Replication: Fairness without Demographics through Adversarially Reweighted Learning

Sk Golam Saroar
saroar@yorku.ca
York University

ABSTRACT

Literature in the field of fairness in Machine Learning (ML) often assumes that protected features such as race and gender are accessible during training or at inference. However, due to privacy concerns and regulations, these features are frequently left out in practise. As a result, in this paper [5], the authors investigated the following question: "How to train an ML model to improve fairness without access to protected features?" They devised the Adversarially Reweighted Learning (ARL) approach to address this issue. In this replication study, I examine the claim that ARL enhances Rawlsian Max-Min fairness for supervised classification when compared to earlier approaches and simple baselines. Using the paper and available code as a guide, I re-implemented all models in PyTorch. I also used my implementation to replicate the hyperparameter search described in the paper. The advantage of ARL over the investigated baselines could not be replicated. Although the results of the replicated ARL do not differ considerably from those of the paper, a stronger baseline performance appears to be negating the advantages of the ARL.

KEYWORDS

Fairness in Machine Learning, Neural Networks, Adversarial Learning, Replication Package

ACM Reference Format:
Sk Golam Saroar. 2022. Replication: Fairness without Demographics through Adversarially Reweighted Learning. In *Proceedings of* . Example Venue, 8 pages. https://doi.org/example/doi

1 INTRODUCTION

In recent years, Machine Learning (ML) systems have become extremely popular for critical decision making, which makes it essential that these systems are fair towards different groups. Sadly, that is not the case, as recent research has raised various issues about fairness including areas such as healthcare systems [3], face detection [1], etc. In response, there has been a surge of research towards fairness in ML. However, most of these studies assume that protected features (for example, race, religion, gender) are present in the dataset and rely on these features to improve fairness. In reality, it is frequently impractical to use protected features due to privacy or regulatory restrictions. For example, regulators such as CFBP require creditors to be fair, but creditors are prohibited from utilizing demographic data for their decision-making. Therefore, addressing fairness without demographics has been characterised as a critical open-problem of significant importance to ML practitioners.

In this paper, the authors propose adversarially reweighted learning (ARL), which provides a foundation for how to pursue fairness without access to demographics. This optimization modeling approach aims to maximize the minimum utility for worst-off protected groups by assuming that these groups are computationally-identifiable through other attributes. ARL considers a minimax game between a learner and adversary, both of which are implemented using a standard feed-forward network. Evaluating ARL on three real-world datasets, the authors showed that, with a significant AUC improvements for worst-case protected groups, ARL outperformed the state-of-the-art alternatives on all the datasets. This paper also presents insights into the inner workings of ARL by analyzing the learned example weights. Through a synthetic study, the authors observed that ARL is quite robust to representation bias but degrades with noisy ground truth labels.

The purpose of this replication study is to test the replicability of the paper. This entails re-implementing the ARL model and experiments in the paper and trying to come up with answers to the following:

- Are my own results in line with the results reported by the authors?
- What challenges did I face when replicating the authors' method?
- Can I support the explicit and implicit assumptions made by the authors to substantiate their method?

The rest of this report is organized as follows. Section 2 presents the experiments. Section 3 describes the results. Section 4 discusses the discrepancies. Section 5 concludes this report.

2 EXPERIMENTS

A few details about the ARL model are missing in the paper, for example, the embedding size, pre-train steps, the optimizer, etc. The authors refer to their publicly available TensorFlow implementation [4] for such details. In order to replicate the paper as closely as possible, I have used their Tensorflow code only as a guide to fill in the missing information. I have used PyTorch for re-implementing the models in my replication study.

This paper's key empirical assertion is that ARL performs better than numerous other methods in terms of performance

, ,
© 2022
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00
https://doi.org/example/doi

and fairness across a variety of datasets. Distributionally Robust Optimization (DRO), Inverse Probability Weighting (IPW), and a simple ERM baseline (Baseline) are the models they consider for comparison. Except for DRO, I generally followed the paper for implementing these models. For DRO, I modified the publicly available implementation [7] to fit the framework I wrote for the other methods. I also performed a grid search to obtain the best hyperparameters.

