) nationality, and c) both from the black box model.

Machine Learning Model. Our study examines both logistic regression (LR) and neural network (NN) models as the underlying functions in the black box model which are implemented using the PyTorch framework. This selection enables us to compare the impacts of a simpler, linear model and a more complex, flexible model on the quality of explanations, aligning with methodologies used in previous studies [6, 17]. The neural network architecture consists of four layers linked by ReLU activation functions and concludes with a final layer with a sigmoid activation function for output, mirroring the setup used for assessing disparities by Dai et al. [17]. The layer configurations are as follows: the first linear layer maps input features to 50 outputs, succeeded by ReLU activation. The subsequent layers expand the dimensionality (50 to 100, and then 100 to 200), each followed by ReLU activations. The final linear layer condenses these 200 inputs to a single output, processed through a sigmoid activation function. We utilize the Adam optimizer with a weight decay of $1e^{-4}$ and train the model using Binary Cross-Entropy Loss over 100 epochs. We represent the NN models that do not include the sensitive attribute in model training as NN$_{\not A}$ and the NN models that include the sensitive attribute as NN$_{A}$ and similar for the LR models as LR$_{\not A}$, and LR$_{A}$, respectively.

Explanation Method. As motivated above, our focus is on LIME for the explanation model, chosen for its extensive application and documented disparities in previous studies [5, 6, 17]. LIME operates by constructing a local surrogate model to interpret specific data points, thereby shedding light on the prediction of the underlying complex model. LIME generates a dataset of perturbations by altering the features of a specific instance, creating a range of variations. The original machine learning model is then used to obtain predictions for the dataset comprising of the perturbed instances. The perturbed instances are weighted according to their similarity to the original instance by comparing the distance to the original instance from which the perturbations are obtained. Subsequently, a simpler, interpretable linear model is trained on this weighted, perturbed dataset [46]. The objective of this simpler model is to approximate the complex model's predictions in the vicinity of the selected instance. The explanation is obtained from the features of this simpler model, identifying the key covariates influencing the specific prediction. The implementation of LIME utilizes LimeTabularExplainer from the 'lime' package in Python, which operates on the training data without discretizing continuous features. The result from this implementation is 1000 perturbed samples for generating the explanation of each instance, considering all dataset features for generating explanations for each test instance.

Settings and Implementation Procedures. The datasets $\mathcal{D}$ are divided into training $\mathcal{D}_{\text{train}}$ and testing $\mathcal{D}_{\text{test}}$ sets, comprising 70% and 30% of the data, respectively. The black box model f is trained using $\mathcal{D}_{\text{train}}$. The test dataset $\mathcal{D}_{\text{test}}$ is further segmented into two groups $\mathcal{D}_{\text{test}}^{A=1}$ for group 1 and $\mathcal{D}_{\text{test}}^{A=0}$ for group 0. We generate explanations using LIME for both groups. The fidelity of these explanations is assessed by comparing the predictions from the black box model f and the explanation model E, using the fidelity metrics defined in Section 4. To evaluate the consistency of predictions between the black box model f and the explanation model E, we conduct five trials, each with different random seeds[1].

5 RESULTS

5.1 Explanation Disparities in Synthetic Simulations

Overall findings on model complexity and inclusion of sensitive attributes. We observe similar characteristics across all objectives between simpler linear models (LR) and complex neural network models (NN). In general, higher disparities in explanation metrics (Maximum Fidelity Gap, Δ_{Acc} and Mean Fidelity Gap, Δ_{Acc}^{group}) are found for models that use higher complexity for the functional form (NN), in comparison to simpler models (LR). Moreover, if the inclusion of the sensitive attribute (A) in training the black box model aligns with the causal structure ($YA \mid C, L$) then A needs to be included in model training, and explanation disparities are smaller than if the inclusion of group information does not align with the causal structure, (if A is not included in the black box model training even though $YA \mid C, L$). Specifically for objectives 1 and 2, excluding the sensitive information aligns with the causal structure, and accordingly, models that include the sensitive attribute LR$_{A}$ and NN$_{A}$ have higher disparities when compared with models that exclude the sensitive attribute, i.e., LR$_{\not A}$ and NN$_{\not A}$, respectively as reflected in Figures 2(a), 2(b). Moreover, we observe that the highest magnitude of Δ_{Acc}^{group} and Δ_{Acc} is for NN and is much larger under conditions of covariate shift (4.52%, 2.25%) and concept shift (27.63%, 14.12%) than either sample size differences (1.6%, 0.82%) or omitted variables (1.6%,0.89%). While we specifically report Δ_{Acc} and Δ_{Acc}^{group} here, we observed similar behavior across Δ_{AUROC}^{group}, Δ_{Error}^{group}, not reported here for brevity.

Objective 1: Observations with respect to variation in sample size of disadvantaged group.

As the proportion of disadvantaged group samples increases, making their representation in the training set closer to the test set, explanation disparity metrics for NN$_{\not A}$ and LR$_{\not A}$ remain approximately consistent, illustrated in Figure 2(a) and Appendix Figure A1(a). For a proportion of 0.05 of the disadvantaged group in the training sample, Δ_{Acc}^{group} for NN$_{\not A}$ and LR$_{\not A}$ are 0.19% and 0.23%, respectively while for a proportion of 0.5 of the disadvantaged group in the training sample, Δ_{Acc}^{group} corresponds to 0.19% and 0.21%, respectively. However, when the sensitive attribute is used for model

[1] Code to replicate experiments is available at https://github.com/ChunaraLab/Disparities-Posthoc_Explanations

training in the case of NN$_A$ and LR$_A$, larger model explanation disparities result. Specifically, for 0.05 proportion of the disadvantaged group, Δ^{group}_{Acc} for NN$_A$ and LR$_A$ is 0.52% and 1.60%, respectively, an increase of 0.33% and 1.37% from NN$_{\not{A}}$ and LR$_{\not{A}}$. At 0.5 proportion of the disadvantaged group, Δ^{group}_{Acc} for NN$_A$ and LR$_A$ drop to 1.1% and 0.21%, respectively. Figure 2(a) and Appendix Figure A1(a) show these findings. Thus, disparities decrease for models that include the sensitive attribute but remain consistent for models that exclude the sensitive attribute. It also should be noted that the difference in black box model performance (accuracy) for the disadvantaged and advantaged groups is consistent for NN$_{\not{A}}$ and LR$_{\not{A}}$, but the difference decreases for NN$_A$ and LR$_A$ as the proportion of the disadvantaged group increases in the training sample, as illustrated in Appendix Figure A2(a).

