Balancing Multiple Fairness Definitions with Adversarial Learning

JULIE CESTARO* and JACKSON OLESON*, New York University, USA

Machine learning models, despite their potential benefits, can perpetuate and amplify existing societal biases present in training data. This raises concerns about fairness, particularly when these models are deployed in high-stakes applications, such as credit scoring, hiring, and criminal justice. A key challenge in mitigating bias is the incompatibility among different fairness definitions. This paper attempts to address this challenge by introducing a novel framework to balance multiple, potentially incompatible fairness definitions using composite adversarial training. We find that there is some promise to this approach, though more work needs to be done to ensure the stability of the system.

CCS Concepts: • **Computing methodologies** → **Machine learning approaches**.

Additional Key Words and Phrases: Adversarial Learning, Debiasing, Composite Adversarial Training

ACM Reference Format:

Julie Cestaro and Jackson Oleson. 2024. Balancing Multiple Fairness Definitions with Adversarial Learning. In *Proceedings of CSCI-GA.3033-112: Fair and Ethical Machine Learning (FML '24).* ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

When studying fairness in machine learning, three main classes of fairness definitions are commonly used: independence, separation, and sufficiency. However, research in this area has widely concluded that it is not possible to simultaneously satisfy multiple fairness definitions if they belong to different classes [1][3][5]. This leaves researchers, auditors, and engineers to commit to a single class of fairness definition and the inevitable risks that follow its benefits. For example, in the now famous analysis of the COMPAS recidivism algorithm, it was found that COMPAS meets the criteria for statistical parity but has highly unequal error rates. In other words, while COMPAS satisfies one fairness standard (from the "independence" class of definitions), it fails to meet another fairness standard (from the "separation" class of definitions).

To combat this trade-off, we investigate the feasibility of enforcing multiple fairness definitions simultaneously. We begin with a supervised classifier whose task is to predict some output variable Y given a set of input covariates X, while remaining unbiased with respect to some protected group membership Z. In measuring the bias of this classifier and subsequently debiasing[1] its predictions, we use the following three definitions used by Zhang et al.[7] in their related work:

*Both authors contributed equally to this research.

[1]Zhang et al.[7] use the word "debiasing" to describe this process, but additionally note that it is perhaps not entirely accurate. These methods remove some amount of bias, but not necessarily all bias, as the word "debiasing" may imply.

Authors' Contact Information: Julie Cestaro, julie.ces@nyu.edu; Jackson Oleson, jo1449@nyu.edu, New York University, New York, New York, USA.

Manuscript submitted to ACM

Definition 1. STATISTICAL PARITY.

A predictor $\hat{Y}$ satisfies statistical parity if $\hat{Y}$ and Z are independent. This means that for all values of the protected variable Z:

$$P(\hat{Y} = \hat{y}) = P(\hat{Y} = \hat{y} \mid Z = z)$$

Definition 2. EQUALITY OF ODDS.

A predictor $\hat{Y}$ satisfies equality of odds if $\hat{Y}$ and Z are conditionally independent given Y. This means that, for all possible values of the true label Y:

$$P(\hat{Y} = \hat{y}) = P(\hat{Y} = \hat{y} \mid Z = z, Y = y)$$

Definition 3. EQUALITY OF OPPORTUNITY.

If the output variable Y is discrete, a predictor $\hat{Y}$ satisfies equality of opportunity with respect to a class y if $\hat{Y}$ and Z are independent conditioned on $Y = y$.

This means that, for a particular value of the true label Y,

$P(\hat{Y} = \hat{y})$ is the same for all values of the protected variable: $P(\hat{Y} = \hat{y} \mid Y = y) = P(\hat{Y} = \hat{y} \mid Z = z, Y = y)$

Because the above definitions fall into different classes of fairness definitions (statistical parity is a measure of independence, equality of odds and equality of opportunity are measures of separation), we would ordinarily expect to choose one to enforce in our system. However, we would like to see a given fairness measure enforced without sacrificing performance under another. For example, if we would like to reduce the disparity across subgroups in access to a resource, we could imagine a scenario in which all subgroups are given equally bad access to a resource and then no one benefits. Because of this, we consider a system in which multiple adversaries work together to balance different fairness definitions.

We begin with the method set by Zhang et al.[7] in which a predictor f is trained to model Y as accurately as possible. An adversary g is then introduced which attempts to predict protected group membership Z based on the prediction $\hat{Y}$ and other relevant data. The gradient of g is then incorporated into the weight update of f to limit the amount of information about Z that is contained in $\hat{Y}$.

