Fairness via Explanation Quality: Evaluating Disparities in the Quality of Post hoc Explanations

Jessica Dai  
Brown University  
jessica.dai@alumni.brown.edu

Sohini Upadhyay  
Harvard University  
supadhyay@g.harvard.edu

Stephen H. Bach  
Brown University  
stephen_bach@brown.edu

Ulrich Aïvodji  
Université du Québec à Montréal  
aivodji.ulrich@uqam.ca

Himabindu Lakkaraju  
Harvard University  
hlakkaraju@hbs.edu

ABSTRACT
As post hoc explanation methods are increasingly being leveraged to explain complex models in high-stakes settings, it becomes critical to ensure that the quality of the resulting explanations is consistently high across all subgroups of a population. For instance, it should not be the case that explanations associated with instances belonging to, e.g., women, are less accurate than those associated with other genders. In this work, we initiate the study of identifying group-based disparities in explanation quality. To this end, we first outline several key properties that contribute to explanation quality—namely, fidelity (accuracy), stability, consistency, and sparsity—and discuss why and how disparities in these properties can be particularly problematic. We then propose an evaluation framework which can quantitatively measure disparities in the quality of explanations. Using this framework, we carry out an empirical analysis with three datasets, six post hoc explanation methods, and different model classes to understand if and when group-based disparities in explanation quality arise. Our results indicate that such disparities are more likely to occur when the models being explained are complex and non-linear. We also observe that certain post hoc explanation methods (e.g., Integrated Gradients, SHAP) are more likely to exhibit disparities. Our work sheds light on previously unexplored ways in which explanation methods may introduce unfairness in real-world decision making.

CCS CONCEPTS
• General and reference → Evaluation; Empirical studies; • Computing methodologies → Machine learning.

KEYWORDS
explainable machine learning, interpretability, fairness, robustness

1 INTRODUCTION
As machine learning (ML) models are increasingly being deployed to make consequential decisions in domains such as healthcare, finance, and policy, there is a growing need to ensure interpretable to ML practitioners and other domain experts (e.g., doctors, policy makers). Only if these practitioners have a clear picture of the behavior of these models can they assess when and how much to rely on them, and detect systematic errors and potential biases in them [28]. However, the increasing complexity as well as the proprietary nature of predictive models make it challenging to understand these complex black boxes, thus motivating the need for tools that can explain them in a faithful and human-interpretable manner. To this end, several techniques have been proposed to explain complex models in a post hoc fashion.

Owing to their generality, post hoc explanation methods are increasingly being used to explain a number of complex models in high-stakes domains such as medicine, finance, law, and science [30, 36, 76]. Therefore, it becomes critical to ensure that the explanations generated by these methods are of high quality. Prior research has studied several notions of explanation quality, including fidelity, stability, consistency, and sparsity [49, 56, 65, 80]. In this work, we focus on disparities in the quality of explanations corresponding to different population subgroups.

To illustrate, consider a healthcare setting in which a predictive model is being used to aid doctors in diagnosing diseases (Figure 1). Suppose the model incorrectly relies on spurious features (e.g., hospital name, zip code) to make predictions. Let us assume that post hoc explanations generated by a state-of-the-art method are being provided to doctors to explain each model prediction. Further, explanations corresponding to men are accurate, and rightly capture the underlying model is relying on spurious features. Therefore, doctors who examine these explanations are likely to mistrust the model in the case of male patients and make better decisions by employing their own discretion. On the other hand, explanations for women are comparatively less accurate, and are incorrectly suggesting that the underlying model is relying on relevant features (e.g., fever, heart rate). So, doctors may trust model predictions in the case of female patients due to the misleading explanations, and may therefore make incorrect decisions leading to unfairness (women are more likely to be incorrectly diagnosed than men) in the overall decision-making process.
Figure 1: Illustrative example highlighting how disparities in explanation quality might lead to sub-optimal decisions, in turn resulting in worse real world outcomes for specific subgroups.

The above scenario, while stylized, demonstrates how disparities in the quality of explanations may induce unfairness in real world decision making, and negatively impact the outcomes of vulnerable populations. Therefore, it is important to assess the extent to which such group-based disparities occur in the quality of the explanations output by state-of-the-art methods. However, there is little to no research that focuses on this critical problem.

In this work, we initiate the study of group-based disparities in explanation quality. More specifically, we make the following key contributions:

- We formulate the problem of detecting group-based disparities in explanation quality.
- We propose a novel evaluation framework which can quantitatively measure disparities in the quality of explanations output by state-of-the-art methods. We show an example of how to use this framework and some initial directions for inquiry using existing common metrics—specifically, fidelity disparity, stability disparity, consistency disparity, and sparsity disparity.
- We leverage the aforementioned framework to carry out a rigorous empirical analysis with six state-of-the-art post hoc explanation methods, three real world datasets, and two different model classes (linear models and deep neural networks) to study if and when group-based disparities in explanation quality arise.

Results from our empirical analysis indicate that disparities in explanation quality are more likely to occur when the models being explained are complex and highly non-linear. Furthermore, we also observed that post hoc explanation methods such as Integrated Gradients and SHAP are more likely to generate explanations that are prone to the aforementioned disparities. Our analysis also revealed that group-based disparities are most prominent in case of two notions of explanation quality—namely, fidelity and sparsity. Our findings not only shed light on the previously unexplored problem of group-based disparities in explanation quality, but also pave the way for rethinking the design and development of explanation methods in a way that such disparities can be minimized.