Objective 2: Variation in covariate shift for the disadvantaged group

Introducing a covariate shift in L, specifically for the disadvantaged group, results in the training distribution of the disadvantaged group not being representative of the test distribution of the disadvantaged group. As the overlap between the training and test distributions of the disadvantaged group increases, overall disparities in the explanation metrics go down. Specifically, for 20% overlap, Δ^{group}_{Acc} for NN$_{\not{A}}$ and LR$_{\not{A}}$ are 0.84% and 0.05%, respectively, while for 100% overlap of the disadvantaged group between the training and test distributions, Δ^{group}_{Acc} corresponds to 0.18% and 0.08%, respectively. However, when the sensitive attribute is used for model training, explanation disparity in Δ^{group}_{Acc} and Δ_{Acc} is higher in comparison to when the sensitive attribute is excluded in model training, where there is incomplete overlap. At 20% overlap, Δ^{group}_{Acc} for NN$_A$ is 4.46% and for LR$_A$ is 2.2%, an increase of 3.62% and 2.15% from NN$_{\not{A}}$ and LR$_{\not{A}}$, respectively. However, at 100% overlap explanation disparities for NN$_A$ and LR$_A$ reduce considerably with Δ^{group}_{Acc} as 0.18% and 0.08%, respectively. At 100% overlap, explanation disparities for NN$_{\not{A}}$ and LR$_{\not{A}}$ are same as NN$_A$ and LR$_A$, respectively. This is illustrated in Figure 2(b) for Δ^{group}_{Acc} and a similar trend is observed in other metric, Δ_{Acc} as shown in Appendix Figure A1(b). The difference in black box model performance metric (accuracy) between the disadvantaged and advantaged groups is consistent for NN$_{\not{A}}$ and LR$_{\not{A}}$, but the difference decreases for NN$_A$ and LR$_A$ as the proportion of the disadvantaged group increases in the training sample as illustrated in Appendix Figure A2(b).

Objective 3: Variation in the magnitude of concept shift

As the magnitude of the concept shift is increased for the disadvantaged group from low to moderate to high, disparities in model explanations for Δ^{group}_{Acc} increase as presented in Figure 2(c). As the concept shift increases from moderate to high, Δ^{group}_{Acc} for LR$_A$ varies from 0.17% to 0.55% (an increase of 0.38%) while that for LR$_{\not{A}}$ varies from 0.09% to 0.04% (a decrease of 0.05%) . For NN$_A$, Δ^{group}_{Acc} increases from 1.47% to 5.92% (an increase of 4.45%). However, in the case of NN$_{\not{A}}$, Δ^{group}_{Acc} increases from 3.70% to 27.63% (an increase of 23.93%) as the concept shift increases from moderate to high. Similar trend is observed across Δ_{Acc} as shown in Appendix Figure A1(c). Thus, as concept shift increases, Δ^{group}_{Acc} and Δ_{Acc} increase for NN$_{\not{A}}$ considerably in comparison to NN$_A$, LR$_{\not{A}}$, and LR$_A$. As concept shift increases, the difference in black box model performance between the advantaged and disadvantaged groups for LR$_A$, LR$_{\not{A}}$, NN$_A$, NN$_{\not{A}}$ also increases as illustrated in Appendix Figure A2(c).

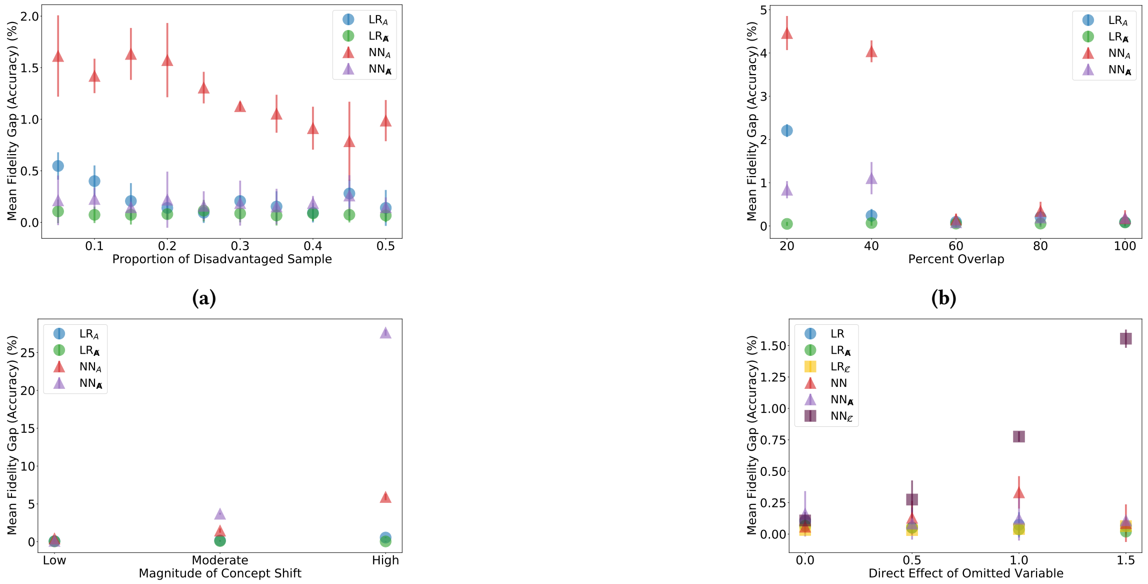
Objective 4: Variation in the direct effect of the omitted variable on the outcome

As the direct effect of the omitted variable C on the outcome increases (from 0 to 1.5), excluding C from model training increases disparities in Δ^{group}_{Acc} when compared to models that include C in model training as shown in Figure 2(d). For a direct effect of magnitude 0.5, Δ^{group}_{Acc} for NN that includes C in model training, represented as NN$_C$, is 0.13% while that of NN excluding C, represented as NN$_{\not{C}}$ is 0.27%. While for LR that includes C in model training, LR$_C$, Δ^{group}_{Acc} is 0.05% for a direct effect of magnitude 0.5, and for LR that excludes C in model training, LR$_{\not{C}}$, Δ^{group}_{Acc} is 0.03. On the contrary, for a direct effect of magnitude 1.5, Δ^{group}_{Acc} for NN$_C$, is 0.09% while that of NN$_{\not{C}}$ is 1.55%, a decrease of 0.04% and an increase of 1.28% in comparison to a direct effect of magnitude 0.5, respectively. While for LR$_C$, Δ^{group}_{Acc} is 0.07% for a direct effect of magnitude 1.5, and for LR$_{\not{C}}$, Δ^{group}_{Acc} is 0.06, an increase of 0.02% and 0.03% from a direct effect of magnitude 0.5, respectively. Thus, Δ^{group}_{Acc} for NN$_{\not{C}}$ increases as the direct effect of C on Y increases. This characteristic is also observed across Δ_{Acc} as shown in Appendix Figure A1(d). The difference in black box model performance of NN$_{\not{C}}$ is higher than that of NN$_C$, when the direct effect of C on Y is non-zero, precisely for values, 0.5,1.0, and 1.5, as shown in Appendix Figure A2(d).