We extend this method to use two adversaries, g_1 and g_2, chosen such that each adversary in the pair is enforcing a complementary fairness definition. We then combine the losses of g_1 and g_2 as a weighted sum and use that sum in the weight update rule of f. This aims to satisfy two fairness measures simultaneously while maintaining the efficacy of the predictor f.

Our results show that this method can make small improvements on the original method from Zhang et al.[7] and continuing to follow the work of Tramer and Boneh[6] may lead us to a more consistent and reliable solution.

The contributions of this paper include:

- A novel application of composite adversarial methods to enforce multiple fairness definitions
- A demonstration of the viability of a composite adversarial framework using a synthetic dataset
- An analysis of the performance of this framework on three benchmark datasets for fairness research

2 Related Work

2.1 Debiasing

As discussed above, our work draws primarily from the work of Zhang et al.[7] and their method for debiasing predictions using an adversary. While novel in structure, their method is stifled by the common limitation in algorithmic

fairness work where one must commit to a single fairness definition to enforce in the system. Our work strives to look beyond that limitation by combining fairness definitions.

2.2 Composite Adversarial Training

In considering how to combine these adversaries, our initial approach is most closely connected to that of Tramer and Boneh[6] in their paper Adversarial Training and Robustness. We found that the questions proposed by Tramer and Boneh[6] with regards to adversarial attacks are fundamentally similar to those that we ask in this paper with regards to fairness definitions. Our experiments in combining adversarial losses as a weighted sum, outlined in Section 3.3, is derived from their "avg" strategy for combining adversarial attacks.

3 Method

The framework of our approach begins with that of Zhang et al.[7]. We train a predictor model f to model Y given the input features X and call those predictions $\hat{Y}$. The adversary g then attempts to predict protected class membership Z based on the predictions $\hat{Y}$ under some condition: either given the true labels Y, subject to some restriction of the data, or both. The objective is to maximize the predictor's ability to determine the output variable Y while minimizing the adversary's ability to predict Z.

3.1 Adversaries

We differentiate between each of the adversaries based on the information to which they have access during training:

- The **statistical parity** adversary only has the predictions $\hat{Y}$.
- The **equality of odds** adversary has both the predictions $\hat{Y}$ and the true labels Y.
- The **equality of opportunity** adversary has both the predictions $\hat{Y}$ and the true labels Y, but only gets a subset of the whole dataset where the true label $Y = y$ and $y \in Y$.
- The **true positive rate (TPR)** adversary is a more specific version of the equal opportunity adversary above where $y = 1$ in the binary case.
- The **false positive rate (FPR)** adversary is a more specific version of the equal opportunity adversary above where $y = 0$, again in the binary case.

3.2 Determining Adversarial Pairs

We first construct two pairs of complementary adversaries so that the classifier will have to explicitly satisfy both. To understand why we do this, consider the equality of odds adversary. The equality of odds adversary has access to all of the available information, so if the classifier is able to maximize the loss of the equality of odds adversary, it is likely to maximize the loss of the other adversaries. To avoid this undesirable scenario, where both adversaries are satisfied because one is simply a less discriminating version of the other, we separate them into complementary pairs.

The first pair consists of one adversary enforcing equality of false positive rate (FPR) and another enforcing equality of true positive rate (TPR). Because each of these adversaries have access to different information (FPR has access to the data where $Y = 0$ and TPR has access to the data where $Y = 1$) they make their predictions for different subsections of the dataset. We validate our method by comparing against a single Equalized Odds adversary which has access to the entire dataset.

The second pair of adversaries consists of one adversary enforcing equal opportunity and one enforcing statistical parity. Like the FPR and TPR adversaries, the equal opportunity and statistical parity adversaries also have access to different information: equal opportunity has access to both the prediction and real outcome but only for a subset of the data, and statistical parity has access to the whole dataset, but not their true labels.