2 RELATED WORK

This work builds on extensive work in explainable machine learning. Crucially, the usage of explanations is often motivated by higher-stakes settings, where the interpretation of the explanation has influence on how a human chooses to interact with the model. Common use-cases for explanations across a variety of stakeholders include model debugging and improvement, monitoring, building confidence, transparency, and auditing [16, 34, 72]. Most notably, explanations can be used for informing and justifying downstream actions for decision-making. For instance, based on the contents of the explanation, an individual might decide to follow the model’s recommendation (or not); or, in a different application setting, an individual might decide to contest the result of a model’s decision (or not). The related literature described in this section ultimately works towards these broader goals of usefulness for explanations, and it is these broader goals that our concerns about performance disparity ultimately address.

Inherently Interpretable Models and Post hoc Explanations

Many approaches learn inherently interpretable models such as rule lists [75, 78], decision trees and decision lists [46], and others [17, 18, 42, 50]. However, complex models such as deep neural networks often achieve higher accuracy than simpler models [59]. Thus, there has been significant interest in constructing post hoc explanations to understand their behavior. To this end, several techniques have been proposed in recent literature to construct post hoc explanations of complex decision models. For instance, LIME, SHAP, Anchors, BayesLIME, and BayesSHAP [51, 59, 60, 65] are considered perturbation-based local explanation methods because they leverage perturbations of individual instances to construct interpretable local approximations (e.g., linear models). On the other hand, methods such as Gradient Input, SmoothGrad, Integrated Gradients, and GradCAM [63, 64, 68, 71] are referred to as gradient-based local explanation methods because they leverage gradients computed with respect to input features of individual instances to explain individual model predictions. An alternate class of methods referred to as global explanation methods attempt to summarize the behavior of black-box models as a whole rather
than in relation to individual data points [12, 44]. A more detailed treatment of this topic is provided in other comprehensive survey articles [9, 23, 33, 48, 53]. In contrast to the aforementioned works, which propose novel methods to explain models and their predictions, our work focuses on understanding group-based disparities in the quality of explanations generated by state-of-the-art post hoc local explanation methods.

Evaluating Explanations Prior research has proposed several metrics to determine if an explanation is reliable. An extensive survey of metrics for evaluating explanation methods can be found in Zhou et al. [80], who propose two high-level goals of explanation methods: interpretability (the clarity, simplicity, and breadth of the explanations), and fidelity (the completeness and soundness of explanations) [19, 32]. Liu et al. [49] provide a synthetic benchmark for explanation evaluation which includes implementations of several metrics, as well as a discussion of how to choose metrics for evaluation. Furthermore, several prior works proposed different metrics for evaluating various aspects of explanation quality, such as fidelity, stability, consistency, and sparsity [6, 31, 35, 44, 49, 56, 65, 80]. Recent research further leveraged the aforementioned properties and metrics to theoretically and empirically analyze the behavior of popular post hoc explanation methods [2, 3, 6, 20, 26, 31, 47, 67].

In addition to quantitative metrics, human-grounded evaluation approaches emphasize how human users perceive and utilize explanations [28]. For example, Lakkaraju and Bastani [43] carry out a user study to understand if misleading explanations can fool domain experts into deploying racially biased models, while Kaur et al. [38] find that explanations are often over-trusted and misused. Similarly, Pourshabzi-Sangdeh et al. [58] find that supposedly-interpretable models can lead to a decreased ability to detect and correct model mistakes, possibly due to information overload. Jesus et al. [37] introduce a method to compare explanation methods based on how subject matter experts perform on specific tasks with the help of explanations. Lage et al. [41] use insights from rigorous human-subject experiments to inform regularizers used in explanation algorithms. Though our work does not involve human subject study, we see this line of human-centered investigation as a critical complement to our quantitative evaluation framework.

Limitations and Vulnerabilities of Post hoc Explanations The aforementioned notions of explanation quality and the corresponding evaluation metrics were also leveraged to analyze the behavior of post hoc explanation methods and their vulnerabilities—e.g., Ghorbani et al. [31] and Slack et al. [66] demonstrated that methods such as LIME and SHAP may result in explanations that are not only inconsistent and unstable, but also prone to adversarial attacks. Furthermore Lakkaraju and Bastani [43] and Slack et al. [66] showed that explanations which do not accurately represent the importance of sensitive attributes (e.g., race, gender) could potentially mislead end users into believing that the underlying models are fair when they are not [4, 43, 66]. This, in turn, could lead to the deployment of unfair models in critical real-world applications. There is also some discussion about whether models which are not inherently interpretable ought to be used in high-stakes decisions at all. Rudin [61] argues that post hoc explanations tend to be unfaithful to the model to the extent that their usefulness is severely compromised. While this line of work demonstrate different ways in which explanations could potentially induce inaccuracies and biases in real-world applications, they do not focus on analyzing group-based disparities in explanation quality, which is our focus.