5.2 Explanation Disparities in Real-World Dataset

In the case of the Adult dataset, increasing the percentage of the disadvantaged group (males) in the training sample shows a decrease in Δ^{group}_{Acc} for LR$_{\not{A}}$ and LR$_A$ while it remains consistent for NN$_A$ and shows a very slight decrease in case of NN$_{\not{A}}$. This is illustrated in Appendix Figure A3(b). Similar characteristics can be seen for Δ_{Acc} in Appendix Figure A3(a). Specifically, for 5% of the disadvantaged group in the training sample, Δ^{group}_{Acc} for LR$_{\not{A}}$ and LR$_A$ is 3.77% and 2.11% respectively that decreases to 1.40% and 1.13% for 50% of the disadvantaged group in the training sample. This corresponds to a decrease of 2.37% for LR$_A$ and 0.98 for LR$_{\not{A}}$. While for NN$_A$, this decrease corresponds to 0.03%, and for NN$_{\not{A}}$, it is 0.91% as the percentage of the disadvantaged group (males) increases from 5% to 50%. This differs from the simulation, where explanation disparities in Δ^{group}_{Acc} remain consistent for LR$_{\not{A}}$ and NN$_{\not{A}}$ decrease for LR$_A$ and NN$_A$ whereas in case of Adult, Δ^{group}_{Acc} for LR$_{\not{A}}$, LR$_A$, and NN$_{\not{A}}$ decrease.

Introducing a covariate shift in L 'hours-per-week,' for the disadvantaged group (males), results in the training distribution of the disadvantaged group not being representative of the test distribution of the disadvantaged group. Δ^{group}_{Acc} increases from 0.46% to 0.53% for LR$_A$, 0.50% to 2.93% for NN$_A$ as percent overlap increases from 20 to 100. On the contrary, Δ^{group}_{Acc} for LR$_{\not{A}}$ decreases from 2.14% to 0.41%, and for NN$_{\not{A}}$ it decreases from 4.61% to 2.94% as the percent overlap increases from 20 to 100. This can be seen in



(a)

(b)

(c)

(d)

Figure 2: Percent Mean Fidelity Gap (Accuracy) of LIME with 95% confidence intervals applied to models built on the synthetic datasets generated for (a) objective 1 - sample size, (b) objective 2 - covariate shift, (c) objective 3 - concept shift, and (d) objective 4 - omitted variables. In (a), we vary the proportion of the disadvantaged group in the training set sample. In (b), we introduce a covariate shift for the disadvantaged group, shifting the overlap between the train and test distributions; and in (c), we vary the magnitude of the concept shift. In (d), we adjust the strength of the direct effect of the omitted variable C. The models that are considered are LR with A, LR_A in blue, LR without A, $\mathrm{LR}_{\cancel{A}}$ in green, NN with A, NN_A in red, and NN without A, $\mathrm{NN}_{\cancel{A}}$ in violet, LR without C, $\mathrm{LR}_{\cancel{C}}$ in yellow, and NN without C, $\mathrm{NN}_{\cancel{C}}$ in plum. Circles represent linear models, and triangles represent neural network models. Notice that the magnitude of the mean fidelity gap is much larger under conditions of (b) covariate shift and (c) concept shift than either (a) sample size differences or (d) omitted variables.

Model	Δ_{Acc}	$\Delta^{group}_{\text{Accuracy}}$
LR_A	0.021 (0.020, 0.021)	0.063 (0.062, 0.064)
$\mathrm{LR}_{\cancel{A}}$	0.015 (0.015, 0.015)	0.046 (0.045, 0.047)
NN_A	0.025 (0.025, 0.026)	0.078 (0.077, 0.079)
$\mathrm{NN}_{\cancel{A}}$	0.022 (0.022, 0.022)	0.067 (0.066, 0.068)

Table 1: Maximum Fidelity Gap (Δ_{Acc}) and Mean Fidelity Gap in Accuracy (Δ^{group}_{Acc}) with (95% Confidence interval) for Adult dataset for LR_A, $\mathrm{LR}_{\cancel{A}}$, NN_A, $\mathrm{NN}_{\cancel{A}}$ for concept shift between 'hours-per-week' and 'income' for male group for Adult (objective 3).

Model	Δ_{Acc}	$\Delta^{group}_{\text{Accuracy}}$
LR_C	0.006 (0.005, 0.006)	0.017 (0.014, 0.020)
$\mathrm{LR}_{\cancel{C}}$	0.004 (0.004, 0.005)	0.013 (0.011, 0.015)
NN_C	0.027 (0.027, 0.028)	0.083 (0.081, 0.084)
$\mathrm{NN}_{\cancel{C}}$	0.023 (0.022, 0.024)	0.070 (0.068, 0.072)

Table 2: Maximum Fidelity Gap (Δ_{Acc}) and Mean Fidelity Gap in Accuracy (Δ^{group}_{Acc}) with (95% Confidence interval) for Adult dataset for LR with 'Nationality' included LR_C, LR with 'Nationality' excluded $\mathrm{LR}_{\cancel{C}}$, NN with 'Nationality' included NN_C and NN with 'Nationality' excluded $\mathrm{NN}_{\cancel{C}}$ for Adult (objective 4).

