Honors Capstone Report: Robustness of Fairness in Machine Learning

Serafina Kamp
serafibk@umich.edu
University of Michigan

ABSTRACT

As machine learning algorithms become widely used in society, certain subgroups are more at risk of being harmed by unfair treatment. Fairness metrics have been proposed to quantify this harm by measuring certain statistics with respect to an evaluation dataset. In this work, we seek to analyze how *robust* these metrics are. That is, we are interested in whether these metrics give the same "fairness score" when measured on different sets of samples from the same distribution. This is important because it gives us insight into how much we can trust the conclusions given by a fairness metric prior to deployment of a model. We design a framework to conduct experiments to test the robustness of a popular fairness metric. We find that, when compared to more traditional performance metrics, it is more sensitive to fluctuations in the evaluation dataset in a variety of settings. Additionally, our work provides a foundation for studying the robustness of fairness metrics in general.

KEYWORDS

Fairness, Machine Learning, Robustness, Bootstrap Sampling

1 PROBLEM AND MOTIVATION

Machine learning algorithms are increasingly used to automate decision processes in society. Sometimes, these decisions are to the detriment of certain demographic groups due to historical biases in human decision making. After all, these algorithms learn from data which is influenced by the human decision making process. So, any existing bias in a dataset may be amplified by a model that learns from that dataset. For example, it has been shown that STEM job advertisements typically show up more for men than for women [23]. This imbalance means less women are being given information about available STEM jobs which could cause disparities in gender representation in those types of jobs down the line.

The study of fairness in machine learning seeks to measure and mitigate biases in algorithms before deployment in society. However, there are currently many different fairness metrics proposed [14, 17] and it has been shown to be impossible to satisfy all of them at once [11, 22]. This can lead to contradictory claims of fairness. Consider a popular dataset in fairness analysis: The COMPAS dataset [2]. This dataset includes information about parolees including demographic and criminal history features and it has been used to train a recidivism risk assessment tool to drive the decision of allowing parole. The tool has been shown to be both fair and unfair based on different metrics which leads to discussions of which metric is actually appropriate in that context [2, 11, 12].

Taking a step back, there is perhaps a more fundamental question we should be asking about these proposed fairness metrics before discerning which one to use in different scenarios. Consider the scenario represented in Figure 1. Suppose we train a model on a dataset and evaluate our model on a held-out test set as usual and find that it satisfies the appropriate fairness metric. Now suppose we create four more held-out test sets by sampling

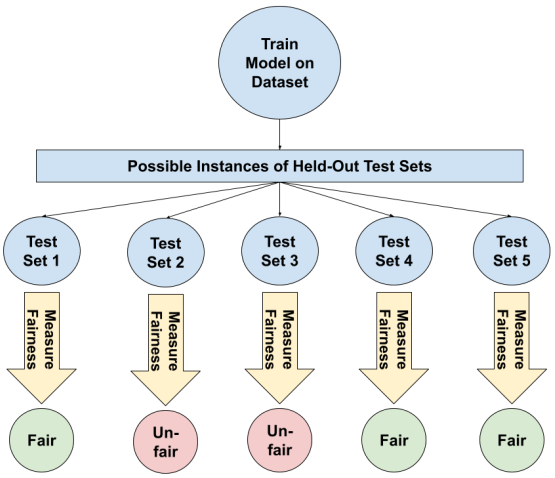


Figure 1: An example of a situation where we would consider a fairness metric to lack robustness.

points from the same distribution as our original test set and get various conclusions of fairness with respsect to the same fairness metric. A key thing to note here is that the underlying distribution of data points has not changed – we have simply sampled a few more sets of points from this distribution. We would say that this metric lacks robustness and it is exactly this that we aim to formalize and study in this work. For simplicity, we focus on one popular fairness metric, **equal opportunity** [17], in this analysis.

