An Empirical Study of the Trade-Offs Between Interpretability and Fairness

Shahin Jabbari [1] **Han-Ching Ou** [1] **Himabindu Lakkaraju** [1] **Milind Tambe** [1]

Abstract

As machine learning models are increasingly being deployed in critical domains such as criminal justice and healthcare, there has been a growing interest in developing algorithms that are interpretable and fair. While there has been a lot of research on each of these topics in isolation, there has been little work on their intersection. In this paper, we present an empirical study for understanding the relationship between model interpretability and fairness. To this end, we propose a novel evaluation framework and outline appropriate evaluation metrics to determine this relationship across various classes of models in both synthetic and real world datasets.

1. Introduction

Over the past decade, there has been an increasing interest in leveraging machine learning (ML) models to aid decision making in critical domains such as healthcare and criminal justice. However, the successful adoption of these models in the real world relies heavily on how well decision makers are able to understand and trust their functionality (Doshi-Velez and Kim, 2017; Lipton, 2016). Consequently, there has been a growing emphasis on building ML models and algorithms that are not only accurate, but also fair and interpretable. This, in turn, has resulted in the emergence of two exciting areas of research within the ML community, *model interpretability* (Lakkaraju et al., 2016; Ribeiro et al., 2016) and *algorithmic fairness* (Berk et al., 2018; Dwork et al., 2012; Hardt et al., 2016). While these two areas share several similarities with respect to their end goals, they have often been treated as separate threads by prior literature.

Model interpretability is advocated as a mean to debug existing ML models, and detect potential biases. On the other hand, the literature on algorithmic fairness argues that solely optimizing for predictive accuracy is one of the predominant causes for the discriminatory behavior of algorithms (Berk et al., 2018). In spite of the extensive research on model interpretability and algorithmic fairness, there has been little work connecting these two directions. While there is a reasonable understanding of the trade-offs between accuracy and interpretability (Ribeiro et al., 2016), and accuracy and fairness (Berk et al., 2017; Feldman et al., 2015; Fish et al., 2016), there has been little work on exploring the trade-offs between fairness and interpretability. A notable exception is the work of Kleinberg and Mullainathan (2019) which we discuss in detail in Section 2.

Our Work To study the trade-offs between model interpretability and fairness, we propose a novel evaluation framework and outline appropriate evaluation metrics to effectively tease apart various confounding effects and determine the relationship between interpretability and fairness across various classes of models (e.g., linear vs. tree based models). More specifically, we carefully construct a variety of synthetic datasets to study the effect of various factors that impact the interpretability-fairness trade-offs, e.g., correlations between the protected, non-protected attributes and class labels, and group imbalance with respect to the protected attributes.

We show that the relationship between interpretability and fairness is complex. More specifically, we found that the trade-offs between fairness and interpretability follow four different trends depending on the correlations between protected, non-protected attributes and class labels. Our analysis reveals that the interpretability-fairness trade-offs do not depend on group imbalance. We further validate our insights on real world datasets. We view our work as a first step in understanding the trade-offs between fairness and interpretability. Our work leaves open several exciting directions for future work which we discuss in Section 4.

2. Related Work

Below, we briefly discuss related work pertaining to interpretability, fairness and their intersection.

Interpretability Many approaches have been proposed to directly learn interpretable models for classification and clustering (Kim et al., 2014; Lakkaraju and Leskovec, 2016; Lakkaraju et al., 2016; Letham et al., 2015). To this end, various classes of models such as decision trees, decision

[1] Harvard University. Correspondence to: Shahin Jabbari <jabbari@seas.harvard.edu>.

ICML Workshop on Human Interpretability in Machine Learning, 2020. Copyright 2020 by the author(s).

lists (Letham et al., 2015), decision sets (Lakkaraju et al., 2016), prototype based models (Bien and Tibshirani, 2009), and generalized additive models (Caruana et al., 2015; Lou et al., 2012) are proposed. However, complex models such as deep neural networks and random forests are often shown to achieve higher accuracy than simpler interpretable models (Ribeiro et al., 2016); thus, there has been an interest in constructing post hoc explanations to understand their behavior (Lundberg and Lee, 2017; Ribeiro et al., 2016; Selvaraju et al., 2017).