Appendix Figure A4(b). This differs from simulation, where increasing overlap decreases Δ^{group}_{Acc} for LR_A NN_A but Δ^{group}_{Acc} remains consistent for $\mathrm{LR}_{\cancel{A}}$ and $\mathrm{NN}_{\cancel{A}}$ while in case of Adult, Δ^{group}_{Acc} increases for LR_A, NN_A but decreases for $\mathrm{LR}_{\cancel{A}}$ and $\mathrm{NN}_{\cancel{A}}$. In the case of concept shift, excluding the sensitive attribute, 'gender,' A for model training results in a 1.5% explanation disparity Δ^{group}_{Acc} for $\mathrm{LR}_{\cancel{A}}$ and 2.2% Δ^{group}_{Acc} for $\mathrm{NN}_{\cancel{A}}$. Including the sensitive attribute results in an explanation disparity of 2.1% for LR_A (an increase of 0.6% from $\mathrm{LR}_{\cancel{A}}$ and 2.5% for NN_A (an increase of 0.3% for NN_A). This is illustrated in Table 1. In the case of simulations, excluding A results in higher disparities for $\mathrm{LR}_{\cancel{A}}$and $\mathrm{NN}_{\cancel{A}}$ in comparison to including A, an opposite trend compared to Adult. Regarding omitted variable bias, excluding 'Nationality' C, which has a direct effect on 'income' Y, results in an explanation disparity in Δ^{group}_{Acc} of 2.29% in comparison to including C in model training, with an explanation disparity of 2.7% (a difference of 0.49%) for NN. While for LR excluding 'Nationality' results in 0.52% explanation disparity,

Δ^{group}_{Acc} in comparison to including it with an explanation disparity, Δ^{group}_{Acc} of 0.55% (lower by 0.03%). This result is illustrated in Table 2. Excluding C results in lower disparities for Adult but higher disparities in the case of simulation. Specifically, in the simulations, excluding C results in an explanation disparity in Δ^{group}_{Acc} of 1.55% in comparison to including it 0.09% (a decrease of 1.46%) for NN and 0.06% in explanation disparity in Δ^{group}_{Acc} for excluding in LR in comparison to 0.06% while including C. Since we cannot vary concept shift and the direct effect of 'Nationality' on the outcome 'Income,' as opposed to the variations in the case of simulations, we only report one value for Δ^{group}_{Acc}, and Δ_{Acc} for all the models for Adult in Tables 1 and 2 but present multiple values for the simulations in Figures 2(c) and 2(d) for objectives 3 and 4, respectively. Black box model performance for Adult is presented in Appendix Figures A5(a) and A5(b) for objectives 1 and 2, and Appendix Tables A1 and A2 for objectives 3 and 4. These results show that there is a larger difference in the accuracy of black box model performance between groups of the sensitive attribute (males and females) for Adult, 13% for NN, and 13.8% for LR in comparison to the simulated data, with 0.10% for NN and 0.18% for LR.

6 DISCUSSION

Our study is the first to examine disparities in LIME explanation method, focusing on the properties of the data and black box model. We examine sample size, covariate shift, concept shift, and omitted variable bias, along with the inclusion of the sensitive attribute in the black box model training and the complexity of the black box model. Our findings show that explanation disparities for both explanation fidelity metrics tested: Maximum Fidelity Gap in accuracy Δ_{Acc} and Mean Fidelity Gap in accuracy Δ^{group}_{Acc} can be based on the characteristics of the data as well as the modeling method. In systematic simulation results, we found no change in Δ_{Acc} and Δ^{group}_{Acc} with increasing disadvantaged samples for models that exclude the sensitive attribute if the exclusion of the sensitive attribute aligns with the causal graph. If the inclusion of the sensitive attribute does not align with the causal graph, Δ_{Acc} and Δ^{group}_{Acc} may depend on the proportion of the disadvantaged group in the training sample. Here, the black box model is likely to learn spurious correlations between the sensitive attribute and outcome, which can affect Δ_{Acc} and Δ^{group}_{Acc}. For limited overlap in the distribution of the disadvantaged sample between the training and test distributions, explanation disparities are higher compared to complete overlap. With a lower overlap, the black box model may not generalize well for the disadvantaged group in the test distribution, resulting in higher disparities in model explanations for Δ_{Acc} and Δ^{group}_{Acc}. Moreover, as the concept shift increases, resulting in a nonlinear relationship between the sensitive attribute and the outcome, linear models that are unable to capture this non-linearity actually have lower Δ_{Acc} and Δ^{group}_{Acc}compared to complex neural network models. If the sensitive attribute is excluded from the training of the complex black-box model, not aligning with the causal structure, Δ_{Acc} and Δ^{group}_{Acc} are higher than if the sensitive attribute is included in the model training for concept shift. Excluding the sensitive attribute may mask its importance in the predictions of the black box model, resulting in higher explanation disparities. Omitting a variable that has a direct effect on the outcome may lead to a higher Δ_{Acc} and Δ^{group}_{Acc} as the magnitude of the direct effect increases, especially for complex models. As the variable has a direct effect on the outcome, excluding it may mask the importance of the variable in predictions of the black box model, resulting in an increase in explanation disparities.

We find that disparities in model explanations for the Adult dataset reflect the issue of $P(Y \mid L, C, A)! = P(Y \mid L, C)$ in real-world datasets, where it is easier to predict the outcome for the advantaged group rather than the disadvantaged group. Meanwhile, for the synthetic simulations presented here, it is equally easy to predict across the disadvantaged and advantaged groups.

In general, in simulations, explanation disparities are higher for models that include the sensitive attribute in comparison to models that exclude the sensitive attribute for objectives 1 and 2. An opposite trend is observed in the case of Adults, where explanation disparities are higher when the sensitive attribute is excluded compared to when it is included. Including the sensitive attribute aligns with the causal graph for Adult but not for simulations which may explain the differences in the observed behavior between simulations and Adult. Specifically, as the simpler linear models are not able to capture the nonlinear relationship in Adult, their overall performance increases with an increased proportion of disadvantaged group samples, resulting in a decrease in Δ_{Acc} and Δ^{group}_{Acc}, irrespective of whether the sensitive attribute is included or not. This differs from simulations where Δ_{Acc} and Δ^{group}_{Acc} only decrease for the models that include the sensitive attribute. We posit that as our simulations comprised linear functional forms, explanation disparities had a similar trend between linear and neural network models; however, in the case of Adults, since the functional form is nonlinear, explanation disparities for the linear models follow a different trend than neural networks. For concept shift, including the sensitive attribute in the black box model training for Adults has a higher Δ_{Acc} and Δ^{group}_{Acc} than excluding it in the case of neural networks for concept shift. On the contrary, in simulations, excluding the sensitive attribute resulted in higher Δ_{Acc} and Δ^{group}_{Acc}. Excluding the sensitive attribute may result in inaccurate explanations of the black box predictions, especially for complex models such as neural networks. A similar characteristic is observed in Adult when 'Nationality' is excluded to assess omitted variable bias, as excluding it can result in inaccurate explanations, especially for the disadvantaged group similar to the simulations. For mitigating disparities in the explanations in the case of datasets like Adult, our analyses reinforce the need to focus on improving the quality of the data and ensure that the complexity in the data is adequately captured by the black box model.