Specifically, in this work we aim to answer the following questions [37]:

RQ1. How can we evaluate the *robustness* of fairness metrics?
RQ2. Is *equal opportunity* a robust fairness metric?
RQ3. How much do different choices of models and features affect the robustness of the fairness metric?
RQ4. Do we see the same trends across different domains?

1.1 Related Work

There has been significant work in quantifying fairness and designing techniques for achieving it [21, 24, 28, 32] as well as in understanding the implications of using fair predictors in practice [34]. The prevalence of bias in fields as wide-ranging as Natural Language Processing [7, 31], vision [8], and health [1] have led to domain-specific analyses on bias detection and consequent work on both building and evaluating fairer datasets [4, 39]. Further, a survey of industry practitioners highlights the need to understand the practical implications of using fairness metrics [19].

There is no single agreed-upon measure of fairness since different contexts may require different criteria of measurement, including exogenous concerns like privacy-preservation [5, 6, 38]. However, while there is no consensus measure of fairness, some tests for evaluating group fairness that have gained widespread acceptance include *demographic parity* [9], *equalized odds* and *equal opportunity* [17]. In the present work, we we focus primarily on the *equal opportunity* fairness metric since there has been significant exploration of models that enforce this constraint [17, 24]. We also use *equalized odds* to derive a fair predictor.

Recent work has analyzed the effects of statistical and adversarial changes in the data distribution. Some of this work has focused on deriving fair models when there is a distributional shift in the data [33], when strategically acting adversaries inject errors in the data [10] or when the data is perturbed to negatively impact a particular subgroup [3, 27].

2 BACKGROUND

2.1 Preliminaries

To learn a predictive model, we use logistic regression both with and without an ℓ_2-norm regularizer [18]. This involves solving the following optimization problem:

$$\min_{\theta,b} \quad C\sum_{i=1}^{n}\log(\exp(-y_i(x_i^T\theta+b))+1)+\frac{1}{2}||\theta||_2^2$$

where (x_i, y_i) are labeled training datapoints, θ, b are the learned parameters, and C is a hyperparameter that controls the degree of regularization.

Each datapoint has a corresponding binary label $\in \{0, 1\}$. For instance, in the COMPAS dataset (see Section 2.2) each datapoint corresponds to an individual and a label of 1 indicates an individual who re-offends within two years. The features that distinguish historically disadvantaged groups are called *sensitive attributes* and the groups themselves are called *protected groups* [17]. Each datapoint includes a sensitive attribute $z \in \{0, 1\}$ that indicates their membership in a protected group. We train the base classifier both including and excluding these sensitive attributes.

We use group fairness measures to evaluate the fairness of the predictor returned by the algorithm. In this work, we focus primarily on analyzing the equal opportunity fairness metric [17], which enforces equal true positive rates (TPR) across each sensitive attribute group. This metric is a weaker notion than the equalized odds fairness metric [17], which enforces equal TPR *and* equal false positive rates (FPR) across each sensitive attribute group. We also experiment with post-processing the predictor by solving a constrained optimization program with the constraints specifying the fairness conditions [17, 29, 30]. A formal definition of the fairness metrics used is given in Section 2.3.

2.2 Datasets

We use the COMPAS dataset [2], the Bank Marketing dataset [26], and the South German Credit (SGC) dataset [13] for our analyses. These datasets are well-known benchmarks that have been frequently used to study algorithmic fairness [24]. Further, the difference in domain and protected attributes between the datasets allows us to analyze the robustness of fairness metrics beyond a single domain.

The COMPAS dataset contains 6150 datapoints with 8 features. The features include demographic information such as age, race, and sex as well as criminal history information such as priors, juvenile offences, and degree of current crime. When assuming a binary sensitive attribute, the dataset is restricted to Caucasian American and African American defendants; given the bias inherent in the dataset, African American defendants are considered to be the protected group. The binary-valued label indicates whether or not the individual has reoffended within two years after being released from prison.