In this section, I start by discussing the models very briefly. Then I give a overview of the datasets and my preprocessing steps, followed by details about how I got the hyperparameters. I build on these steps to compare my replication with the paper, which is discussed in section 3.

2.1 Models

ARL, DRO, and IPW have one thing in common: they all try to improve fairness by changing the optimization objective. The inference model's network architecture is the same for all approaches. I employed a simple fully-connected feed forward neural network to train on the three datasets. The network, as described in the paper, is made up of two hidden layers of 64 and 32 units each, with ReLU activation functions added after each hidden layer. The output layer consists of one output unit for the binary classification tasks.

2.1.1 Adversarially Reweighted Learning (ARL). ARL is made up of two simple feed-forward networks, one for the Learner and one for the Adversary. Each individual in the sample is assigned to an unobserved protected subgroup S, such as race or gender, in a binary classification task. The model's goal is to maximise the minimal utility U across all groups $s \in S$. Protected group membership S, on the other hand, is not available during training or inference. For classification tasks, the Learner is a basic iterative learning algorithm that aims to minimise the predicted loss at each training step. The Adversary tries to find the computationally-identifiable subgroups in which the Learner makes substantial mistakes, then maximises the weighted loss in those regions. As a result, the Learner is compelled to improve in those areas. Both networks are alternatively trained.

Implementation: The Learner is a fully-connected two-layer feed-forward network with 64 and 32 units in the hidden layers respectively, and a sigmoid activation function. I implemented the Adversary using a single linear layer with 32 hidden units and a sigmoid activation function. As per the TensorFlow implementation, I used categorical embeddings of size 32 and Adagrad optimizer for both models.

[**Discrepancy**:The authors employed two consecutive linear layers in their code, which is not fully aligned with the paper. Their code also reveals that the learner network was pre-trained, which they do not mention in the paper.]

2.1.2 Distributionally Robust Optimization (DRO). Similar to ARL, DRO also aims to improve Rawlsian Max-Min Fairness without access to demographics. DRO considers any worst-case distribution exceeding a given size α as a potential protected group and optimizes for improving the worst-case performance of any set of examples exceeding size α. DRO thus implicitly decides about the size of the worst-case group. By focusing on improving any worst-case distributions, DRO runs the risk of focusing the optimization on noisy outliers. ARL, on the other hand, relies on the concept of computational-identifiability.

2.1.3 Inverse Probability Weighting (IPW). Rather than learning the weights to reweight the loss function dynamically, a simpler strategy is to weight each sample with the inverse probability of obtaining this sample from the dataset. This gives the dataset's underrepresented groups a higher weight. The IPW(S) method calculates probability based solely on membership in protected categories (such as race or gender), resulting in weights of the form 1/p(s). In addition to the protected group (S), we can also consider the label (Y), resulting in the IPW(S+Y) model. The probabilities are computed as the combined probabilities of being a member of a protected group and belonging to a specific label, for example being a "black female" (S) and "passing the bar exam" (Y). This results in weights of the form 1/p(s, y).

The choice of IPW as one of the experiments for this study is very interesting. This provides a opportunity to compare the ARL's approach to increasing fairness without demographics against a method that increases fairness with demographics.

2.1.4 Empiricial Risk Minimization (Baseline). This model is identical to the ARL learner. The authors say they raised the number of hidden units in this model to account for the adversary's enhanced capabilities in ARL. However, it was not clear from the paper how they achieved this, nor was it obvious from their code that they did so. A single more unit would result in more parameters than ARL based on parameter count. I did not add any more units because the simple fully-connected model already matched ARL's performance (see Table 4).

2.2 Datasets

2.2.1 Overview. In this replication study, I have used the same real-world, publicly available datasets as used in the paper: adult income (UCI Adult) [2], law school admission (LSAC) [8], and criminal recidivism (COMPAS) [6]. Each dataset describes a binary prediction task. In the UCI Adult dataset, the task is to predict if an individual's income is above 50 thousand dollars. The LSAC dataset is used to predict whether a candidate would pass the bar exam. The COMPAS dataset consists of criminal records comprising offender's criminal history and the task is to predict the recidivism of offenders. All three datasets include information about sensitive demographic attributes such as race and gender which makes them suitable for evaluating model fairness. The datasets also exhibit imbalance in these sensitive attributes.