Need for benchmark datasets for developing fair explanation methods

Given the importance of data, benchmark datasets for assessing explanation disparity metrics would help but currently do not exist. For developing such benchmark datasets, knowledge of the causal graph can aid in understanding if including sensitive information is relevant to the task and can also highlight which variables can be omitted in the model training to ensure explanations are accurate, especially those that do not directly affect the outcome. Further, systematically designed benchmark datasets that allow for varying complexities in the functional form between the covariates and the outcome will be useful in order to assess the explanation disparities of black box models with varying complexity. Finally, benchmark datasets can be used to assess how different the test distributions can be from the training distributions to ensure that explanation disparities are within the desired range. For example, in the simulations, around 60% overlap results in a considerable drop in explanation disparities in comparison to 20% overlap. These types of examinations can help demonstrate how well a particular explanation method generalizes based on the overlap between train and test distributions.

Limitations of the study

While we highlight the potential factors in the data-generating process and model training that can result in explanation disparities, our study concentrates on the LIME explanation method. Although LIME has widespread use, and previous research focuses on disparities in LIME explanations, other explanation methods, such as SHAP, may also exhibit disparity challenges that depend on the properties of the data and the black box model. Future work should extend our investigation to other explanation methods, such as SHAP. Moreover, LIME explanations have inherent limitations, such as uncertainty in perturbation processes. Additionally, the computational cost of LIME and the selection of hyperparameters like kernel width and regularization parameters are crucial to LIME and can influence explanations. Prior methods developed for addressing these challenges also need to be audited with respect to data and model properties. While we restrict to simple causal graphs without unmeasured confounding between the sensitive attribute and the outcome, further efforts to include unmeasured confounding can provide additional insights. However, in considering the properties of the data and models, this work takes an important initial step towards incorporating broader aspects into the assessment of explanation disparities.

ACKNOWLEDGMENTS

This work was supported by grant 1845487 from the National Science Foundation. VM would like to acknowledge funding through the Google PhD fellowship.

REFERENCES

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *Advances in neural information processing systems* 31 (2018).

[2] Julius Adebayo, Michael Muelly, Hal Abelson, and Been Kim. 2022. Post hoc explanations may be ineffective for detecting unknown spurious correlation. *arXiv preprint arXiv:2212.04629* (2022).

[3] Ibrahim Adeshola and Adeola Praise Adepoju. 2023. The opportunities and challenges of ChatGPT in education. *Interactive Learning Environments* (2023), 1–14.

[4] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. 2018. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*. 559–560.

[5] Johannes Allgaier, Lena Mulansky, Rachel Lea Draelos, and Rüdiger Pryss. 2023. How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare. *Artificial Intelligence in Medicine* 143 (2023), 102616.

[6] Aparna Balagopalan, Haoran Zhang, Kimia Hamidieh, Thomas Hartvigsen, Frank Rudzicz, and Marzyeh Ghassemi. 2022. The road to explainability is paved with bias: Measuring the fairness of explanations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1194–1206.

[7] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

[8] Tom Begley, Tobias Schwedes, Christopher Frye, and Ilya Feige. 2020. Explainability for fair machine learning. *arXiv preprint arXiv:2010.07389* (2020).

[9] Jacob Bien and Robert Tibshirani. 2009. Classification by set cover: The prototype vector machine. *arXiv preprint arXiv:0908.2284* (2009).

[10] Nadia Burkart and Marco F Huber. 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research* 70 (2021), 245–317.

[11] Niklas Bussmann, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. 2021. Explainable machine learning in credit risk management. *Computational Economics* 57 (2021), 203–216.

[12] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 1721–1730.

[13] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. 2021. Ethical machine learning in healthcare. *Annual review of biomedical data science* 4 (2021), 123–144.

[14] Richard J Chen, Judy J Wang, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Ming Y Lu, Sharifa Sahai, and Faisal Mahmood. 2023. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature biomedical engineering* 7, 6 (2023), 719–742.

[15] Silvia Chiappa. 2019. Path-specific counterfactual fairness. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 7801–7808.

[16] Mark Craven and Jude Shavlik. 1995. Extracting tree-structured representations of trained networks. *Advances in neural information processing systems* 8 (1995).

[17] Jessica Dai, Sohini Upadhyay, Ulrich Aivodji, Stephen H Bach, and Himabindu Lakkaraju. 2022. Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 203–214.

[18] Jessica Dai, Sohini Upadhyay, Stephen H Bach, and Himabindu Lakkaraju. 2021. What will it take to generate fairness-preserving explanations? *arXiv preprint arXiv:2106.13346* (2021).

[19] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[20] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. 2018. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 0210–0215.

[21] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaao5580.

[22] Dheeru Dua, Casey Graff, et al. 2017. UCI machine learning repository. (2017).

[23] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery* 36, 6 (2022), 2074–2152.

[24] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[25] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. 2020. A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings* 2020 (2020), 191.

[26] Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? *arXiv preprint arXiv:2005.01831* (2020).

[27] Nathan Kallus and Angela Zhou. 2018. Residual unfairness in fair machine learning from prejudiced data. In *International Conference on Machine Learning*. PMLR, 2439–2448.

[28] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*. Springer, 35–50.

[29] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. 2020. Model-agnostic counterfactual explanations for consequential decisions. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 895–905.

[30] Giona Kleinberg, Michael J Diaz, Sai Batchu, and Brandon Lucke-Wold. 2022. Racial underrepresentation in dermatological datasets leads to biased machine learning models and inequitable healthcare. *Journal of biomed research* 3, 1 (2022), 42.

[31] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1675–1684.

[32] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. 2019. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. *arXiv preprint arXiv:1907.09294* (2019).

[33] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, and David Madigan. 2015. Interpretable classifiers using rules and bayesian analysis: Building a better

stroke prediction model. (2015).

[34] Moshe Lichman et al. 2013. UCI machine learning repository.

[35] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2020. Explainable ai: A review of machine learning interpretability methods. *Entropy* 23, 1 (2020), 18.

[36] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. 2021. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*. PMLR, 6781–6792.

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

[38] Fernando Martínez-Plumed, Cèsar Ferri, David Nieves, and José Hernández-Orallo. 2019. Fairness and missing values. *arXiv preprint arXiv:1905.12728* (2019).

[39] Vishwali Mhasawade and Rumi Chunara. 2021. Causal multi-level fairness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 784–794.

[40] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[41] Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. 2012. A unifying view on dataset shift in classification. *Pattern recognition* 45, 1 (2012), 521–530.

[42] Hongseok Namkoong, Steve Yadlowsky, et al. 2023. Diagnosing Model Performance Under Distribution Shift. *arXiv preprint arXiv:2303.02011* (2023).

[43] Dana Pessach and Erez Shmueli. 2023. Algorithmic fairness. In *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*. Springer, 867–886.