The Bank Marketing dataset [26] contains 45211 datapoints with 15 features. The features include demographic information such as age, job, and education, seasonal data such as day and month, and financial data such as balance and whether an individual has any personal loans. Following prior work [40], the sensitive attribute is age where ages between 25 and 60 are considered protected. A positive outcome is when an individual subscribes to a term deposit.

The SGC dataset [13] contains 1000 datapoints with 20 features. The features of this dataset include demographic information such as age, sex, and marriage status, financial standing information such as credit history, savings account amount, and homeowner status, and, finally, information about the requested loan such as loan amount, purpose of loan, and duration of loan. Consistent with prior work [16, 20], we use age as the sensitive attribute for this dataset where an age of 25 years or younger are considered the protected group. The outcome for this dataset is a binary variable indicating whether or not the loan contract has been fulfilled after the duration of the loan.

2.3 Metrics

Accuracy. For a given model, we measure its performance using accuracy defined[1] as $Acc = \frac{1}{N}\sum_{i=1}^{N}[[\hat{y}_i = y_i]]$ where $\hat{y}_i$ is the outcome predicted by the model, y_i is the true outcome and N is the number of samples we are evaluating [18]. We use accuracy as a

[1]We use $[[]]$ to denote the Iverson bracket which returns a value of 1 if the predicate contained within is true and 0 otherwise.

benchmark to evaluate the robustness of our fairness metric.

Equal Opportunity and Equalized Odds. A predictor is said to satisfy equal opportunity if and only if $\Pr(\hat{y} = 1|z = 1, y = 1) = \Pr(\hat{y} = 1|z = 0, y = 1)$ where z is a sensitive attribute. For example, in the COMPAS dataset, this can be interpreted as requiring the predictor to be agnostic to race for individuals who reoffend. We also consider a model where the predictor is modified to satisfy the stricter measure of *equalized odds* [17], that additionally enforces equal false positive rates. Formally, equalized odds requires the following to hold: $\forall a \in \{0, 1\} \quad \Pr(y = 1|z = 1, y = a) = \Pr(y = 1|z = 0, y = a)$.

Degree of Fairness and Direction of Unfairness. We measure the extent to which a model deviates from equal opportunity so that our fairness metric has the same support as accuracy, our benchmark metric. We define the *degree of fairness* of the predictor as: $1 - |\Pr(\hat{y} = 1|z = 1, y = 1) - \Pr(\hat{y} = 1|z = 0, y = 1)|$. The range of this measure is the unit interval $[0, 1]$ where a higher value indicates a fairer model. To identify the subgroup against which a predictor is biased, we define the *direction of unfairness* as $\text{sign}[\Pr(\hat{y} = 1|z = 1, y = 1) - \Pr(\hat{y} = 1|z = 0, y = 1)]$. For example, in the COMPAS dataset, $z = 1$ indicates an African American defendant and $z = 0$ indicates a Caucasian American defendant. So, a *positive* direction of unfairness corresponds to unfairness towards the protected group (in this case, African American defendants).

2.4 Model Choices

We learn twelve different models on the training data to evaluate their effects on both the mean and variance of fairness and performance metrics. In particular, we train a logistic regression classifier both with and without an ℓ_2-norm regularizer and both including and excluding sensitive attributes while training. In addition to these four models, we learn modified models by post-processing each of these models to separately satisfy first equal opportunity and then equalized odds.

3 UNIQUENESS OF APPROACH

3.1 Experimental Design

In order to split the datasets into training and held-out sets, we first randomly shuffle each dataset. For each dataset, we also ensure that the proportion of positive examples, the proportion of protected class, and the proportion of positive examples within the protected class are all preserved across the training and testing set. Then, we separately train the twelve models described in Section 2.4. For models trained with regularization we used 5-fold cross-validation to choose the hyperparameter that determines how much we penalize model complexity.

For the COMPAS dataset, we trained each model on 5000 points and held out 1150 for evaluation. For the Bank Marketing dataset, we trained each model on 25000 points and held out roughly 20000 points for evaluation. For the South German Credit dataset, we trained each model on 600 points and held out 400 for evaluation.