The Intersection of Fairness and Explainability While fairness and explainability are often described as qualities which “responsible” models ought to have simultaneously [62, 73], there is surprisingly little work exploring the intersection of the two. Begley et al. [13] extend SHAP as a tool to explain observed (un)fairness. Several recent works [1, 55, 69, 70] propose methods to make specific algorithms simultaneously fair and explainable. As fairness is often explicitly referred to as a justification or motivation for the use of explanation methods [14, 21, 72], it is critical to also evaluate explanations from the perspective of fairness. Here, we make a distinction between whether an explanation accurately portrays the fairness of the underlying model, and whether the explanation method performs equally well on all groups. Several works address the former question—e.g., Aïvodji et al. [4] introduce the notion of “fairwashing,” in which a rule-based explanation method can be adversarially constructed such that an unfair model decision can be rationalized. Meanwhile, Slack et al. [66] illustrate how a model can be adversarially constructed such that LIME and SHAP hide the explicitly unfair behavior of the model. Alikhahem et al. [5] finds that existing explainability tools are ill-suited for evaluating fairness performance. Finally, Dodge et al. [25] conduct a human study investigating how explanations can be used to interpret the fairness performance of a model. While the aforementioned works address the question of whether an explanation accurately portrays the fairness of the underlying model, we investigate whether the explanation method performs equally well on all the groups. In concurrent work, Balagopalan et al. [10] are motivated by a similar question. Notably, they focus solely on fidelity, and empirically validate our findings of disparity in existing explanation methods.

3 OUR FRAMEWORK

3.1 Preliminaries and Notation

Let us consider a dataset \( \mathcal{D} = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\} \) where \( x^{(i)} \in \mathbb{R}^d \) is a vector of feature values and represents an instance in the data, and \( y^{(i)} \in \mathcal{Y} \) is the corresponding class label. Since our goal is to unearth group-based disparities in the quality of post hoc explanations, we consider the case where the dataset \( \mathcal{D} \) comprises of some feature \( s \) which corresponds to a discrete sensitive attribute (e.g., gender, race, age).

Let \( f : \mathbb{R}^d \rightarrow \mathcal{Y} \) be a predictive model which can make predictions on instances in \( \mathcal{D} \). Let \( E : (x, f) \mapsto w \in \mathbb{R}^d \) be a local explanation method which takes as input an instance \( x \) and the predictive model \( f \), and outputs a vector of feature importances \( w \). Note that the \( j \)-th element of this vector (i.e., \( w_j \)) represents how important the \( j \)-th feature is to the prediction \( f(x) \). Finally, let \( M : (x, f, E) \mapsto \mathbb{R} \) denote a metric which measures some facet of the quality (e.g., fidelity or stability etc.) of a given local explanation \( E(x, f) \).
3.2 Metrics for Evaluating Explanations

In their survey, Zhou et al. [80] highlight two characteristics of a high-quality explanation: first, how well the explanation approximates the model, and second, human-understandability of the explanation. We focus on three metrics—fidelity, stability, and consistency—that measure the accuracy of the explanation’s approximation, as well as a fourth—sparsity—that measures how easily the explanation is understood by a user.

3.2.1 Fidelity. Perhaps the most natural question when considering explanation quality is the extent to which it accurately represents the underlying decision-making process. In other words, are the “important features” designated in the explanation truly important features for the model? Explanations that accurately designate the important features to the underlying model are said to have high fidelity.

In our framework, we select two measures of fidelity: ground truth fidelity, which is applicable to models that represent their features as linear weights comparable with explanations, and prediction gap fidelity, which is applicable to a wider range of models like neural networks.

The ground truth fidelity metric applies solely to machine learning models which inherently encode some notion of feature importance as a vector of $d$ weights, such as in the coefficients of a logistic regression model. While therefore limited in practical applicability across most models, this metric is a useful baseline for determining whether an explanation method is correct when there exists a ground truth definition of features and their importance. Specifically, suppose we have ground truth weights $\omega_i$ such as the weights of a logistic regression model. The ground truth fidelity metric is defined as the percent of the top $k$ features from explanation $w$ which are also in the top $k$ features from $\omega$.

**Definition 3.1.** Let $\omega \in \mathbb{R}^d$ be a vector denoting the importance of features in a model $f$, where $\omega_i > \omega_j$ implies that the $i$th feature is more important in $f$ than the $j$th feature. Let $\text{top}(k, \cdot)$ be a function that returns the indices of the $k$ largest elements of a given input vector. Then, given local explanation $w = E(x, f)$ the **ground truth fidelity** for a given value of $k$ is defined as follows:

$$M_{\text{ground truth}}(x, f, E) = \frac{|\text{top}(k, w) \cap \text{top}(k, \omega)|}{k}.$$ 

The above definition requires that some agreed-upon measure of feature importance $\omega$ is available. For models that do not have such a quantity, such as deep neural networks, we measure prediction gap fidelity. The intuition behind this fidelity metric is that the model’s prediction should not change substantially if minor alterations are made to features that are not important. If the prediction does change substantially, then this indicates that the explanation failed to capture features that were in fact important to the prediction, and is thus low fidelity. This follows a similar intuition to saliency in Dabkowski and Gal [24], and insertion/deletion in Petsiuk et al. [56]. This metric is applicable for probabilistic classifiers $f(x) = g(h(x))$, where $h : \mathbb{R}^d \to [0, 1]$ generates predicted probabilities and $g : [0,1] \to \{0, 1\}$ maps the probability to a classification. The $k$ most important features of the model for $x$ are identified from $w = E(x, f)$. Then, a small amount of Gaussian noise with variance $\sigma$ is added to all features of $x$ outside the top $k$ to produce $\tilde{x}$, and the difference between $h(x)$ and $h(\tilde{x})$ is taken. The “prediction gap” for the single $x$ is the expected value of this difference; practically, the procedure of adding noise and recording the gap in prediction—is repeated a total of $m$ times, and the “prediction gap” is computed as the average of all $m$ gaps.