[44] Stephen Robert Pfohl, Natalie Harris, Chirag Nagpal, David Madras, Vishwali Mhasawade, Olawale Elijah Salaudeen, Katherine A Heller, Sanmi Koyejo, and Alexander Nicholas D'Amour. 2023. Understanding subgroup performance differences of fair predictors using causal models. In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*.

[45] Gregory Plumb, Denali Molitor, and Ameet S Talwalkar. 2018. Model agnostic supervised local explanations. *Advances in neural information processing systems* 31 (2018).

[46] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[47] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[48] María Agustina Ricci Lara, Candelaria Mosquera, Enzo Ferrante, and Rodrigo Echeveste. 2023. Towards Unraveling Calibration Biases in Medical Image Analysis. In *Workshop on Clinical Image-Based Procedures*. Springer, 132–141.

[49] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.

[50] Amit Sangroya, Mouli Rastogi, C Anantaram, and Lovekesh Vig. 2020. Guided-LIME: Structured Sampling based Hybrid Approach towards Explaining Blackbox Machine Learning Models.. In *CIKM (Workshops)*.

[51] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[52] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. 2021. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 3–13.

[53] Dylan Slack, Anna Hilgard, Sameer Singh, and Himabindu Lakkaraju. 2021. Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in neural information processing systems* 34 (2021), 9391–9404.

[54] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825* (2017).

[55] Adarsh Subbaswamy and Suchi Saria. 2018. Counterfactual Normalization: Proactively Addressing Dataset Shift Using Causal Mechanisms.. In *UAI*. 947–957.

[56] Adarsh Subbaswamy and Suchi Saria. 2020. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics* 21, 2 (2020), 345–352.

[57] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. 2018. Distill-and-compare: Auditing black-box models using transparent model distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 303–310.

[58] Lukas Tuggener, Mohammadreza Amirian, Katharina Rombach, Stefan Lörwald, Anastasia Varlet, Christian Westermann, and Thilo Stadelmann. 2019. Automated machine learning in practice: state of the art and recent results. In *2019 6th Swiss Conference on Data Science (SDS)*. IEEE, 31–36.

[59] Mythreyi Velmurugan, Chun Ouyang, Catarina Moreira, and Renuka Sindhgatta. 2021. Developing a fidelity evaluation approach for interpretable machine learning. *arXiv preprint arXiv:2106.08492* (2021).

[60] Darshali A Vyas, Leo G Eisenstein, and David S Jones. 2020. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. , 874–882 pages.

[61] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

[62] Fulton Wang and Cynthia Rudin. 2015. Falling rule lists. In *Artificial intelligence and statistics*. PMLR, 1013–1022.

[63] Yanchen Wang and Lisa Singh. 2021. Analyzing the impact of missing values and selection bias on fairness. *International Journal of Data Science and Analytics* 12, 2 (2021), 101–119.

[64] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564* (2023).

[65] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579* (2015).

[66] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*. PMLR, 962–970.

[67] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. 2017. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society Series A: Statistics in Society* 180, 3 (2017), 689–722.

[68] Yijun Zhao, Xiaoyu Chen, Haoran Xue, and Gary M Weiss. 2023. A machine learning approach to graduate admissions and the role of letters of recommendation. *Plos one* 18, 10 (2023), e0291107.

[69] Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. 2021. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics* 10, 5 (2021), 593.

[70] Zhengze Zhou, Giles Hooker, and Fei Wang. 2021. S-lime: Stabilized-lime for model explanation. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 2429–2438.

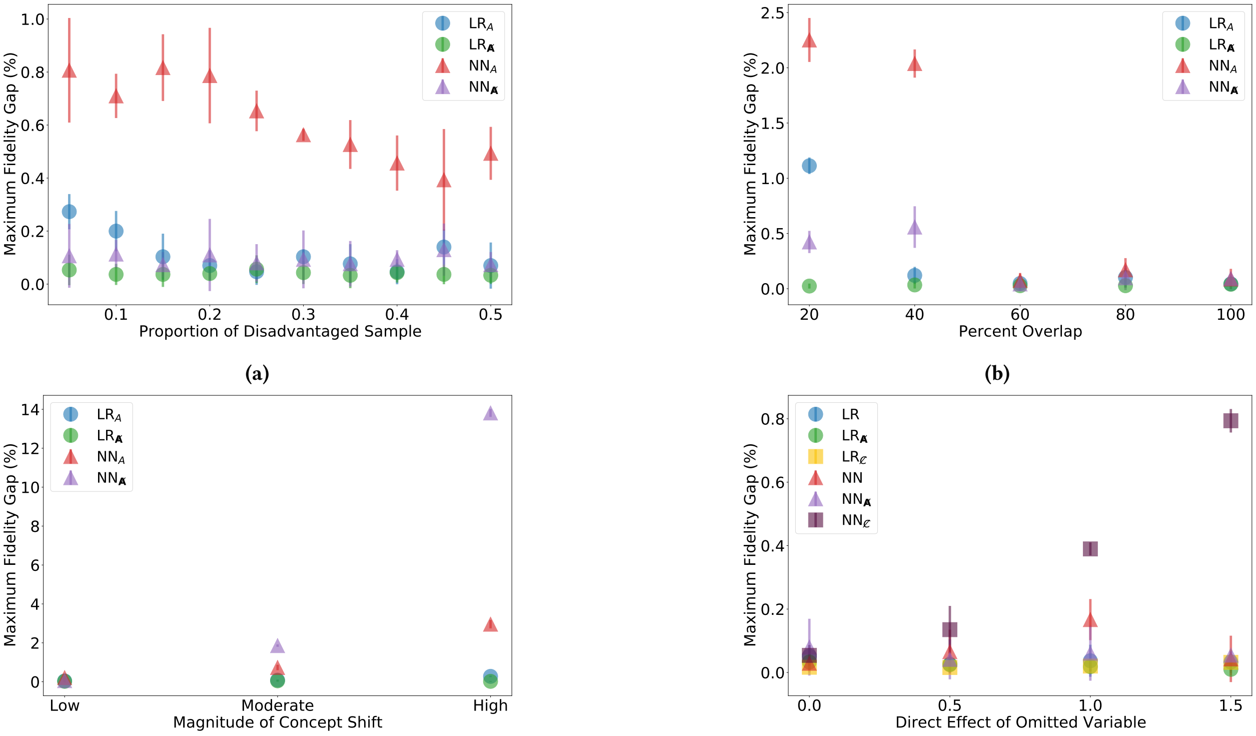
A APPENDIX

A.1 Additional Performance Metric Results for Simulation

Results for Δ_{Acc} for all 4 objectives are presented in Appendix Figure A1. We also provide detailed results for the simulation pertaining to the black box model performance for all 4 objectives in Appendix Figure A2.