3.3 Adversarial Framework

The original architecture for the adversarial system set forth by Zhang et al.[7] consists of updating the gradient of the predictor ($\nabla_W L_P$) at each training step by subtracting the gradient of the adversary ($\nabla_W L_A$), thereby incentivising the predictor to maximize the loss of the adversary. They also include a projection term to prevent the predictor from inadvertently helping the adversary. Without the projection term, the gradients of the predictor can be in a direction that helps the adversary also decrease its loss.

$$\nabla_W L_P - \text{proj}_{\nabla_W L_A} \nabla_W L_P - \alpha \nabla_W L_A$$

We experimented with extending this system to incorporate a second fairness definition following the work of Tramer and Boneh[6]. Consider either of the adversarial pairs outlined in Section 3.2 above. We will refer to each adversary in a given adversarial pair as g_i and its corresponding loss as L_{A_i}. In this method, we alter the loss of the adversary (L_A in the equation above) to be a weighted sum of the losses of the adversaries in an adversarial pair.

$$L_{A_{sum}} = (c_1 * L_{A_1}) + (c_2 * L_{A_2})$$

For the adversarial pair of statistical parity and equal opportunity evaluated on the synthetic dataset, we set $c_1 = c_2 = 0.9$ where c_1 and c_2 are the weights we put on the adversarial gradients in the update step. For the adversarial pair of TPR and FPR evaluated on the benchmark datasets, we set $c_1 = c_2 = 0.5$. This loss is calculated prior to the gradient calculation step and then we incorporate this as one term when the gradient of the predictor is updated as follows:

$$\nabla_W L_P - \text{proj}_{\nabla_W L_A} \nabla_W L_P - \alpha \nabla_W L_{A_{sum}}$$

4 Evaluation

4.1 Datasets

We perform our evaluation on a synthetic dataset and three datasets commonly used in fairness literature[2]. Our synthetic dataset is evaluated using the adversarial pair of statistical parity and equalized opportunity. The three benchmark datasets are evaluated using the adversarial pair of FPR and TPR. Each of the the three benchmark datasets pre-define a privileged class and an unprivileged class, which we reference in our performance metric formulations in Section 4.2 below.

4.1.1 Synthetic Data. We recreate the synthetic dataset given by Zhang et al.[7], with a few modifications. The dataset is a synthetic training example $(x^{(i)}, y^{(i)}, z^{(i)})_{i=1}^{n}$. For each i we let the protected variable $r \in \{0, 1\}$ be picked uniformly at random. We let $v \sim \mathcal{N}(r_i, 1)$ and $u, w \sim \mathcal{N}(v_i, 1)$ be independent. We let $x^{(i)} = (r, u)$, $y^{(i)} = \mathcal{I}(w > 0)$, $z = r$. To increase accuracy, we can optionally set $x^{(i)} = (r, u, f, g)$ where $f \sim \mathcal{N}(y_i, 1)$ and $g \sim \mathcal{N}(y_i, 0.2)$.

4.1.2 UCI Adult. The UCI Adult Dataset[2], also known as the Census Income Dataset, is used to predict whether an individual's annual income exceeds 50,000 based on various demographic and work-related attributes. It includes features such as age, education, occupation, marital status, race, and sex. We use sex as the basis for the privileged and unprivileged class where Male is the privileged class and Female is the unprivileged class.

4.1.3 German Credit. The German Credit Dataset is designed to assess the creditworthiness of loan applicants by predicting whether they are a "good" or "bad" credit risk. It consists of attributes like age, credit history, loan purpose, and employment status. We use sex as the basis for the privileged and unprivileged class where Male is the privileged class and Female is the unprivileged class.

4.1.4 COMPAS. The COMPAS Dataset is used to predict recidivism risk—whether a defendant is likely to reoffend—based on features such as age, prior convictions, charge degree, and criminal history. We use race as the basis for the privileged and unprivileged class where Caucasian is the privileged class and Non-Caucasian is the unprivileged class.

4.2 Metrics

We evaluate the efficacy of our method using the following five metrics, some of which are based on the pre-defined privileged and unprivileged classes described for each dataset in Section 4.1:

4.2.1 Classification accuracy. This is the classic metric for understanding if primary predictor is still useful after performing our adversarial debiasing process. We calculate classification accuracy in the standard way:

$$\frac{\text{true positives} + \text{true negatives}}{\text{total predictions}}$$

4.2.2 Balanced classification accuracy. This metric is useful for understanding how our adversarial framework impacted the overall true positive rate (TPR) and overall true negative rate (TNR) of the primary predictor. We calculate balanced classification accuracy as the average of TPR and TNR:

$$0.5 * (\text{TPR} + \text{TNR})$$

4.2.3 Disparate impact. This metric helps us to understand how our adversarial framework impacted the predictions for the privileged class and the unprivileged class. We calculate disparate impact as follows:

$$\frac{Pr(\hat{Y} = 1 | Z = \text{unprivileged})}{Pr(\hat{Y} = 1 | Z = \text{privileged})}$$

4.2.4 Difference of equal opportunity. This metric helps us to understand how our adversarial framework impacted the true positive rate (TPR) specifically for the privileged class and the unprivileged class. We calculate difference of equal opportunity as follows:

$$TPR_{Z=\text{unprivileged}} - TPR_{Z=\text{privileged}}$$

Note that when this metric is positive the true positive rate is higher for the unprivileged class, and when it is negative the true positive rate is higher for the privileged class.