Table 1: Overview of the used datasets, including protected features, and protected groups.

Dataset	Size	No. of Features	Protected Features	Protected Groups	Prediction Task
Adult	48842	15	Race, Sex	$\{White, Black\} \times \{Male, Female\}$	Income above 50k?
LSAC	26551	12	Race, Sex	$\{White, Black\} \times \{Male, Female\}$	Pass bar exam?
COMPAS	7214	11	Race, Sex	$\{White, Black\} \times \{Male, Female\}$	Recidivate in 2 years?

Table 2: Imbalance in the datasets between different protected groups.

Dataset	Size	#Male	#Female	%Male	%Female	#White	#Black	%White	%Black
Adult	48842	32650	16192	0.668	0.332	41762	4685	0.855	0.096
LSAC	26551	14873	11678	0.560	0.440	21936	1790	0.826	0.067
COMPAS	7214	5819	1395	0.807	0.193	2454	3696	0.340	0.512

Table 1 shows an overview of the used datasets, including number of features, the protected features, protected groups and prediction task of each dataset. Table 2 displays a further breakdown of the protected groups in each dataset. We can see that Female and Black population are underrepresented in the datasets.

2.2.2 Preprocessing. The data preprocessing step is the only step where I have essentially reused the authors' provided code. The reasons are: (1) I wanted to train my models on the exact data as the authors used (2) Data preprocessing is a small and trivial task considering the overall volume of work in this replication study, and (3) I have had to make enough changes (discussed below) to the authors' preprocessing code to consider it as my own contribution for this section.

As part of their replication package, the authors provided instructions for downloading and pre-processing the datasets. However, they did not document all the necessary packages required for pre-processing the data, resulting in ambiguity about the correct version of these missing packages. As a result, I have used the latest versions of these packages. This, in turn, presented some deprecation issues, for example, *reindex_axis* being changed to *reindex*.

Besides, the train and test files in the UCI Adult dataset contained spaces between columns. The test file was unusable as each row in the target column (income) had a period (50k.). The test file also contained a first line that was not a datapoint. For the LSAC dataset, no explanation was provided on how to convert the original SAS files to CSV files. I wrote the necessary code to overcome all these issues.

For the UCI adult dataset, a split of training data and testing data is already provided. For the other two datasets, I have followed the paper and randomly split them into 70% training data and 30% test data. For the main experiments, the protected features, race and gender, were removed from the dataset because these demographics were not available to the models for training, validation, or testing. During evaluation, each data element was categorized as being a member of one of the following groups: "not black and male", "not black and female", "black and male", "black and female". The group with the fewest members among these is referred to as *minority*. Similar to the paper, I have transformed all categorical attributes using one-hot encoding, and standardized all features vectors to have zero mean and unit standard deviation.

2.3 Hyperparameters

I have followed the paper to use the same experimental setup, data split, and parameter tuning techniques for all the methods. Following the authors' code, I pretrained the learner for ARL for 250 steps, used AdaGrad as the optimizer, and ran each grid search for 5000 training steps, with early stopping if the overall AUC on the validation set had not improved for 10 epochs.

For each approach, I chose the best learning-rate, and batch size by performing a grid search over an exhaustive hyper parameter space given by batch size (32, 64, 128, 256, 512) and learning rate (0.001, 0.01, 0.1, 1, 2, 5). For ARL, I have also searched over the adversary learning rate (0.001, 0.01, 0.1, 1, 2, 5), and for DRO, the parameter η was searched over (0.5, 0.6, 0.7, 0.8, 0.9, 1). These search spaces were taken from the paper. Finally, all the parameters were chosen via 5-fold cross validation by optimizing for best overall AUC.

Table 3 shows the optimal hyperparameters for each dataset and model combination. The grid-search took about 15-16 hours on my computer with 16GB memory and eight CPUs.

3 RESULTS

I report the performance of my own implementation on the same datasets as used by the authors. Based on my own results and the process of obtaining these, I evaluate the replicability of the paper. All results in this section were obtained by training from scratch with the obtained optimal hyperparameters using ten different random seeds, then evaluating on the test set and averaging the results.