We evaluate the performance and fairness of each model on multiple test datasets generated from the held-out dataset using bootstrap sampling. Bootstrap sampling allows us to approximate the true distribution our test sample was drawn from [15]. So, crucially, these experiments are not evaluating robustness under a distribution shift, but rather robustness when we simply take multiple sets of samples from the same distribution. Each sample set was the same size as the held-out set and was created by uniformly picking a point from the held-out set with replacement. We created 800 such sampled datasets for each evaluation and then measured accuracy and degree of fairness on each sample dataset as described in Section 2.3. A schematic of this approach is shown in Figure 2.

We compute both the mean and variance of the degree of fairness and accuracy metrics. We then

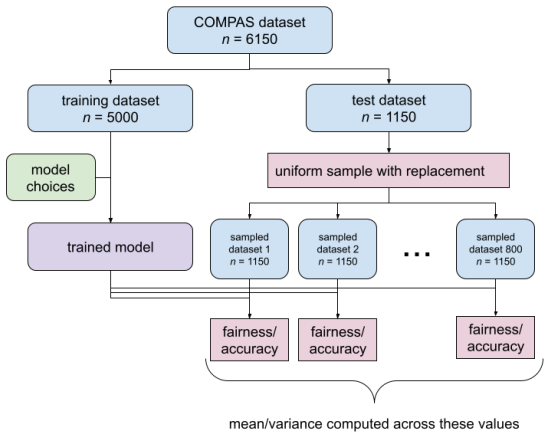


Figure 2: The flow of each of our experiments using the COMPAS dataset as an example. The same process is repeated for the SGC and Bank datasets.

compare the variance of these metrics over these 800 datasets in multiple ways.

3.2 Robustness Evaluation Technique

First, we numerically compute the variance achieved by these metrics and tabulate it for comparison across all twelve models (see Tables 3, 4, and 5).

Next, we create scatter plots and histograms of the values of both metrics for each of the bootstrap sampled datasets for a visual representation of the distribution of these measures. For the scatter plots, we use the same scale for both axes. A larger spread along a particular axis, therefore, indicates a larger variance along that metric. See Figure 6 for the scatter plots and see Figure 7 for the histograms for two models on the COMPAS dataset.

Lastly, we translate both measures from the $[0, 1]$ to the $(-\infty, +\infty)$ interval by first centering to 0.5 mean and then applying the logit function to the values so obtained[2]. We see that the mapped values broadly follow a normal distribution. We then compute the variance of these mapped values and apply

[2]Datasets with unit fairness were withheld in the F-test analysis to prevent degenerate cases. However, these accounted for less than 1.5% of all 800 sample datasets.

	No Sensitive		Sensitive	
Model	Accuracy	Deg of Fairness	Accuracy	Deg of Fairness
LogReg	62.16 (2.17)	78.32 (20.66)	62.73 (2.06)	61.50 (20.57)
LogReg + L2	62.16 (2.17)	78.32 (20.66)	62.37 (2.12)	65.01 (20.61)
EqOpp	58.72 (2.12)	96.54 (6.37)	56.63 (2.00)	96.27 (8.27)
EqOpp + L2	58.71 (2.12)	96.53 (6.42)	56.41 (2.07)	95.41 (10.53)
EqOdds	58.62 (2.09)	96.29 (7.78)	56.95 (1.99)	95.76 (9.65)
EqOdds + L2	58.61 (2.10)	96.30 (7.76)	56.89 (2.07)	95.40 (11.56)

Figure 3: Mean (and variance) values in percentage for accuracy and degree of fairness for the COMPAS dataset reported for Logistic regression (LogReg); postprocessing for equal opportunity (EqOpp) and equalized odds (EqOdds); L2 indicates regularization.