[2]We acknowledge the work of Ding et al.[4] which demonstrates the idiosyncratic properties and largely outdated information in the UCI Adult Dataset. We choose to use this dataset even in light of these findings due to its benchmark status.

4.2.5 Difference of average odds. This metric helps us to understand how our adversarial framework impacted equality of odds for the privileged class and the unprivileged class. We calculate difference of average odds as follows:

$$\frac{1}{2}\left[(FPR_{Z=\text{unprivileged}} - FPR_{Z=\text{privileged}}) + (TPR_{Z=\text{unprivileged}} - TPR_{Z=\text{privileged}})\right]$$

Note that a value of 0 in this metric indicates equality of odds.

4.3 Baselines for Adversarial Pairs

Recall the adversarial pairs outlined in Section 3.2. When evaluating the efficacy of a given pair of adversaries, we compare them to the impact of a single adversary operating alone as a baseline, rather than to the plain model without debiasing.

When evaluating the pair of adversaries in which one attempts to balance TPR and one attempts to balance FPR, we evaluate against a single Equalized Odds adversary. Fundamentally, attempting to satisfy equality of odds can be broken down into balancing TPR in addition to FPR. Therefore we want to know if separately enforcing TPR and FPR improves on equality of odds.

When evaluating the pair of adversaries in which one attempts to satisfy statistical parity and one attempts to satisfy equality of opportunity, we evaluate against each of these adversaries acting alone. We hypothesize that adding an additional adversary enforcing an additional fairness definition will improve the general fairness of the predictor at the expense of the individual fairness metric associated with the single adversary.

5 Empirical Results

We collected two types of results for this paper: first demonstrating that multiple fairness definitions satisfied via the weighted sum approach is viable via a synthetic dataset, and then applying this framework to three benchmark datasets.

We implemented our framework for the benchmark datasets as an extension of the existing implementation of the Zhang et al.[7] method in the AIF360 library[2] and ran those experiments within this existing framework. Overall, we found the adversarial performance on the benchmark datasets to be fairly inconsistent, both with the original adversary framework and with our extended methodology using a weighted sum of two losses. Because of this, we performed five trials each with a single adversary as well as with our weighted sum framework for each of the three datasets and then plotted the results for comparison. As stated in Section 4.3, we compare each adversarial pair to its relevant baseline in our results below. The discrete datapoints for the figures related to the benchmark datasets can be found in Appendix A.

5.1 Performance on Synthetic Data

In parallel with Zhang et al.[7], a predictor model trained in lockstep with an adversary model learns to predict y without the use of protected variable z and achieves statistical parity. We introduce a second adversarial logistic regression model with the goal to achieve equality of opportunity. We show in Table 1 that the ability of the multi-adversary model to achieve statistical parity and equal opportunity declines slightly, but still performs well in these respective cases and is also able to maintain high accuracy.

5.2 Performance on UCI Adult Dataset

We refer to the graphs in Figure 1 for the following analysis. We notice that there are minimal changes to the classification accuracy and balanced classification accuracy of the primary predictor between the single adversary and the adversarial pair. Additionally, we see some undesirable movement in disparate impact, equal opportunity difference, and average

Table 1. Performance on Synthetic Data

Evaluation Metric	Parity Adversary	EO Adversary	Multiple Adversaries
Accuracy	0.997	0.990	0.964
Balanced Classification Accuracy	0.930	0.964	0.943
Disparate Impact	0.818	1.041	0.770
Difference of Equal Opportunity	0.058	0.028	0.038
Difference of Average Odds	0.109	0.023	0.071

odds difference. Ideally we would like to see a disparate impact metric close to 1, but the adversarial pair performs lower than the single adversary. We also strive for equal opportunity difference and average odds difference to be close to zero, but the adversarial pair produces values for these metrics that are farther away from zero than the single adversary.