I begin this section with the evaluation metrics. Then, I first compare ARL with the DRO model and the Baseline model, following that with comparison against two variants of the IPW model. I have presented these comparisons separately to be in accordance with the paper. Next, I present the difference between my result and the authors'. Finally, in the

Table 3: The hyperparameters with the best overall AUC for each dataset and training method.

Dataset	Method	Batch size	Primary learning rate	Adversary learning rate	η
Adult	Baseline	64	0.1	-	-
Adult	ARL	512	0.01	0.01	-
Adult	DRO	512	0.1	-	0.5
Adult	IPW(S)	128	0.01	-	-
Adult	IPW(S+Y)	128	0.1	-	-
LSAC	Baseline	256	0.1	-	-
LSAC	ARL	256	0.1	0.001	-
LSAC	DRO	64	0.1	-	0.5
LSAC	IPW(S)	256	0.1	-	-
LSAC	IPW(S+Y)	64	0.01	-	-
COMPAS	Baseline	32	0.01	-	-
COMPAS	ARL	32	0.01	0.001	-
COMPAS	DRO	64	0.01	-	0.6
COMPAS	IPW(S)	64	0.01	-	-
COMPAS	IPW(S+Y)	64	0.01	-	-

last two sub-sections, I discuss computational-identifiability and the adversary outputs, in order to verify the authors' insights about the ARL.

3.1 Evaluation Metrics

Similar to the paper, I have used AUC as the utility metric as it is robust to class imbalance, yielding an unbiased metric of performance. Using accuracy as an evaluation metric would present a bias towards the largest groups in the data, since correctly predicting for the largest groups would contribute most to the overall performance. In contrast, it is not easy to receive high AUC for trivial predictions. To evaluate fairness, I stratified the test data by groups (see section 2.2.2), computed AUC per protected group, and reported the following:

- AUC(avg): micro-average over all protected group AUCs,
- AUC(macro-avg): macro-average over all protected group AUCs,
- AUC(min): minimum AUC over all protected groups, and
- AUC(minority): AUC reported for the smallest protected group in the dataset

As mentioned in section 2.2.2, the protected features are removed from the dataset, and are only used to compute subgroup AUC in order to evaluate fairness.

3.2 ARL vs DRO vs Baseline

The findings of comparing ARL to the DRO model and the baseline model are presented in Table 4. These results are based on average AUC across ten runs. For each dataset and metric, we can see that all approaches get relatively comparable results. Many of the changes between these methods are insignificant, in other words, there is no clear winner.

3.3 ARL vs IPW

In Table 5, I compare the results of ARL to two variants of the naive reweighting strategy IPW, i) IPW(S): uses only protected groups to compute the inverse probability weights, and ii) IPW(S+Y): uses protected groups and labels. On any of the datasets, ARL does not outperform other approaches, similar to the results in Table 4.

3.4 My Results vs Results in the Paper

Table 6 shows the difference in percentage points between my results and those reported in the paper. Positive numbers mean that I achieved better results while negative means my results were worse. The negative numbers are shown in red. For the UCI Adult and the LSAC datasets, all the methods perform better in my replication compared to the paper. However, for ARL (shown in gray), this increase is the minimum, almost negligible. On the other hand, DRO performs significantly better in my replication than shown in the paper, specially for the LSAC dataset (shown in bold text). This explains why in my results, ARL does not have an advantage like it does in the paper: ARL does not necessarily perform any worse, (except partially on the COMPAS data) but the other methods perform much better. Figure 1 and Figure 2 also show these results using charts. I further explain these differences in section 4.