Fairness The initial literature on algorithmic fairness emphasize heavily on outlining the precise definitions of fairness (Hardt et al., 2016). Several competing and contrasting notions of fairness emerge during this phase which can be broadly categorized into: 1) *group fairness* which emphasizes that protected groups should receive similar treatment as that of advantaged groups (Hardt et al., 2016) 2) *individual fairness* which requires that *similar* individuals to be treated similarly (Dwork et al., 2012), and 3) counterfactual fairness which captures the intuition that a decision pertaining to an individual is fair if it is the same in the actual world and a counterfactual world where the individual belonged to a different demographic group (Kusner et al., 2017). We refer the reader to a recent survey (Berk et al., 2018).

Connections between Fairness and Interpretability Most relevant to this paper is the work by Kleinberg and Mullainathan (2019). They consider a selection problem in which a decision maker aims to select a fixed fraction of applicants e.g., to give them loans. This is quite different than a classification setting where there is no restriction on the fraction of instances that are positively labeled by a classifier. Kleinberg and Mullainathan (2019) assume each applicant has a true score (e.g., a true credit-worthiness score in the context of loan applications) and that this true score is independent of the sensitive attribute. This assumption also does not typically hold in a classification setting. Furthermore, they propose a notion of advantage which states that the ratio of the advantaged applicants to disadvantaged applicants is increasing as we focus on applicants with higher and higher scores. Under these assumptions they demonstrate theoretically that simpler (and hence more interpretable) models are strictly improvable i.e., there exists a more complex model that is both strictly more accurate and also strictly improves the fairness (or *equity*) of the simpler model in terms of admitting a higher number of disadvantaged applicants. Although our setting is not directly comparable to them, we show that simpler models can indeed be fairer than more complex models in certain situations. Kleinberg and Mullainathan (2019) also show that simpler models create incentives to use information about individuals' membership in the disadvantaged group.

3. Results

Preliminaries We focus on binary classification tasks and assume that the data points are divided into two groups using a binary protected attribute. In our analysis, we focus on two commonly used definitions of fairness: statistical parity (Pedreshi et al., 2008) and equality of opportunity (Hardt et al., 2016). We leave the study of other fairness measures for future work. For interpretability, instead of choosing a context-specific metric, we use the number of features available to the classifier as our measure of model complexity. Fewer the number of features available to the classifier, lower the complexity of the classifier and higher the interpretability. This measure is also generic enough to cater to the diverse model classes we study in this work. *We note that using the number of features as a measure of interpretability is rather too strong and simplistic, but we believe that this assumption is a good starting point in exploring the trade-offs between fairness and model interpretability especially when focusing on simple classification techniques such as linear models or decision trees.*

Generating Synthetic Data We generate synthetic datasets with two classes (lables) and each class is allocated one normally-distributed cluster of data points. This construction allows us the flexibility to vary the center and standard deviation σ^2 of each cluster and thereby explore the effect of the separability of the classes. We then add a new binary-valued feature which corresponds to the protected attribute using the following procedure. We assign each data point with a positive class label to the advantaged (disadvantaged) group with probability p $(1-p)$. Similarly, for each negatively labeled data point, we assign the point to belong to the advantaged (disadvantaged) group with probability $1-p$ (p). This construction coupled with the fact that we have equal number of positively and negatively labeled data points guarantees that the number of the data points that belong to the advantaged group is equal to the number of the data points that belongs to the disadvantaged group. Furthermore, focusing on values of $p > 1/2$ ensures that the advantaged group has a higher fraction of positively labeled data points compared to the disadvantaged group.

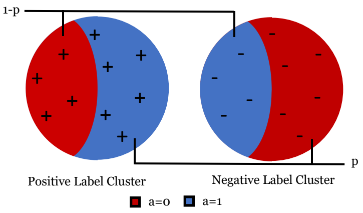

Figure 1. An illustration of how we assign the protected attribute to the clusters when $\sigma^2 \approx 0$.