**Definition 3.2.** Let $f$ be a model such that $f(x) = g(h(x))$, where $h : \mathcal{X} \to [0,1]$ generates predicted probabilities and $g : [0,1] \to \{0,1\}$ maps the probability to a classification. Let $\text{top}(k, \cdot)$ be a function that returns the indices of the $k$ largest elements of a given input vector. Then, given local explanation $w = E(x, f)$ and variance parameter $\sigma$, define $\tilde{x}$ such that

$$\tilde{x}_i = x_i + \sigma [i \notin \text{top}(k, w)] \cdot z_i \quad \text{where} \quad z_i \sim \mathcal{N}(0, \sigma).$$

Then **prediction gap fidelity** is defined as follows:

$$M_{\text{pred. gap}}(x, f, E) = \mathbb{E}[\|h(x) - h(\tilde{x})\|].$$

We approximate prediction gap fidelity empirically. We generate $\tilde{x}$ $m$ times, producing $\{\tilde{x}_j \mid 1 \leq j \leq m\}$. Then prediction gap fidelity is approximated as follows:

$$M_{\text{pred. gap}}(x, f, E) = \frac{1}{m} \sum_{j=1}^{m} ||h(x) - h(\tilde{x}_j)||.$$ 

3.2.2 Stability. The idea that similar points should receive similar explanations—and the observation that popular explanations often do not satisfy this property—has been discussed extensively (common terms for this principle, in addition to stability, include robustness and insensitivity) [7, 15, 77]. Alvarez-Melis and Jaakkola [7] note that instability has been observed even when the underlying model is stable—calling into question the veracity of those explanations. In other words, given a set of differing explanations for similar points, it is not credible that those explanations are all correct.

To measure an explanation’s stability at a point $x$, we add some noise to $x$ to generate similar points, calculate explanations for those similar points, then take the average L1 distance between $x$’s explanation and those of the similar points. This is similar to the definition of prediction gap fidelity, except now we add noise to all features and measure the difference in explanations rather than model predictions. This definition exactly mirrors the definition of stability in Yeh et al. [77] and others [6, 15].

**Definition 3.3.** Given a model $f$, local explanation $w = E(x, f)$, and variance parameter $\sigma$, define $\tilde{x}$ such that

$$\tilde{x}_i = x_i + z_i \quad \text{where} \quad z_i \sim \mathcal{N}(0, \sigma).$$

Then **instability** is defined as follows:

$$M_{\text{instability}}(x, f, E) = \mathbb{E}[\|E(x, f) - E(\tilde{x}, f)\|].$$

We approximate instability empirically. We generate $\tilde{x}$ $m$ times, producing $\{\tilde{x}_j \mid 1 \leq j \leq m\}$. Then instability is approximated as follows:

$$M_{\text{instability}}(x, f, E) = \frac{1}{m} \sum_{j=1}^{m} ||E(x, f) - E(\tilde{x}_j, f)||.$$ 


3.2.3 Consistency. Consistency captures the intuition that if an explanation for a single point is calculated multiple times, each of the calculated explanations should be similar. Inconsistent explanations for the same input \( x \) suggest that these explanations may be unreliable. This metric is motivated by concerns raised by empirical observations of the performance of many stochastic post-hoc explanation methods—requesting many explanations for the same point often resulted in drastically different explanations [6, 45, 45, 65, 79]. If the same point may result in a variety of very different explanations, it is unlikely that any individual explanation is correct, indicating a low approximation quality.

To measure consistency for a point \( x \), we calculate several explanations for that same point; then, we take the average L1 distance between the first explanation and each of the new explanations.

**Definition 3.4.** Let an explanation method \( E \) be stochastic, such that \( E_j \) indicates an explanation generated with a random seed \( j \). The empirical inconsistency is defined as follows:

\[
\mathcal{N}_{\text{inconsistency}}(x, f, E) = \frac{1}{m} \sum_{j=1}^{m} \| E_0(x, f) - E_j(x, f) \|_1.
\]

3.2.4 Sparsity. Finally, we are also interested in how easily an explanation can be understood and interpreted by users. Extensive previous work drawing from cognitive psychology has discussed the question of whether a given explanation is actually understood by a human user [39, 54, 74]. One common theme is complexity leading to higher cognitive load. Interpretability is measured in Bhatt et al. [15] as the entropy of the explanation, where equal attribution to all features is considered to be the least-interpretable. In Lakkaraju et al. [44], interpretability is measured by counting discrete concepts.

We measure sparsity by counting the number of features with an attributed importance greater than a threshold \( t \).

**Definition 3.5.** Let \( t \) be the threshold for which a feature importance is considered significant. Given local explanation \( w = E(x, f) \) and threshold parameter \( t \), complexity is defined as follows:

\[
\mathcal{N}_{\text{complexity}}(x, f, E) = \sum_{i} \| w_i > t \|.
\]

3.3 Measuring Disparity

For any given metric, we can empirically estimate disparity as follows. For each dataset, we partition the data into groups according to the sensitive attribute. In the datasets used for our experiments, we split the data into group 0 (majority group) and group 1 (minority group), though our framework is generic enough to handle multi-valued sensitive attributes. For each of these groups, we compute the mean values of the metric by averaging the metric values across instances in the respective groups. If there is a significant difference (as measured by a hypothesis test) between the mean metric values for each group, then there is a disparity in the quality of the explanations for that metric.