A.2 Detailed Results for Adult

Here, we present results for Explanation Fidelity Metrics, Δ_{Acc}^{group} and Δ_{Acc} for objective 1 (varying percentage of the disadvantaged group, males in the training distribution) and objective 2 (varying overlap in the distribution of 'hours-per-week,' L for the disadvantaged group, males between the training and test distributions in Appendix Figures A3 and A4, respectively. Moreover, we provide additional results for the Adult dataset for the black box model performance, a disparity in the prediction accuracy of the black box model between the advantaged and disadvantaged groups across all 4 objectives. We present these prediction disparities in Appendix Figures A5(a), A5(b) for objectives 1, 2 and in Tables A1 and A2 for 3, and 4, respectively.



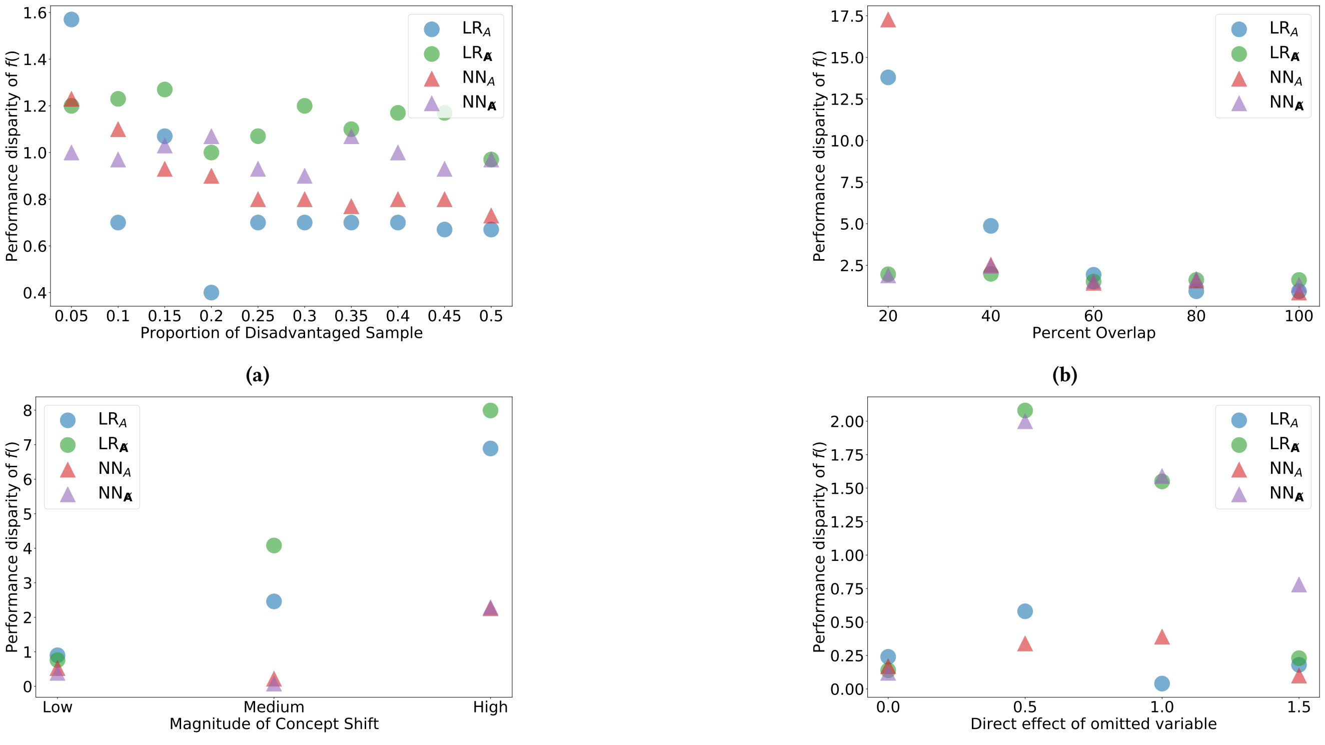
(a)

(b)

(c)

(d)

Figure A1: Percent Maximum Fidelity Gap of LIME applied to models built on the synthetic datasets generated for (a) objective 1 - sample size, (b) objective 2 - covariate shift, (c) objective 3 - concept shift, and (d) objective 4 - omitted variables for LR with A, LR$_A$ in blue, LR without A, LR$_{\not A}$ in green, NN with A, NN$_A$ in red, and NN without A, NN$_{\not A}$ in violet, LR without C, LR$_{\not C}$ in yellow, and NN without C, NN$_{\not C}$ in plum. Circles represent linear models, and triangles represent neural network models.



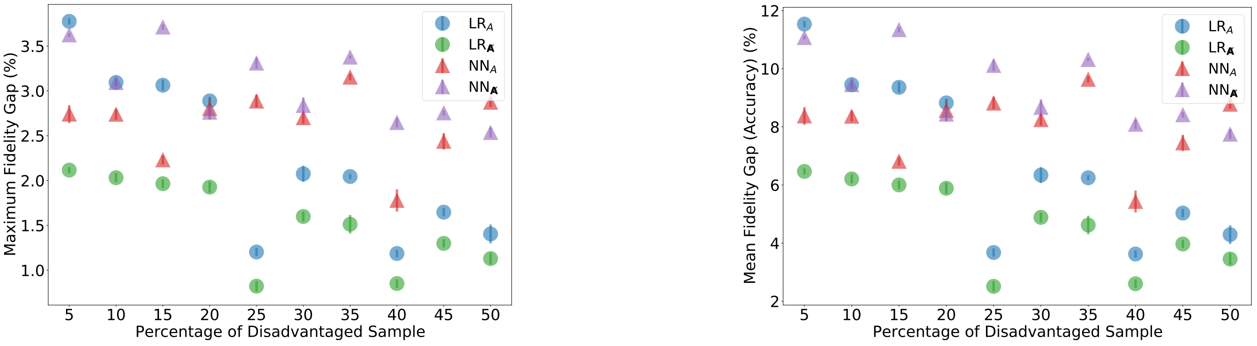
(a)

(b)

(c)

(d)