The conditional group membership probability p controls the correlation between the protected attribute and the class label as p determines how predictive the protected attribute is of the class label. When $p \to 1/2$, the protected attribute reveals less information about the class label and hence the predictive power of the protected attribute is lower. On the other hand, as $p \to 1$, the protected attribute fully determines the class label. See Figure 1 for an illustration.

The σ^2 represents the spread of the data points around the cluster centers and can model the correlation between the non-protected attributes and the class label. In the extreme case where $\sigma^2 \to 0$, the data points pertaining to each cluster are very tightly knit and the non-protected attributes completely determine the class labels of the data points (see Figure 1). On the other hand, when $\sigma^2 \to \infty$, the data points pertaining to each cluster are spread out almost randomly across the feature space and the non-protected attributes do not convey any information about the class labels. By varying σ^2, we can interpolate between these two extremes.

The group membership ratio r denotes the fraction of data points belonging to the advantaged vs. disadvantaged groups, and thereby models group imbalance. Group imbalance is often touted to be one of the primary causes of bias in algorithmic decision making in practice (Kleinberg et al., 2017). We only focus on values of $r \geq 1$ and the higher the value of r, the more imbalanced the resulting dataset. This choice guarantees that the advantaged group also has a higher number of data points than the disadvantaged group. Therefore, we use the terms advantaged (disadvantaged) with majority (minority) interchangeably. To simulate group imbalances, we randomly choose data points to delete from the minority group until the desired ratio r is achieved. While doing so, we ensure that the correlation between the protected attribute and the class label remains unaffected by only deleting those randomly chosen data points that do not change this correlation significantly.

Classifiers An important factor that affects the trade-offs between interpretability and fairness is the classifier C. We experiment with several classifiers such as logistic regression, naive Bayes, decision trees, SVMs, neural networks, and random forests. These classifiers span a variety of model classes such as linear vs. non-linear models, rule-based vs. feature-importance based models, ensemble vs. stand alone models. We use scikit-learn implementations with default parameter settings to construct these classifiers.

Plotting Trade-Off Curves The fundamental building block of our analysis is a *trade-off curve* which is a plot depicting how model fairness is affected as we vary model complexity (or interpretability) under different conditions. We use the number of features available to a classifier as a measure of model complexity in this work. To track fairness, we keep track of the *fairness violation* δ i.e., the degree in which the fairness definitions are violated – the higher the δ is the more the fairness is violated by the classifier. Given these metrics, the trade-off curves can be plotted by having the number of features available to the classifier (complexity) on the x-axis and fairness violation δ on the y-axis. Each plot has a curves for statistical parity and another for equality of opportunity. We can make one such plot for each combination of C, p, σ^2, and r. For a given fixed set of values for the four parameters, we first generate a synthetic dataset. Then, we allow the classifier C to use at most one feature (complexity = 1). To this end, we perform a 5-fold cross validation and choose the feature that maximizes accuracy. We then train the classifier using this one feature and compute δ. We repeat the above step K times. At each step $1 \leq k \leq K$, we choose k best features that maximize the accuracy using 5-fold cross validation, train the classifier C using those features only, and compute the corresponding fairness violations so that we can track these quantities to understand their trends. We also track other metrics such as accuracy and F-1 score of the classifier. All results are out-of-sample and averaged using 5-fold cross validation.

We study the effects of varying p, σ^2 and r on the trade-off curves. For space considerations we only consider the effect of varying r in this manuscript. We wrap up by extending our observations to the real world datasets.

3.1. Correlation between Protected Attribute & Class Label

To model the correlation between the protected attribute and class label, we vary p while fixing other parameters to track the trends in accuracy, F-1 score, and fairness violations as a function of the number of allowed features. Figure 2 shows the trade-off curves for $\sigma^2 = 10$, $r = 10$, $C =$ logistic regression and different values of p.