3.4 Consequences of Disparity

Though our broad proposal is to measure disparities in explanation quality in general, in this section we discuss some possible consequences of disparity in the specific metrics defined in the previous sections. We emphasize that disparity in additional metrics—or even different implementations of the same metrics—may have implications beyond what is discussed here; furthermore, this discussion is meant primarily as a preliminary exploration, and more tailored study (for example, with human subjects) for each case described here is warranted.

3.4.1 Fidelity Disparity. A significant disparity in explanation fidelity, or the closeness of an explanation to the model’s actual behavior, can have concrete ramifications. For example, as in Figure 1, a domain expert such as doctor could make systematically worse decisions for a group of people because of a disparity in explanation quality.

3.4.2 Stability Disparity. Like disparity in fidelity, disparity in stability can be consequential. Suppose the machine learning model that a doctor uses an explanation method that is not equally stable across groups. This disparity might lead to them rely, rightly or wrongly, more often on the model being explained for one group or the other. Measuring disparity in stability at development time, meanwhile, is important not just to prevent the above scenario, but to potentially identify problems with the model itself. Dooley et al. [27] observed that models can be unstable across groups; disparity in explanation stability may be a warning sign.

3.4.3 Consistency Disparity. Disparity in consistency is also problematic when considering the user’s interpretation of the explanations. A user may themselves be testing the explanation method by requesting an explanation multiple times, choosing to use the explanation for a given point only if they observe that the multiple explanations they received for that point are similar. If consistency disparity exists, therefore, the user may end up using explanations to aid decisionmaking much more often for one group than another. Returning to the example of the clinical setting, this may ultimately result in meaningfully disparate diagnoses.

3.4.4 Sparsity Disparity. Unlike the three metrics discussed previously, disparity in sparsity across demographic groups is not inherently problematic. In fact, it may reflect the model’s true behavior, e.g., if the model has a more complex decision boundary for one group than another, or if the model does rely more features for one group than another. Checking for disparity in sparsity, therefore, can be informative about the underlying model. One possible adverse consequence of disparity in sparsity is that there may be higher cognitive load when examining explanations for one group than another. However, if it is the case that the model truly is relying on more information to make predictions for one group, artificially ensuring equal sparsity may result in a substantial tradeoff with fidelity.

4 EMPIRICAL ANALYSIS

In this section, we demonstrate how to employ our explanation evaluation framework and interpret its outputs. First we generate a range of post-hoc explanations for various data and model combinations. We apply our evaluation framework to these explanations, computing the metrics of interest and analyzing them across data subgroups. Finally we examine the disparities that emerge, highlighting trends that may be of interest to practitioners.
4.1 Experimental Setup

4.1.1 Data. Our evaluation framework is of particular importance in settings where explanation disparities emerge across protected attribute classes like sex and race. Accordingly, we choose the following benchmark datasets: German Credit [29], Student Performance [22, 29], and COMPAS [8]. Though we acknowledge the limitations of COMPAS as a benchmark dataset [11] (and indeed, “fairness benchmarks” in general), we choose to evaluate and report explanation performance on it because we do not claim to successfully “achieve” any notion of fairness with respect to classification or explanation. We evaluate explanation disparities across one protected attribute in each dataset, namely sex in both German Credit and Student Performance and race in COMPAS. Datasets are divided into training and testing sets via an 80/20 random stratified split with respect to the target label.

4.1.2 Models. We investigate whether explanation disparities can vary based on the complexity of model being explained. We train both linear and non-linear models, Logistic Regression (LR) and a 3-layer neural network (NN) respectively. For the latter, we use 50, 100, and 200 nodes in each consecutive layer, ReLU activation, binary cross entropy loss, Adam optimizer, and 100 training epochs.

4.1.3 Explanation Methods. We investigate whether different explanation methods can produce different disparities. Most post-hoc explanation methods can be categorized into two families of approaches. One approach generates explanations by constructing locally interpretable model approximations whereas the other generates explanations by analyzing the behavior of the model’s gradient. We choose LIME [59], SHAP [51], and MAPLE [57] from the former and SmoothGrad [68], Integrated Gradients (IntGrad) [71], and Vanilla Gradients (VanGrad) [64] from the latter. We employ the authors’ implementations for LIME [59] and MAPLE [57]. We use Captum library [49] implementations of SHAP (sample size 1000), IntGrad, and VanillaGrad. We use default hyperparameter settings throughout unless otherwise specified. We implement SmoothGrad in PyTorch with standard normal noise and sample size 1000.

4.1.4 Metric Hyperparameters. For both ground truth fidelity and prediction gap fidelity we fix $k = 5$. Additionally, for prediction gap fidelity we fix $m = 1000$ and $\sigma = 0.1$. For both instability and inconsistency we fix $m = 5$. For sparsity we fix $t = 0.01$.

4.1.5 Setting and Implementation Details. For each dataset $D$, we begin by splitting the data into training set $D_{\text{train}}$ and testing set $D_{\text{test}}$ with respect to a random seed, as detailed in 4.1.1. Next we train predictive model $f$ on $D_{\text{train}}$ and partition $D_{\text{test}}$ with respect to the value of the salient protected attribute into $D_0$ and $D_1$, where $D_0$ and $D_1$ correspond to elements in $D_{\text{test}}$ with protected attribute values 0 and 1 respectively. For each explanation method $E$ and framework metric $M$, we generate $M_0 = \{ M(x, f, E) | x \in D_0 \}$ and $M_1 = \{ M(x, f, E) | x \in D_1 \}$. By comparing $M_0$ and $M_1$, we can determine whether there is a disparity in explanation performance. To understand whether $M_0$ and $M_1$ consistently differ, we perform 5 trials, repeating this procedure with different random seeds for the data split.