3.5 Computational Identifiability

The authors hypothesized that unobserved protected groups are correlated with observed features and class label. Therefore, although they are unobserved, they can be computationally-identifiable. In order to test this hypothesis, they trained a linear model to predict the protected features (race and sex) based on the other features and target values. The authors did not share any additional information about their training process. For this task, I only perform a single run because

Table 4: Main results: ARL vs DRO vs Baseline

Dataset	Method	AUC avg	AUC macro-avg	AUC min	AUC minority
Adult	Baseline	**0.9103**	**0.9198**	**0.8857**	0.9445
Adult	DRO	0.9096	0.9190	0.8849	**0.9446**
Adult	ARL	0.9101	0.9193	0.8855	0.9435
LSAC	Baseline	**0.8317**	**0.8253**	**0.8080**	0.8360
LSAC	DRO	0.8279	0.8236	0.8058	0.8344
LSAC	ARL	0.8266	0.8202	0.8051	**0.8364**
COMPAS	Baseline	0.7337	**0.7317**	0.6958	0.7409
COMPAS	DRO	**0.7338**	0.7312	**0.6976**	0.7422
COMPAS	ARL	0.7330	0.7312	0.6956	**0.7445**

Table 5: ARL vs Inverse Probability Weight

Dataset	Method	AUC avg	AUC macro-avg	AUC min	AUC minority
Adult	IPW(S)	0.9080	0.9176	0.8819	0.9425
Adult	IPW(S+Y)	0.9100	**0.9196**	0.8848	**0.9448**
Adult	ARL	**0.9101**	0.9193	**0.8855**	0.9435
LSAC	IPW(S)	0.8150	0.8083	0.7918	0.8160
LSAC	IPW(S+Y)	**0.8373**	**0.8321**	**0.8149**	**0.8420**
LSAC	ARL	0.8266	0.8202	0.8051	0.8364
COMPAS	IPW(S)	0.7264	0.7245	0.6877	0.7324
COMPAS	IPW(S+Y)	0.7310	0.7285	0.6942	0.7374
COMPAS	ARL	**0.7330**	**0.7312**	**0.6956**	**0.7445**

Table 6: Difference (percentage point) between my results and those reported in the paper.

Dataset	Method	AUC avg	AUC macro-avg	AUC min	AUC minority
Adult	Baseline	1.23	2.88	1.87	6.95
Adult	DRO	1.06	1.10	1.59	1.16
Adult	IPW(S)	1.10	2.56	0.59	5.95
Adult	IPW(S+Y)	1.30	1.06	0.78	1.28
Adult	ARL	0.31	0.43	0.45	0.14
LSAC	Baseline	1.87	1.23	1.80	1.20
LSAC	DRO	**11.89**	**11.36**	**12.28**	**10.54**
LSAC	IPW(S)	2.1	1.93	1.98	4.10
LSAC	IPW(S+Y)	3.83	3.41	3.09	5.70
LSAC	ARL	0.36	0.02	0.71	0.44
COMPAS	Baseline	-1.43	0.17	2.18	-3.31
COMPAS	DRO	3.48	5.32	8.16	3.82
COMPAS	IPW(S)	-1.76	-0.25	0.87	-2.66
COMPAS	IPW(S+Y)	0.40	0.45	1.62	-2.66
COMPAS	ARL	-1.00	0.42	3.76	-4.05

the goal is simply to test if the adversary is capable of accomplishing this task, obtaining a precise accuracy is not the focus.

The best hyperparameters discovered by the grid search in section 2.3 are used for training, validation, and testing. Table 7 displays the test accuracies for predicting race and sex. The findings are mostly in line with those of the original study.

However, that is not the whole picture. Although this was not shown in the paper, per-group accuracies from the table (also see Figure 3) illustrate that the linear model gets its performance by focusing on the majority class, or even just predicting the majority class all of the time. This casts doubt on the adversary's ability to accurately identify and prioritise these protected groups.

Table 7: Identifying groups using ARL

	Group	Adult	LSAC	COMPAS
Race	Total	0.904	0.921	0.604
	Not black	0.999	0.982	0.428
	Black	0.001	0.075	0.762
Sex	Total	0.794	0.557	0.802
	Male	0.957	0.877	1.0
	Female	0.327	0.141	0.0