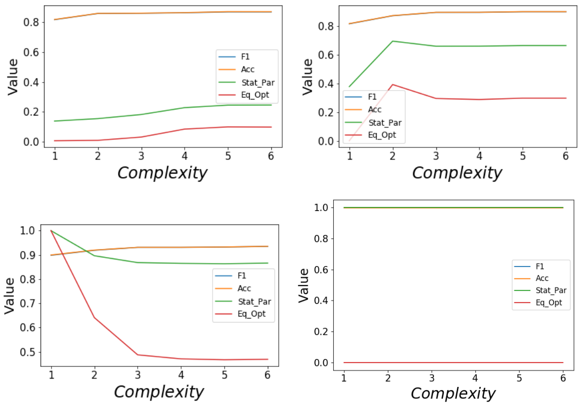

Figure 2. The effect of increasing the predictive power of the protected attribute p. $C =$ logistic regression. $\sigma^2 = 10$ and $r = 2$. $p = 0.6$ (upper left), $p = 0.8$ (upper right), $p = 0.9$ (lower left) and $p = 0.999$ (lower right).

As expected, F-1 scores and accuracies of classifiers increase monotonically as their complexity increases. On the other hand, we found that diverse trends emerge with respect to fairness violations. In particular, we see 4 different trends in Figure 2: (1) fairness violation increases as model complexity increases (upper left), (2) fairness violation increases initially and then decreases (upper right), (3) fairness violation decreases (lower left), and (4) fairness violation remains constant (lower right).

A simple calculation shows the statistical parity violation is $p-(1-p) = 2p-1$ which is increasing in p. If the clusters are perfectly separable i.e., $\sigma^2 = 0$, we expect the statistical parity violation to reach $2p-1$ when all the features are accessible to the classifier. Figure 2 shows that the statistical parity violation gets close to $2p-1$ for different values of p but does not always reach that exact value (as σ^2 is fairly small but still non-zero). While we cannot as easily quantify the value that the equality of opportunity violation should converge to, it is evident from the Figure 2 that this violation also exhibits a similar trend and increases as p increases.

In trend (1) (upper left), both fairness violations increase as we allow the classifier access to more features. We note that the protected attribute only becomes part of the feature set when we allow the model complexity to be very high indicating that it is not strongly predictive of the class label. This, in turn, implies that adding the protected attribute to the feature set of the classifier will not drastically change the fairness violations.

In trend (2) (upper right), the fairness violations increase initially but then decrease as complexity becomes larger. We note that the protected attribute is chosen as the second feature (sharp increase in the fairness violations). In contrast to trend (1), adding the protected attribute in this case results in a significant increase both in accuracy and fairness violations. After the addition of the protected attribute, adding more features does not seem to affect things by much.

In trend (3) (lower left), the fairness violations decrease as the model complexity increases. As p gets closer to 1, the protected attribute becomes the strongest predictor of the class label and hence would be selected first. This results in statistical parity and equality of opportunity violations of 1 initially. This is because, when using the protected attribute as the only feature, all the data points in the minority group are labeled positively by the classifier while all the data points in the majority group are labeled negatively. Furthermore the false negative rate in the minority group is 0 while the false negative rate in the majority group is 1. It can also be seen that adding other non-protected features after this point decreases the fairness violations.

In trend (4) (lower right), the fairness violations do not change as we increase the complexity. At $p = 0.999$, the protected attribute (almost) perfectly predicts the class label. This also implies that adding any other non-protected features has no impact on fairness violations.

We observe a phase transition from trend (1) to (4) as we increase p. We also experimented with other values of σ^2 and r and observe a similar transition as p increases. While the experimental results reported here were with logistic regression, there is robustness in our results when using other classifiers such as naive Bayes or decision trees.

3.2. Varying Several Parameters Simultaneously

While we have omitted the experimental results regarding varying σ^2 and r, we describe the high level insights. We observe a similar phase transition from trend (1) to trend (4) as we increase σ^2. Our observations also exhibit robustness with respect to the choice of classification algorithm C. The trade-off curves seem to be independent of the imbalance parameter r. While this is surprising at first glance, it can be explained by the observation that the distribution of non-protected attributes is the same for both groups. For future work, we plan to modify our data generation process to study the role of r in a more meaningful way.