	No Sensitive		Sensitive	
Model	Accuracy	Deg of Fairness	Accuracy	Deg of Fairness
LogReg	78.45 (4.42)	90.18 (31.65)	77.30 (4.07)	85.07 (43.75)
LogReg + L2	77.20 (4.55)	91.94 (27.48)	78.53 (4.00)	88.07 (34.59)
EqOpp	76.49 (4.35)	93.51 (22.38)	73.58 (4.05)	92.14 (29.91)
EqOpp + L2	75.22 (4.55)	94.56 (16.92)	74.40 (3.88)	94.74 (16.28)
EqOdds	75.23 (4.34)	93.14 (23.70)	73.65 (4.08)	92.04 (30.34)
EqOdds + L2	74.15 (4.61)	94.27 (18.50)	74.48 (3.87)	94.99 (14.98)

Figure 4: Mean (and variance) values in percentage for accuracy and degree of fairness for the SGC dataset reported for Logistic regression (LogReg); postprocessing for equal opportunity (EqOpp) and equalized odds (EqOdds); L2 indicates regularization.

the F-test [35] to determine the significance of the difference in variances with high confidence[3]. The results are shown in Table 1.

We describe our results in the next section. Not all graphs and tables are included, but similar trends were observed in those that were omitted. We felt the included graphs best represented the trends and patterns we observed.

4 RESULTS AND CONTRIBUTIONS

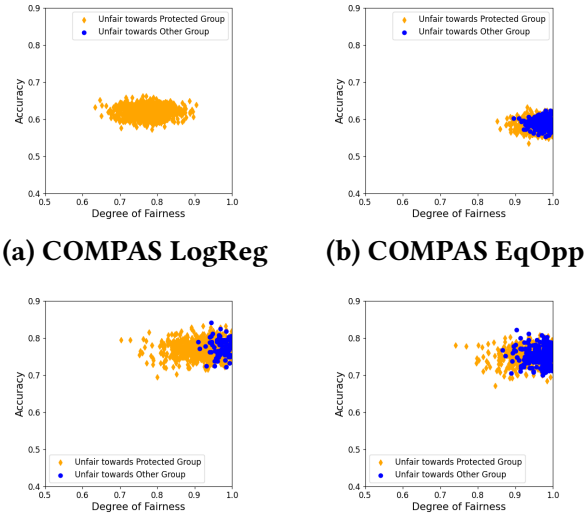
4.1 Variance of Fairness and Performance Metrics

As shown in Tables 3, 4, and 5 we note that the variance in degree of fairness is higher than for

[3]While the independence assumption does not strictly hold, the F-test gives us one more means of comparison.

Model	No Sensitive		Sensitive	
	Accuracy	Deg of Fairness	Accuracy	Deg of Fairness
LogReg	88.74 (0.05)	86.98 (3.10)	88.80 (0.04)	89.88 (3.86)
LogReg + L2	88.75 (0.05)	87.18 (2.94)	88.53 (0.05)	87.51 (3.05)
EqOpp	88.35 (0.05)	92.65 (2.20)	88.52 (0.05)	97.52 (2.50)
EqOpp + L2	88.33 (0.05)	92.36 (2.08)	88.35 (0.05)	95.22 (2.20)
EqOdds	72.32 (0.05)	93.57 (4.63)	77.40 (0.05)	94.89 (4.26)
EqOdds + L2	71.27 (0.05)	93.52 (4.77)	74.17 (0.05)	94.48 (4.51)

Figure 5: Mean (and variance) values in percentage for accuracy and degree of fairness for the Bank dataset reported for Logistic regression (LogReg); postprocessing for equal opportunity (EqOpp) and equalized odds (EqOdds); L2 indicates regularization.