5.3 Performance on German Credit Dataset

We refer to the graphs in Figure 2 for the following analysis. We note that, as with the UCI Adult dataset, there are minimal changes to the classification accuracy and balanced classification accuracy of the primary predictor between the single adversary and the adversarial pair. However, unlike the results on the UCI Adult dataset, we see more promising results in the other metrics. We see that the adversarial pair resulted in a disparate impact much more consistently close to 1 than the single adversary. We also see that the equal opportunity difference and average odds difference are much more consistently close to zero, which is also desirable.

5.4 Performance on COMPAS Dataset

We refer to the graphs in Figure 3 for the following analysis. We again note minimal changes to the classification accuracy and balanced classification accuracy of the primary predictor between the single adversary and the adversarial pair. As with the UCI Adult dataset, we see similar undesirable decreases in disparate impact with the adversarial pair, and we also see that equal opportunity difference and average odds difference move further away from zero.

6 Discussion

Our results show that there is some promise to the notion that multiple fairness definitions can be enforced simultaneously. Our evaluation on the synthetic example showed promising results and demonstrated the feasibility of a multi-adversary system allowing for a slight dip in performance on any individual adversary in lieu of balanced performance for multiple fairness definitions. We also saw an overall improvement in the performance of the adversarial pairs on the German Credit Dataset for all of the fairness metrics (Figure 2). We did not see large decreases in overall classification accuracy on any of the datasets with the use of our system. The inconsistencies in performance suggest that our model may be sensitive to characteristics of the dataset and that there may exist some configuration that can account for this in future iterations of this framework.

7 Conclusions

In this paper we applied Tramer and Boneh's[6] "avg" method to extend the work of Zhang et al.[7] in adversarial debiasing. We introduce the notion of adversarial pairs, in which two adversaries attempt to satisfy complementary measures of fairness while preserving the accuracy of the original predictor. We evaluated our framework on a synthetic