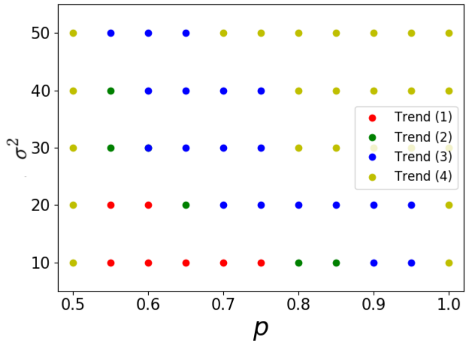

Figure 3. An illustration of the transition between trends as a function of p and σ^2. The classification algorithm that is used is logistic regression and $r = 2$.

Since our trade-off curves are independent of the choice of imbalance parameter r and classification algorithm$C$, to study all parameters simultaneously, we only consider varying p and σ^2. Figure 3 shows the results for logistic regression and $r = 2$. The x-axis captures various values of p, and the y-axis captures different values of σ^2. For each pair of p and σ^2 values, we plot the trade-off curves for fairness violations and label them with one the 4 trends we discussed in Section 3.1. For any fixed σ^2 (p), we see a smooth transition between trends as we increase p (σ^2). While the exact threshold where the transition happens depends on the specific values of p and σ^2, for moderately low values of p and σ^2 (bottom left) the trade-off curves are in trend (1). We will discuss this in more details next.

3.3. Validating Our Insights with Real World Datasets

We start our real world experiments by introducing the datasets. Our first dataset is COMPAS which is collected by ProPublica (Kaggle, b) and captures detailed information about the criminal history, jail and prison time, demographic attributes, and COMPAS risk scores for 7214 defendants. The protected attribute in this dataset is race with African-Americans being the minority group and whites being the majority group. Each defendant in the data is labeled either as high-risk or low-risk for recidivism.

Our second dataset is the Adult dataset (Kaggle, a) with information about the income level, demographic and socio-economic attributes of $48,842$ individuals. The protected attribute in this case is gender with females being the minority group and males being the majority group. Each individual in this dataset is labeled with "$>$ 50K" or "$\leq$ 50K" depending on the individual's income level.

Similar to our synthetic data experiments, we vary the number of features and also apply a 5-fold cross validation in the real world experiments. Since we consider all possible subsets of features of size k, when choosing k best features, our computations become intractable quickly in case of real world datasets. To remedy this, we use the built-in function `SelectKBest` from scikit-learn library to select the K most informative features for each of the datasets from which the k best features will in turn be chosen using 5-fold cross validation. We set $K = 7$ in our experiments.

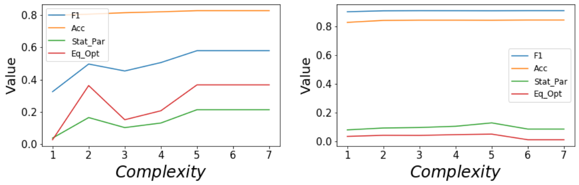

Figure 4. The trade-off curves in real world data using logistic regression as a classifier. Adult dataset (left) and COMPAS (right).

Figure 4 shows the trade-off curves for the Adult dataset (left) and the COMPAS dataset (right). These curves follow trends (1) and (4), respectively. To better understand these observations, we put them in the context of our synthetic experiments . More specifically, we estimate p and σ^2 on the real world datasets and determine if the trade-off curves on these real world datasets match those of the synthetic datasets with similar properties.