4.2 Results & Insights

We compute set averages $\overline{M}_0$ and $\overline{M}_1$ for each metric, model, explanation method, and dataset combination. To identify when explanation disparity is statistically significant, we perform Mann-Whitney U tests [52] on each pair of $\overline{M}_0$ and $\overline{M}_1$ distributions and report the resulting p-values in Table 1. Of the 162 experimental combinations analyzed, 18.5% exhibited statistically significant explanation disparity. In Table 2 we aggregate these instances across metrics, counting the number of times significant explanation disparity occurs in each of the 36 explanation, model, and dataset combinations. Across these 36 combinations, explanation disparity occurs in at least one metric 56% of the time. Next we analyze explanation disparity across each of our framework metrics and summarize additional aggregate trends.

4.2.1 Fidelity disparity. As detailed in subsections (a) and (b) of Table 1, significant prediction gap fidelity disparity occurred 30.6% of the time. Significant ground truth fidelity disparity also occurred 11.1% of the time. Across both ground truth fidelity disparity and prediction gap fidelity disparity, significant fidelity disparity occurred most often with SHAP explanations. For prediction gap fidelity disparity, which can be computed for both LR and NN models, first notice that the prediction gap values are larger across both groups 0 and 1 in the NN setting, indicating worse fidelity. Next notice that the majority of significant disparities occur in the more complex NN model setting (63.6%). In addition to being more frequently significant, prediction gap fidelity explanation disparity appears more severe in complex model settings. This is illustrated in Figure 2 by the wider gaps between $\overline{M}_0$ and $\overline{M}_1$ means in complex NN model settings when compared to their simpler LR counterparts. Moreover, even when disparities are not statistically significant, we notice more severe disparity in complex NN model settings when compared to their simpler LR counterparts. In these NN settings with observably larger disparity, the lack of statistical significance is possibly due to larger variances. These larger variances themselves suggest that prediction gap fidelity is a less robust metric on complex models. In general, the trends of fidelity disparity being more frequently significant and more severe in complex models align with our expectations of explanation methods that employ local linear approximations.

4.2.2 Stability disparity. As reported in Table 1 (d), significant stability disparity occurred 13.9% of the time and most frequently with VanillaGrad explanations. The majority (80%) of instances of significant explanation disparity occurred on the complex NN model setting. In Figure 3, we notice that instability in general, across both groups on the German Credit dataset. On the German Credit dataset, like with fidelity disparity, in Figure 3, we notice that even when disparities are not significant, wider gaps between $\overline{M}_0$ and $\overline{M}_1$ means in complex NN model settings when compared to their simpler LR counterparts, indicating more severe disparity. Similarly, larger variances are again observed on more complex models, suggesting stability is a less robust metric in these settings.

4.2.3 Consistency disparity. As reported in Table 1 (e), significant consistency disparity occurred in one setting, LIME on the LR model trained on the German Credit dataset. Yet as observed in Figure 4, though there exists a statistically significant consistency...
Table 1: P-values from Mann-Whitney U tests with null hypothesis that average metric values for group 0 ($m_0$) and group 1 ($m_1$) are equal. Of the 162 explanation/metric/model/dataset combinations analyzed, 18.5% exhibited statistically significant explanation disparity (bolded). Note that the Ground Truth (Fidelity) metric can only be computed for linear models (LR).

|         | LIME | SHAP | SmoothGrad | IntGrad | VanillaGrad | Maple |
|---------|------|------|------------|---------|-------------|-------|
| German Credit |     |      |            |         |             |       |
| LR      | 0.424 | 0.008 | 1.0        | 0.008   | 1.0         | 0.905 |
| NN      | 1.0   | 0.421 | 1.0        | 0.421   | 1.0         | 1.0   |
| COMPAS  | 0.841 | 0.151 | 1.0        | 0.131   | 1.0         | 0.401 |

(a) Ground truth — 2/18 significant

|         | LIME | SHAP | SmoothGrad | IntGrad | VanillaGrad | Maple |
|---------|------|------|------------|---------|-------------|-------|
| German Credit |     |      |            |         |             |       |
| LR      | 0.032 | 0.056 | 0.032      | 0.056   | 0.032       | 0.421 |
| NN      | 0.421 | 0.222 | 0.016      | 0.008   | 0.016       | 0.675 |
| COMPAS  | 0.095 | 0.016 | 0.008      | 0.016   | 0.008       | 0.222 |

(b) Prediction Gap — 11/36 significant

|         | LIME | SHAP | SmoothGrad | IntGrad | VanillaGrad | Maple |
|---------|------|------|------------|---------|-------------|-------|
| German Credit |     |      |            |         |             |       |
| LR      | 0.100 | 0.008 | 1.0        | 0.008   | 0.690       | 0.690 |
| NN      | 0.690 | 0.222 | 0.016      | 0.008   | 0.016       | 0.675 |
| COMPAS  | 0.310 | 0.151 | 0.022      | 0.310   | 0.310       | 0.348 |

(c) Sparsity — 11/36 significant

|         | LIME | SHAP | SmoothGrad | IntGrad | VanillaGrad | Maple |
|---------|------|------|------------|---------|-------------|-------|
| German Credit |     |      |            |         |             |       |
| LR      | 0.222 | 0.222 | 0.548      | 1.0     | 0.008       | 1.0   |
| NN      | 0.690 | 0.100 | 0.056      | 0.310   | 0.100       | 0.841 |
| COMPAS  | 0.421 | 0.222 | 0.222      | 0.008   | 0.008       | 0.690 |