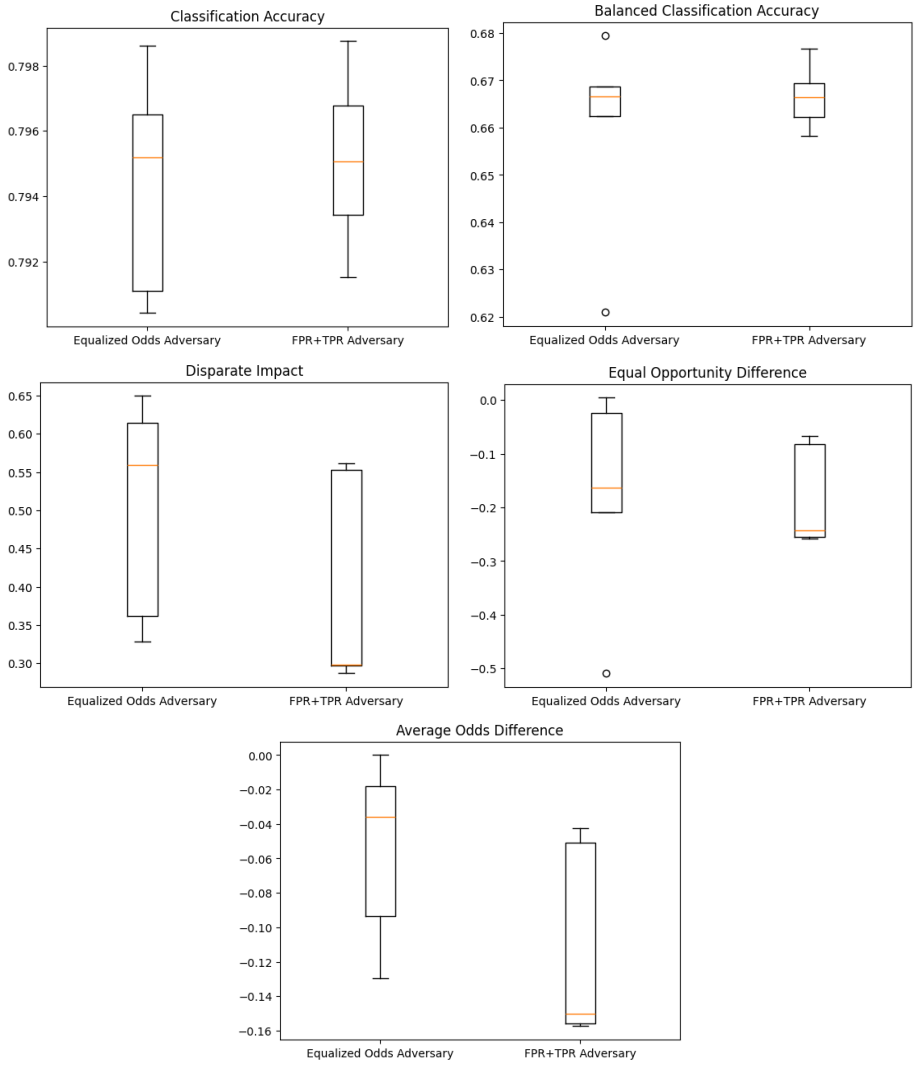

Fig. 1. Trial Results on the UCI Adult Dataset

dataset based on that of Zhang et al.[7] and three different benchmark datasets, ensuring that our basis for comparison was appropriate for the adversarial pair in use. Though we were hoping for more consistent and obvious improvement over the traditional single adversary framework across all datasets, we saw mild improvements in our evaluation on our synthetic dataset as well as on the German Credit Dataset. This work contributes to ongoing efforts towards developing more equitable machine learning systems, although we recognize there is more work to be done.

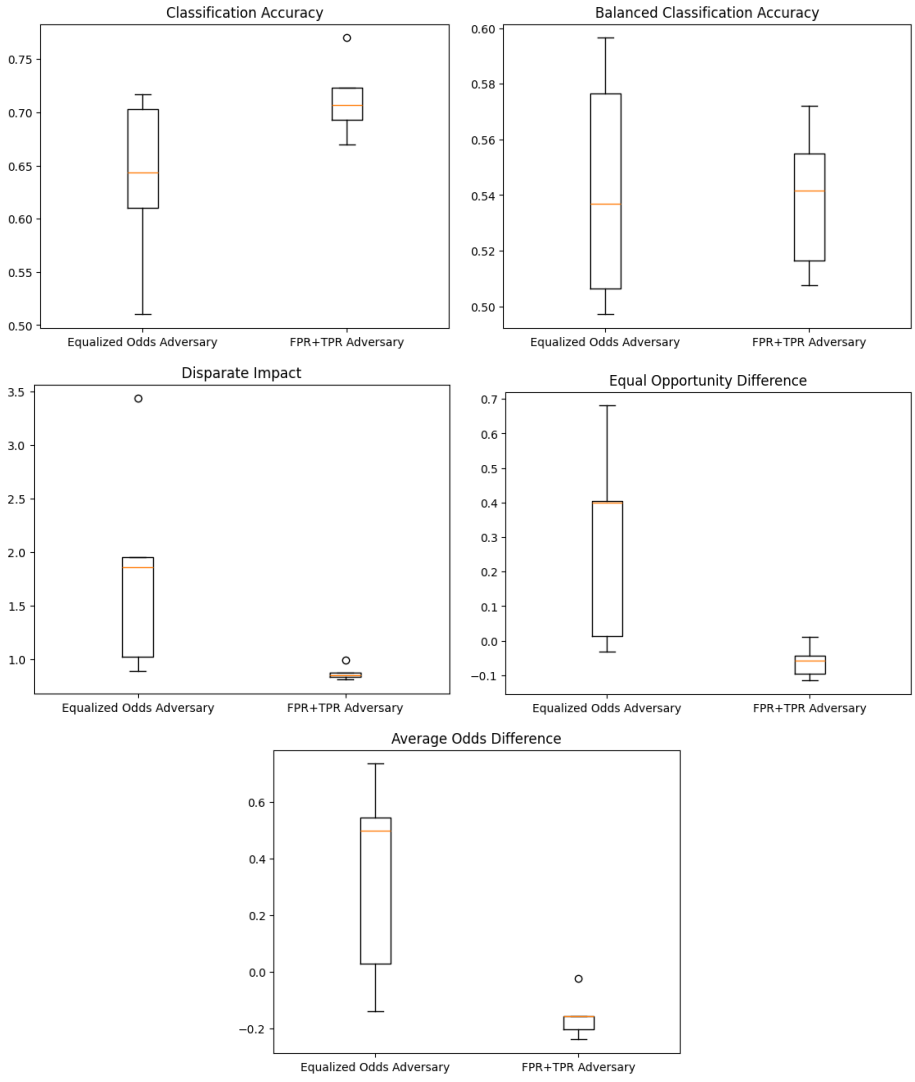

Fig. 2. Trial Results on the German Credit Dataset

7.1 Limitations and Future Work

We acknowledge the limitations of this work, particularly the inconsistent results observed across different datasets, which reflect the known sensitivity of the adversarial debiasing method on which ours is based[7] and the sensitivity of adversarial training in general.