(a) COMPAS LogReg **(b) COMPAS EqOpp**

(c) SGC LogReg **(d) SGC EqOpp**

Figure 6: Scatter plot for degree of fairness and accuracy. Orange diamonds indicate unfairness towards protected group, blue dots indicate unfairness towards the other group. Plots shown for the COMPAS and SGC datasets for Logistic regression (LogReg); postprocessing for equal opportunity (EqOpp) trained with regularization and without sensitive attributes.

accuracy. The blue boxes in these tables are the variances values for accuracy and the red boxes are the variance values for degree of fairness. The difference is most apparent in Table 5 which we believe is due to the larger dataset size. We show that this difference in variance is statistically significant for various significance levels (given by α values) in Table 1. We report values for the logistic regression base classifier with regularization trained on data with sensitive attributes both before and after post-processing for fairness constraints[4]. This indicates that the fairness metric of equal opportunity is not as robust as accuracy across the sampled test sets.

Once we post-process for fairness constraints, we see that, as expected, mean degree of fairness improves. We also note that the variance in degree of fairness reduces significantly, especially for the COMPAS dataset (see Table 3). This effect with the COMPAS dataset can be seen visually in Figure 7. We note, however, that the variance of degree of fairness is still statistically significant higher than the variance of accuracy for all models as seen in Table 1.

When comparing the effect of incorporating different fairness constraints, we note that both equalized odds as well as equal opportunity yield fairly similar results for degree of fairness. Typically, we observe that for models with post-processing for fairness constraints, means of degrees of fairness are within at most 1% of each other. We also observe that in most cases equal opportunity and equalized odds have comparable magnitudes of variance in degree of fairness. However, in the case of unregularized base classifiers, equal opportunity has a smaller degree of fairness variance; a likely explanation for this lies in our measure of degree of fairness which explicitly checks for deviation from the equal opportunity measure.

The effects of incorporating fairness constraints on accuracy have been previously observed [25]. This is corroborated in our experiments as we observe a trade-off between accuracy and degree of fairness. In all cases, adding a fairness constraint reduced overall accuracy; however, the effect on

[4]While we do not report results on all models due to space constraints, the omitted results are similar to reported values

Data set	Model	Ratio	$\alpha = 0.025$ 1.1488	$\alpha = 0.001$ 1.2446
COMPAS	LogReg	9.722	✓	✓
SGC	LogReg	8.648	✓	✓
Bank	LogReg	61	✓	✓
COMPAS	EqOpp	5.087	✓	✓
SGC	EqOpp	4.196	✓	✓
Bank	EqOpp	44	✓	✓
COMPAS	EqOdds	5.585	✓	✓
SGC	EqOdds	3.871	✓	✓
Bank	EqOdds	90.2	✓	✓

Table 1: F-test for statistical significance of the difference between performance and fairness variances reported for Logistic regression (LogReg); postprocessing for equal opportunity (EqOpp) and equalized odds (EqOdds). All models include sensitive attributes and a regularizer term. ✓ indicates that the ratio is higher than the F critical value, implying that the difference is statistically significant

its variance was typically minimal and inconsistent in direction indicating that adding fairness constraints does not seem to affect stability of the performance measure. Amongst models that were optimized for fairness, we notice that their mean accuracy is quite similar, being within at most 1% of each other's performance. This can be explained by the relationship between the fairness constraints and the degree of fairness measure. Another important trend we note is that higher mean degree of fairness generally corresponds to lower degree of fairness variance.

The effects of both including sensitive attributes in training the model, and adding a regularization term in the objective function, are mixed. The best performing models for accuracy are logistic regression models with access to sensitive attributes; perhaps unsurprisingly however, these are often among the worst performing with respect to the mean and variance of degree of fairness. We also want to acknowledge that non-sensitive attribute features can be highly correlated with the sensitive attribute features such that removing the sensitive attribute feature does not mitigate bias. There are proposed techniques that deal with this problem which are worth exploring over naively removing the sensitive attribute [41].