(d) Stability — 5/36 significant

|         | LIME | SHAP | SmoothGrad | IntGrad | VanillaGrad | Maple |
|---------|------|------|------------|---------|-------------|-------|
| German Credit |     |      |            |         |             |       |
| LR      | 0.016 | 1.0   | 1.0        | 1.0     | 1.0         | 0.690 |
| NN      | 0.548 | 1.0   | 1.0        | 0.841   | 1.0         | 1.0   |
| COMPAS  | 0.310 | 0.672 | 1.0        | 1.0     | 1.0         | 0.841 |

(e) Consistency — 1/36 significant

Table 2: Here we aggregate the instances of significant explanation disparity reported in Table 1 across metrics, counting the number of times significant explanation disparity occurs in the explanation/model/dataset combinations. Across these 36 combinations, explanation disparity occurs in at least one metric 56% of the time.

|         | LIME | SHAP | SmoothGrad | IntGrad | VanillaGrad | Maple |
|---------|------|------|------------|---------|-------------|-------|
| German Credit |     |      |            |         |             |       |
| LR      | 2    | 2    | 1          | 2       | 2           | 0     |
| NN      | 0    | 0    | 1          | 1       | 1           | 0     |
| Student Performance |     |      |            |         |             |       |
| LR      | 0    | 0    | 0          | 1       | 0           | 0     |
| NN      | 0    | 2    | 0          | 2       | 0           | 1     |
| COMPAS  | 1    | 2    | 0          | 1       | 1           | 0     |
| NN      | 0    | 2    | 1          | 2       | 2           | 0     |

4.2.4 Sparsity disparity. As reported in Table 1 (c), sparsity disparity occurred 30.6% of the time and most frequently with IntGrad explanations. Notice that, in contrast with previously observed trends,
Figure 2: In each plot we compare the average prediction gap values for group 0 in blue ($m_0$) and group 1 in green ($m_1$) for each of the explanation methods ($\mathcal{E}$) listed on the x-axis. In the top row is German Credit (left), Student Performance (middle), COMPAS (right) with the LR predictive model. In the bottom row is German Credit (left), Student Performance (middle), COMPAS (right) with the NN predictive model. Differences between group 0 and group 1 distributions indicate explanation disparity.

Figure 3: In each plot we compare the average stability values for group 0 in blue ($m_0$) and group 1 in green ($m_1$) for each of the explanation methods ($\mathcal{E}$) listed on the x-axis. In the top row is German Credit (left), Student Performance (middle), COMPAS (right) with the LR predictive model. In the bottom row is German Credit (left), Student Performance (middle), COMPAS (right) with the NN predictive model. Differences between the group 0 and group 1 distributions indicate explanation disparity.

instances of significant sparsity disparity are split more evenly between the NN and LR model settings, 45.5% and 54.5% respectively. However, in Figure 5 we notice that the variance in average sparsity is typically larger in NN settings than in LR counterparts, similarly to those for fidelity disparity, perhaps contributing to the fewer instances of statistically significant disparity.
4.2.5 Additional aggregate trends. To identify which explanation methods exhibit the most explanation disparity across our experiments, we read Table 1 columnwise, noting that each explanation method is employed on 27 metric, model, and dataset combinations. IntGrad exhibited significant explanation disparity on 33.3% of the combinations, followed by SHAP (29.6%), VanillaGrad (22.2%), LIME...
(11.1%), SmoothGrad (11.1%), and MAPLE (3.7%). In Table 2 we also notice that IntGrad is the only explanation method with nonzero entries in every row. This means that IntGrad exhibits significant explanation disparity in at least one dimension in every model/dataset setting tested and is the only explanation method that does so. As detailed in Section 5, investigating the underlying mechanisms in explanation methods that give rise to these disparities is a critical direction for future research.

To determine whether explanations disparity occurs across multiple metrics at once, we refer to Table 2. Of the 20 explanation, model, and metric settings in Table 2 with non-zero entries, indicating significant explanation disparity is observed at least once, in one half disparities occur in only one metric, and in the other half disparities co-occur in two metrics. This behavior suggests that practitioners should not rely on disparity in one metric to serve as a proxy for disparity along all metrics.

5 DISCUSSION

In this work, we proposed an evaluation framework for unearthing disparities in post hoc explanation quality. To the best of our knowledge, this work is the first to study the problem of group-based disparities across a variety of metrics for explanation quality. This is especially important when the black-box model is making predictions about humans, given the consequential use of explanations in this context: explanations are often used in conjunction with a model in order to shape decisions, and disparity in explanation quality across demographic groups may be yet another way that adverse downstream outcomes become baked into the technical system. There are many cases where disparity in performance does not occur, however. The prevalence of significant disparity throughout datasets, metrics, and methods in our results suggest that measuring disparity in explanation quality is still critical: more than half of all dataset-model combinations exhibited disparity.

Our work has lasting implications for several stakeholders. For researchers working on novel explanation methods, it may be useful to apply our framework on a test suite of datasets and models to identify whether the explanation method is susceptible to disparity in general, and if so, with respect to what metrics. For practitioners who are developing application-specific models and explanation methods, our framework may be useful to identify whether the resulting explanations exhibit quality disparities in the context of the application.