Although we would have ideally liked to see more improvement from our adversarial pairs method, it is worth noting that Tramer and Boneh[6] saw similarly unremarkable performance with their "avg" method, to which our adversarial pairs are related. Given that, a natural next step for this work is to attempt to replicate their affine attack method[6] with our adversarial pairs.

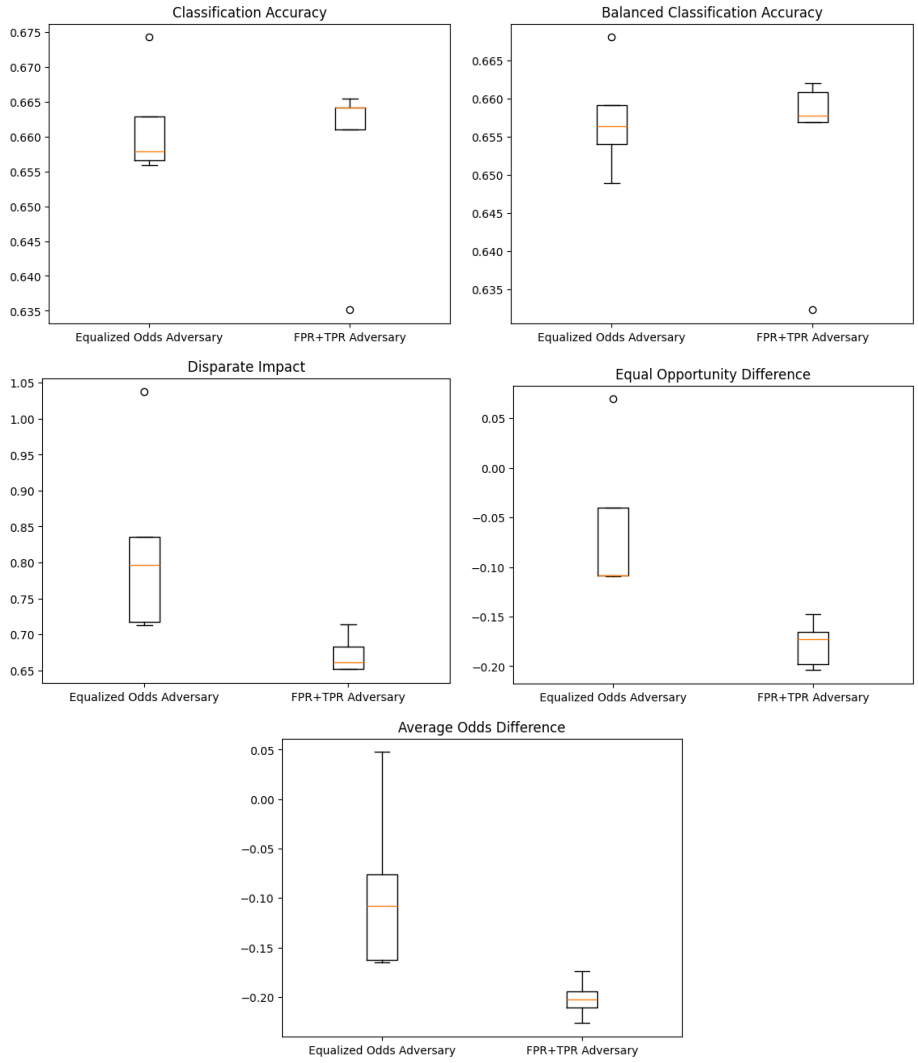

Fig. 3. Trial Results on the COMPAS Dataset

Furthermore, in our adversarial framework, we statically set the weights on each loss to be 0.5 in the TPR and FPR case and 0.9 in the statistical parity and equal opportunity case, as we describe in Section 3.3. It would be very interesting to see results for a spectrum of different weight combinations. Perhaps there is a weight combination that performs better for certain settings.

Acknowledgments

Thank you to Daniel, without whom this paper would not exist – mainly because he is the one who asked us to write it.

References

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities.* MIT Press. 65–66 pages.