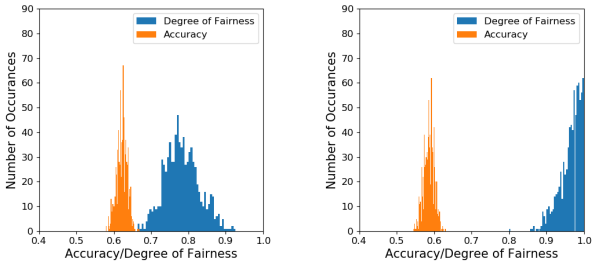
We also note that regularization has a significant effect on variance of degree of fairness especially when post-processing for fairness in the SGC dataset (Table 4) as compared to the COMPAS dataset (Table 3). This can be likely explained by the difference in sizes of the two datasets.

4.2 Direction of Unfairness

In addition to looking at the general trends of fairness, we also explore the direction of unfairness in these models for the SGC and COMPAS datasets. In Figure 6, we show a scatter plot of the 800 bootstrapped sampled test datasets (for both SGC and COMPAS datasets) along the accuracy and degree of fairness axes. As observable from the plots, generally the models are unfair towards the protected groups. Fairness constraints help shift the entire distribution to more fair outcomes, but we still see that most of the unfairness is to the detriment of protected groups. The plots for other models are omitted, but they show similar results as well.

5 CONCLUSION

In this work, we have provided a framework for evaluating the robustness of fairness metrics across uncertainty in test data. To do this, we resample test data using bootstrap sampling and compute both the mean and variance of degree of fairness and accuracy. This allows us to compare the variations across these metrics for different learning models. We train a logistic regression model for binary classification with and without a regularizer, as well as with and without sensitive attributes. We also post-process these models to separately satisfy two separate fairness constraints. We evaluate these twelve models separately on 800 bootstrapped test datasets to measure the variability as well as the mean of both a performance metric



(a) LogReg, no regularization, no sensitive attribute

(b) EqOdds, no regularization, no sensitive attribute

Figure 7: Histogram showing the difference in mean and variance of degree of fairness and accuracy scores for different models on the COMPAS dataset. Figure 7a includes scores for logistic regression without regularization and without sensitive attributes. Figure 7b is trained on the same settings as Figure 7a, but with the addition of post-processing for equalized odds fairness constraint.

and a fairness metric. We show that the equality of opportunity fairness metric is less robust to variations in the test data than the accuracy performance metric. We highlight that current post-processing methods for improving fairness can affect mean fairness and reduce fairness variance; by and large, however, the variance of fairness still remains significantly higher than that of performance. We show that variance in model fairness is typically to the detriment of protected groups, making fairness variance analysis an important part of developing robust and fair machine learning models.

This lack of robustness conclusion is of significance for the machine learning community because such algorithms are used for making decisions that affect people. A lack of fairness could have detrimental consequences to historically disadvantaged groups. As such, machine learning practitioners need to be confident in backing up their claim of the fairness of a machine learning model. Therefore, it is crucial to consider robustness when figuring out how to evaluate a model's fairness.

6 FUTURE WORK

Since we have only analyzed one fairness metric so far, we are interested in expanding this work to evaluate other kinds of fairness metrics including predictive parity rates or generalized entropy indices [36], as well as individual fairness metrics, such as Lipschitz conditions constraints [14]. From there, we would be interested in doing a more in-depth investigation as to why certain fairness metrics lack robustness. This will hopefully provide insight as to how we might remedy a lack of robustness.

I am going to be starting a PhD in Computer Science with a focus on fairness in machine learning next semester, and I hope to answer these questions and more during this next phase of my academic career.

7 ACKNOWLEDGEMENTS

I would like to thank Dr. Sindhu Kutty for all of her guidance on this project. I have learned so much in the past year, and I'm excited for future collaborations. I would also like to thank Andong Luis Li Zhao and Jillian Lew for their contributions to this project and our publication associated with this work.

REFERENCES

[1] Rediet Abebe and Kira Goldner. 2018. Mechanism Design for Social Good. *AI Matters* 4, 3 (Oct. 2018), 27–34.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. Propublica.

[3] Pranjal Awasthi, Matthäus Kleindessner, and Jamie Morgenstern. 2020. Equalized odds postprocessing under imperfect group information. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy] (Proceedings of Machine Learning Research, Vol. 108)*, Silvia Chiappa and Roberto Calandra (Eds.). PMLR, 1770–1780.