If disparity is in fact identified, what actions should stakeholders take? As discussed in Section 3, disparity in fidelity and consistency always indicates that the explanation method is amplifying or generating problems. Disparity in stability and sparsity, on the other hand, may either indicate an issue with the explanation method, or an issue with the underlying model to be explained. Further research is necessary in order to provide more prescriptive next steps. Our work points to several follow-up research directions: First, what conditions give rise to observed disparity? Are there ways in which we can characterize datasets, models, or explanation methods such that we can more systematically identify the causes of measured explanation disparity? Second, human-grounded evaluations of our setting are necessary. Given significant explanation disparities, how is human decision making affected? Finally, it would be interesting to develop explanation methods which are less susceptible to such disparities while still maintaining overall utility.

ACKNOWLEDGMENTS

The authors would like to thank all the funding agencies supporting this work. This work is supported in part by the NSF awards IIS-2008461 and IIS-2040989, and research awards from Amazon, Harvard Data Science Institute, Bayer, and Google. HL would also like to thank Sujatha and Mohan Lakkharaju for their support and encouragement. The views expressed here are those of the authors and do not reflect the official policy or position of the funding agencies. JD completed this work while an undergraduate at Brown University.

REFERENCES

[1] Behnouf Abdollahi and Olla Nasrassl. 2018. Transparency in fair machine learning: the case of explainable recommender systems. In Human and machine learning. Springer, 21–35.
[2] Julius Adelbayo, Justin Gilmer, Michael Mueller, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. arXiv preprint arXiv:1810.03920 (2018).
[3] Sushant Agarwal, Shahin Jabbari, Chirag Agarwal, Sohini Udapidhyay, Zhiwei Steven Wu, and Himabindu Lakkharaju. 2021. Towards the Unification and Robustness of Perturbation and Gradient-Based Explanations. (2021).
[4] Utech Airodi, Hiromi Arai, Olivier Fortineau, Sebastien Gamba, Satsuki Hara, and Alain Tapp. 2019. Fairwashing: the risk of rationalization. In International Conference on Machine Learning, PMLR, 161–170.
[5] Kiana Alkhamdadi, Brianna Richardson, Emma Drobina, and Juan E Gilbert. 2021. Can Explainable AI Explain Unfairness? A Framework for Evaluating Explainable AI. arXiv preprint arXiv:2106.07483 (2021).
[6] David Alvarez-Melis and Tommi S Jakkola. 2018. On the robustness of interpretability methods. arXiv preprint arXiv:1806.08049 (2018).
[7] David Alvarez-Melis and Tommi S Jakkola. 2018. Towards robust interpretability with self-explaining neural networks. arXiv preprint arXiv:1806.07538 (2018).
[8] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. ProPublica (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[9] Alejandro Barredo Arrieta, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Ben-netot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58 (2020), 82–115.
[10] Aparna Balagopalan, Haoran Zhang, Kimia Hamidieh, Thomas Hartsiegn, Frank Rudzicz, and Maryelyn Ghassemi. 2021. The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations. ACM Conference on Fairness, Accountability, and Transparency (2022).
[11] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. 2021. It’s COMPSAicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks. arXiv preprint arXiv:2106.05498 (2021).
[12] Ostert Bastani, Carolyn Kim, and Hamsa Bastani. 2017. Interpretability via model extraction. arXiv preprint arXiv:1706.09773 (2017).
[13] Tom Begley, Tobias Schwedes, Christopher Frye, and Ilya Feige. 2020. Explainability for fair machine learning. arXiv preprint arXiv:2010.07389 (2020).
[14] Umang Bhattacharya, Adrian Weller, and Alice Xiand. 2020. Machine learning explainability for external stakeholders. arXiv preprint arXiv:2007.05408 (2020).
[15] Umang Bhatt, Adri van Weller, and José MF Moura. 2020. Evaluating and aggregating feature-based model explanations. arXiv preprint arXiv:2005.00631 (2020).
[16] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 648–657.
[17] Jacob Bien and Robert Tibshirani. 2009. Classification by set cover: The prototype vector machine. arXiv preprint arXiv:0908.2284 (2009).
[18] 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 ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 1721–1730.
[19] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. Electronics 8, 8.
[70] Alexander Stevens, Peter Deruyck, Ziboud Van Veldhoven, and Jan Vantienen. 2020. Explainability and Fairness in Machine Learning: Improve Fair End-to-end Lending for Kiva. In 2020 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 1241–1248.

[71] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In International Conference on Machine Learning. PMLR, 3319–3328.

[72] Harini Suresh, Steven R Gomez, Kevin K Nam, and Arvind Satyanarayan. 2021. Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 1–16.

[73] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad Van Moorsel. 2020. The relationship between trust in AI and trustworthy machine learning technologies. In Proceedings of the 2020 conference on human factors in computing systems. 272–283.

[74] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In Proceedings of the 2019 CHI conference on human factors in computing systems. 1–15.

[75] Fulton Wang and Cynthia Rudin. 2015. Falling rule lists. In Artificial Intelligence and Statistics. PMLR, 1013–1022.

[76] Leanne S Whittmore, Anthe George, and Corey M Hudson. 2016. Mapping chemical performance on molecular structures using locally interpretable explanations. CoRR, abs/1611.07443 (2016).

[77] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. 2019. On the (in) fidelity and sensitivity of explanations. Advances in Neural Information Processing Systems 32 (2019), 10967–10978.

[78] 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.

[79] Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell. 2019. "Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations. arXiv preprint arXiv:1904.12992 (2019).

[80] 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.