[2] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. https://arxiv.org/abs/1810.01943

[3] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163. https://doi.org/10.1089/big.2016.0047 arXiv:https://doi.org/10.1089/big.2016.0047 PMID: 28632438.

[4] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2022. Retiring Adult: New Datasets for Fair Machine Learning. arXiv:2108.04884 [cs.LG] https://arxiv.org/abs/2108.04884

[5] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. arXiv:1609.05807 [cs.LG] https://arxiv.org/abs/1609.05807

[6] Florian Tramer and Dan Boneh. 2019. Adversarial Training and Robustness for Multiple Perturbations. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/5d4ae76f053f8f2516ad12961ef7fe97-Paper.pdf

[7] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. arXiv:1801.07593 [cs.LG] https://arxiv.org/abs/1801.07593

Table 2. Performance on UCI Adult Dataset

Adversarial Framework	Accuracy	Balanced Accuracy	Disparate Impact	Equal Opportunity Difference	Avg Odds Difference
Equalized Odds Trial 1	0.790418	0.620919	0.361358	-0.163762	-0.093745
Equalized Odds Trial 2	0.795196	0.662501	0.649369	0.004374	-0.000032
Equalized Odds Trial 3	0.791101	0.668691	0.614141	-0.024199	-0.017945
Equalized Odds Trial 4	0.796492	0.679392	0.559284	-0.509660	-0.036039
Equalized Odds Trial 5	0.798608	0.666636	0.328075	-0.208639	-0.129651
TPR and FPR Trial 1	0.791510	0.666484	0.561367	-0.082822	-0.051050
TPR and FPR Trial 2	0.793421	0.658196	0.296683	-0.257454	-0.157459
TPR and FPR Trial 3	0.798744	0.669322	0.286756	-0.254364	-0.155608
TPR and FPR Trial 4	0.796765	0.676743	0.552745	-0.067559	-0.042595
TPR and FPR Trial 5	0.795059	0.662186	0.298276	-0.242733	-0.150193

Table 3. Performance on German Credit Dataset

Adversarial Framework	Accuracy	Balanced Accuracy	Disparate Impact	Equal Opportunity Difference	Avg Odds Difference
Equalized Odds Trial 1	0.716667	0.536985	0.891827	-0.032244	-0.140834
Equalized Odds Trial 2	0.610000	0.576593	1.858407	0.404110	0.498930
Equalized Odds Trial 3	0.510000	0.497105	3.431034	0.680272	0.734367
Equalized Odds Trial 4	0.643333	0.596463	1.953271	0.400000	0.543750
Equalized Odds Trial 5	0.703333	0.506497	1.020305	0.013158	0.026987
TPR and FPR Trial 1	0.693333	0.507496	0.988636	0.009769	-0.023321
TPR and FPR Trial 2	0.723333	0.554996	0.871613	-0.042991	-0.156470
TPR and FPR Trial 3	0.706667	0.541686	0.807692	-0.096154	-0.240385
TPR and FPR Trial 4	0.670000	0.516513	0.846764	-0.115800	-0.157900
TPR and FPR Trial 5	0.770000	0.572133	0.834138	-0.056917	-0.204595

Table 4. Performance on COMPAS Dataset

Adversarial Framework	Accuracy	Balanced Accuracy	Disparate Impact	Equal Opportunity Difference	Avg Odds Difference
Equalized Odds Trial 1	0.790418	0.620919	0.361358	-0.163762	-0.093745
Equalized Odds Trial 2	0.795196	0.662501	0.649369	0.004374	-0.000032
Equalized Odds Trial 3	0.791101	0.668691	0.614141	-0.024199	-0.017945
Equalized Odds Trial 4	0.796492	0.679392	0.559284	-0.509660	-0.036039
Equalized Odds Trial 5	0.798608	0.666636	0.328075	-0.208639	-0.129651
TPR and FPR Trial 1	0.791510	0.666484	0.561367	-0.082822	-0.051050
TPR and FPR Trial 2	0.793421	0.658196	0.296683	-0.257454	-0.157459
TPR and FPR Trial 3	0.798744	0.669322	0.286756	-0.254364	-0.155608
TPR and FPR Trial 4	0.796765	0.676743	0.552745	-0.067559	-0.042595
TPR and FPR Trial 5	0.795059	0.662186	0.298276	-0.242733	-0.150193

A Tabular Data for Figures

Received 13 December 2024