In Table 1, `Corr-Protected` corresponds to the correlation between the protected attribute and the class label and is computed as the Pearson product-moment correlation coefficient. p denotes the conditional group membership probability and corresponds to the value of `Corr-Protected` for each dataset. This is obtained by first generating synthetic datasets with various values of p and then computing the Pearson product-moment correlation coefficient for each of those datasets, and picking the p whose corresponding correlation coefficient on the synthetic dataset matches with that of the real world dataset's `Corr-Protected` value. `Pred-Non-Protected` denotes how well the non-protected attributes can predict the class label. This is analogous to the σ^2 parameter in the case of the synthetic data experiments. To compute the value of `Pred-Non-Protected` for a particular dataset, we train Gradient Boosted Trees, a highly expressive classifier, using only the non-protected attributes and the resulting accuracy is the `Pred-Non-Protected` value of that dataset.

Let us consider the Adult dataset for which trade-off curves are shown in Figure 4 and key characteristics are summarized in Table 1. As shown in Figure 3, multiple trends can emerge when $p = 0.6$ depending on σ^2. Similar to the computation of p above, we can leverage synthetic datasets to check what value of σ^2 will result in the same predictive power as `Pred-Non-Protected`. It turns out that the corresponding σ^2 in case of the Adult dataset is 10. We can see from Figure 3 that when $p = 0.6$ and $\sigma^2 = 10$, the corresponding trade-off curves follow trend (1). It is easy to see that trade-off curves for the Adult dataset also follow the same trend (Figure 4 (left)). Performing a similar analysis as above, we find out that $p = 0.55$ and $\sigma^2 = 12$ for COMPAS dataset. We can see from Figure 3 that when $p = 0.55$ and $\sigma^2 = 12$, the neighboring trade-off curves follow trend (1) or trend (4). Figure 4 (right) shows that the trade-off curve for COMPAS follows trend (4), there by, validating our understanding of the fairness violations.

Dataset	Corr-Protected	p	Pred-Non-Protected
COMPAS	-0.12	0.55	0.84
Adult	-0.21	0.6	0.86

Table 1. Information about the datasets. See text for more details.

4. Future Work

There are many possible areas for future work in addition to what we point out earlier. First, it would be valuable to develop a theory to accompany our experimental results. This theory can be used to explain under what circumstances we observe each of the trade-off trends. Second, we should extend our experimental results to more complex real-world domains. The state of the art classification algorithms for COMPAS and Adult datasets use a small set of features so the set of possible trade-offs are limited to begin with. Lastly, it would be insightful to study more complicated correlations in the data e.g., between protected and non-protected attributes. This is useful in practice where the protected attributes are prohibited to be used by law.

References

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. In *4th Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2017.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 2018.

Jacob Bien and Robert Tibshirani. Classification by set cover: The prototype vector machine. *arXiv:0908.2284*, 2009.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Knowledge Discovery and Data Mining*, pages 1721–1730, 2015.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv:1702.08608*, 2017.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science*, pages 214–226, 2012.

Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2015.

Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 144–152, 2016.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

Kaggle. Adult income dataset. https://www.kaggle.com/wenruliu/adult-income-dataset, a.

Kaggle. Compas recidivism racial bias. https://www.kaggle.com/danofer/compass#propublica_data_for_fairml.csv, b.

Been Kim, Cynthia Rudin, and Julie Shah. The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Advances in Neural Information Processing Systems*, pages 1952–1960. 2014.

Jon Kleinberg and Sendhil Mullainathan. Simplicity creates inequity: Implications for fairness, stereotypes, and interpretability. In *Proceedings of the 20th ACM Conference on Economics and Computation*, pages 807–808, 2019.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Conference on Innovations in Theoretical Computer Science*, pages 43:1–43:23, 2017.

Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

Himabindu Lakkaraju and Jure Leskovec. Confusions over time: An interpretable bayesian model to characterize trends in decision making. In *Advances in Neural Information Processing Systems*, pages 3261–3269, 2016.

Himabindu Lakkaraju, Stephen Bach, and Jure Leskovec. 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*, pages 1675–1684, 2016.

Benjamin Letham, Cynthia Rudin, Tyler McCormick, and David Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics*, 2015.

Zachary Lipton. The mythos of model interpretability. *arXiv:1606.03490*, 2016.

Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–158, 2012.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774. 2017.

Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568, 2008.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD*, pages 1135–1144, 2016.

Ramprasaath Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.