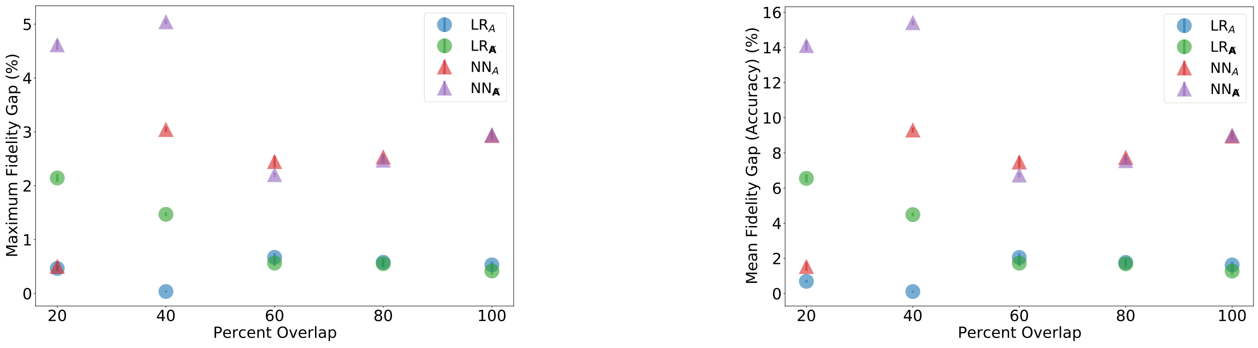
Figure A2: Performance disparity of $f()$ calculated as Accuracy for A = 1 - Accuracy for A = 0 on the synthetic datasets generated with an increasing (a) proportion of the disadvantaged sample (objective 1), (b) overlap between the distribution of L for $A = 0$ between training and test distributions, (c) concept shift, and (d) direct effect of omitted variable C for LR with A, LR$_A$ in blue, LR without A, LR$_{\not A}$ in green, NN with A, NN$_A$ in red, and NN without A, NN$_{\not A}$ in violet. Circles represent linear models, and triangles represent neural network models.



(a)

(b)

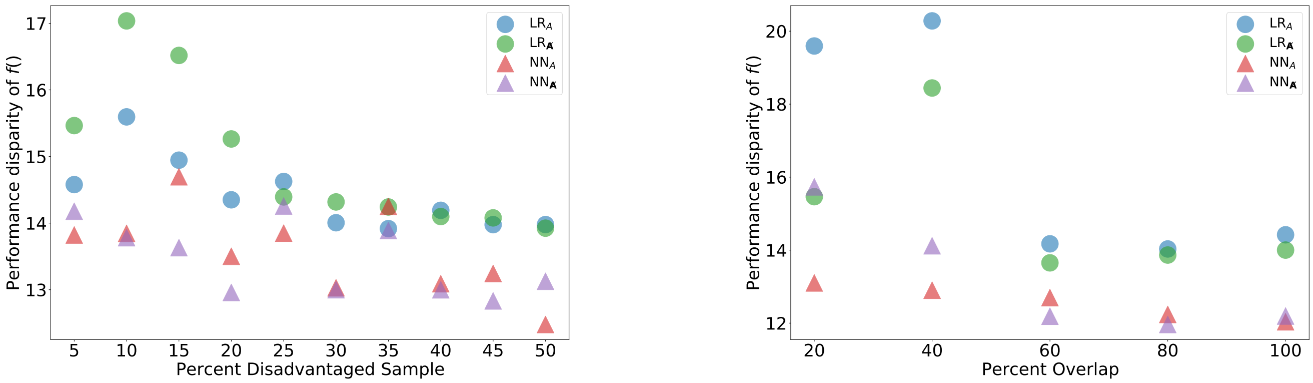
Figure A3: (a) Percent Maximum Fidelity Gap, Δ_{Acc}, (b) mean fidelity gap in accuracy, Δ_{Acc}^{group} of LIME on Adult dataset with variation in the proportion of the 'males' (A) in the training sample (objective 1) for LR with A, LR$_A$ in blue, LR without A, LR$_{\not A}$ in green, NN with A, NN$_A$ in red, and NN without A, NN$_{\not A}$ in violet. Circles represent linear models, and triangles represent neural network models.



(a)

(b)

Figure A4: (a) Percent Maximum Fidelity Gap, Δ_{Acc}, (b) mean fidelity gap in accuracy, Δ_{Acc}^{group} of LIME on the Adult dataset with variation in the overlap (covariate shift) in the distribution of the 'males' (A) in the training sample and the test set (objective 2) for LR with A, LR$_A$ in blue, LR without A, LR$_{\not A}$ in green, NN with A, NN$_A$ in red, and NN without A, NN$_{\not A}$ in violet. Circles represent linear models, and triangles represent neural network models.



(a)

(b)

Figure A5: Performance disparity of $f()$, black box model, calculated as Accuracy for A = 1 - Accuracy for A = 0 for Adult (a) with varying proportion of 'male' group (objective 1), and (b) with varying overlap (covariate shift in 'hours-per-week,' L) between train and test distribution for disadvantaged 'male' group (objective 2). Circles represent linear models, and triangles represent neural network models.

Model	$\mathrm{Acc}_{A=0}$	$\mathrm{Acc}_{A=1}$	$\lvert \mathrm{Acc}_{A=1} - \mathrm{Acc}_{A=1} \rvert$
LR_{C}	76.36	90.69	14.33
$\mathrm{LR}_{\not{C}}$	76.64	90.87	14.23
NN_{C}	78.78	91.33	12.55
$\mathrm{NN}_{\not{C}}$	79.00	91.92	12.92

Table A2: Black box model performance with respect to percentage accuracy for disadvantaged ($A = 0$) and advantaged ($A = 1$) groups with the difference in accuracy across groups for LR with 'Nationality' included LR_{C}, LR with 'Nationality' excluded $\mathrm{LR}_{\not{C}}$, NN with 'Nationality' included NN_{C}, and NN with 'Nationality' excluded $\mathrm{NN}_{\not{C}}$ for Adult (objective 4).

Model	$\mathrm{Acc}_{A=0}$	$\mathrm{Acc}_{A=1}$	$\lvert \mathrm{Acc}_{A=1} - \mathrm{Acc}_{A=1} \rvert$
LR_{A}	76.41	90.21	13.80
$\mathrm{LR}_{\not{A}}$	76.30	90.50	14.21
NN_{A}	78.10	91.10	13.01
$\mathrm{NN}_{\not{A}}$	78.30	91.50	13.22

Table A1: Black box model performance with respect to percentage accuracy for disadvantaged ($A = 0$) and advantaged ($A = 1$) groups with the difference in accuracy across groups for LR_{A}, $\mathrm{LR}_{\not{A}}$, NN_{A}, $\mathrm{NN}_{\not{A}}$ for concept shift between 'hours-per-week' and 'income' for male group for Adult (objective 3)