[4] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. 2021. It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks. *CoRR* abs/2106.05498 (2021).

[5] Solon Barocas, Moritz Hardt, and Arvind Narayanan. [n.d.]. *Fairness and Machine Learning*.

[6] Reuben Binns. 2020. On the Apparent Conflict between Individual and Group Fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 514–524.

[7] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 4356–4364.

[8] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 77–91.

[9] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building Classifiers with Independency Constraints. In *2009 IEEE International Conference on Data Mining Workshops*. 13–18.

[10] L. Elisa Celis, Anay Mehrotra, and Nisheeth K. Vishnoi. 2021. Fair Classification with Adversarial Perturbations. *CoRR* abs/2106.05964 (2021). arXiv:2106.05964

[11] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[12] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc* 7, 4 (2016).

[13] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[14] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[15] Bradley Efron and Robert Tibshirani. 1993. *An Introduction to the Bootstrap*. Springer.

[16] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A Comparative Study of Fairness-Enhancing Interventions in Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 329–338.

[17] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc.

[18] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*. Springer.

[19] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. *Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?* Association for Computing Machinery, New York, NY, USA, 1–16.

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

[21] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. Algorithmic Fairness. *AEA Papers and Proceedings* 108 (May 2018), 22–27.

[22] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA (LIPIcs, Vol. 67)*, Christos H. Papadimitriou (Ed.). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 43:1–43:23.

[23] Anja Lambrecht and Catherine Tucker. 2019. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management science* 65, 7 (2019), 2966–2981.

[24] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. *CoRR* abs/1908.09635 (2019). arXiv:1908.09635

[25] Aditya Krishna Menon and Robert C Williamson. 2018. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 107–118.

[26] Sérgio Moro, Paulo Cortez, and Paulo Rita. 2014. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems* 62 (2014), 22–31.

[27] Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P. Dickerson. 2021. Fairness Through Robustness: Investigating Robustness Disparity in Deep Learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 466–477.

[28] David C. Parkes, Rakesh V. Vohra, and et al. 2019. Algorithmic and Economic Perspectives on Fairness. *CoRR* abs/1909.05282 (2019). arXiv:1909.05282

[29] Geoff Pleiss. 2013. Code and data for the experiments in "On Fairness and Calibration".

[30] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On Fairness and Calibration. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc.

[31] Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation Sensitivity Analysis to Detect Unintended Model Biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 5739–5744.

[32] Ashesh Rambachan, Jon Kleinberg, Jens Ludwig, and Sendhil Mullainathan. 2020. An Economic Perspective on Algorithmic Fairness. *AEA Papers and Proceedings* 110 (May 2020), 91–95.

[33] Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D. Ziebart. 2021. Robust Fairness Under Covariate Shift. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 11 (May 2021), 9419–9427.

[34] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. 2020. How do fairness definitions fare? Testing public attitudes towards three algorithmic definitions of fairness in loan allocations. *Artif. Intell.* 283 (2020), 103238.

[35] George W. Snecdecor and William G. Cochran. 1991. *Statistical Methods*. Wiley-Blackwell.

[36] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2239–2248.

[37] **Kamp, Serafina**, Andong Luis Li Zhao, and Sindhu Kutty. 2021. Robustness of Fairness: An Experimental Analysis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 591–606.

[38] Cuong Tran, Ferdinando Fioretto, and Pascal Van Hentenryck. 2021. Differentially Private and Fair Deep Learning: A Lagrangian Dual Approach. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 11 (May 2021), 9932–9939.

[39] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 547–558.

[40] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2019. Fairness Constraints: A Flexible Approach for Fair Classification. *Journal of Machine Learning Research* 20, 75 (2019), 1–42.

[41] Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. 2022. Towards Fair Classifiers Without Sensitive Attributes: Exploring Biases in Related Features. (2022).