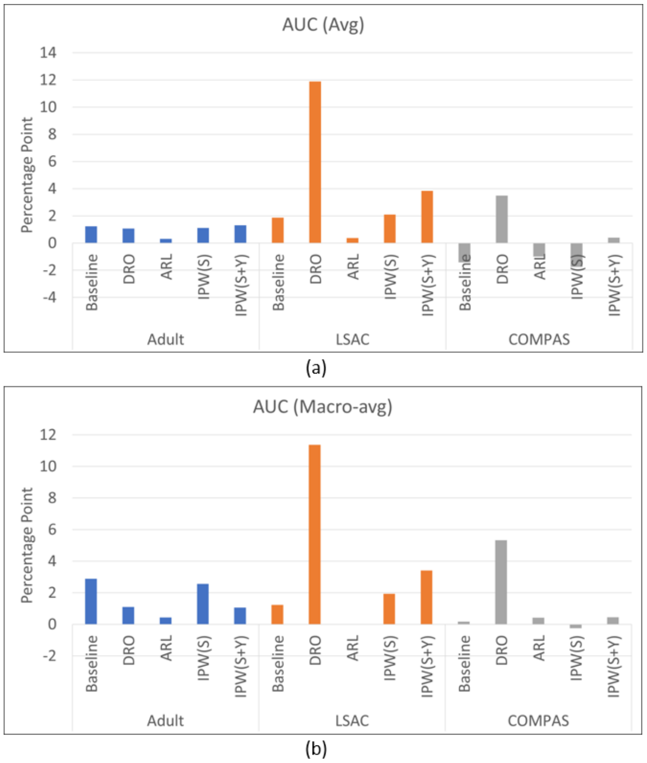

Figure 1: Percentage Point Difference in (a) Avg AUC and (b) Macro-avg AUC between my results and those reported in the paper

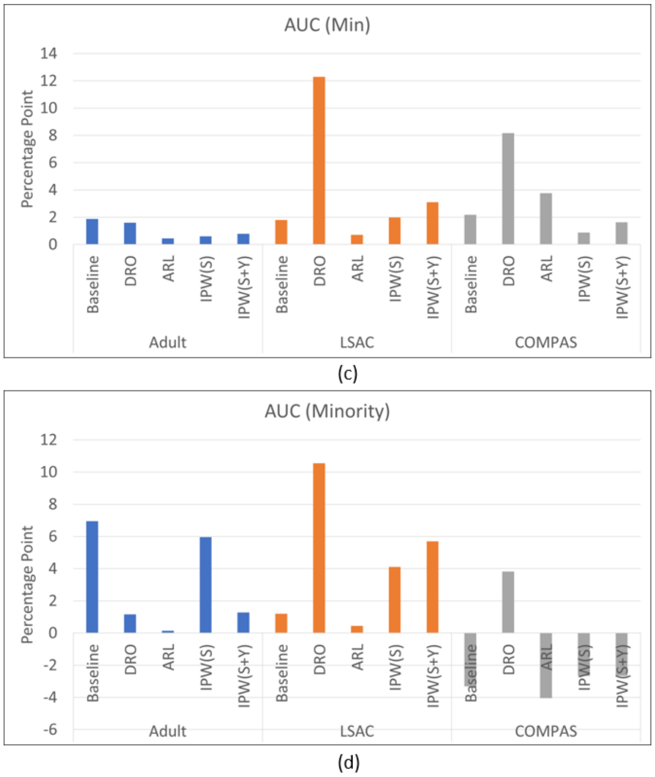

Figure 2: Percentage Point Difference in (c) Minimum AUC and (d) Minority AUC between my results and those reported in the paper

3.6 Does the adversary learn meaningful weights?

The authors plotted the weights returned by the adversary on the UCI Adult training set to see if the weights learnt by ARL are meaningful. I ran the same task, but on the test set of the UCI Adult data. On the test weights, I performed KDE with a bandwidth parameter of 0.3 to generate continuous distributions across λ. Figure 4 shows weights assigned by ARL into four quadrants of a confusion matrix. Each subplot displays the weight on x-axis and the density on y-axis.

The computed densities are rather close to the authors' findings. The UCI Adult dataset has a significant class imbalance, and as we can see, the adversary gives the less common class higher weights even when there is no error (Figure 4 bottom-right). The plots for the misclassifications cases (top-right, bottom-left) show larger weights as well. As a result, I was able to back up the authors' assertion that ARL learns to give more weight to underrepresented classes.

However, there are two things to note here. First, although the misclassified examples are getting larger weights, they are more significant for the Male groups, and less for the Female groups. The adversary is clearly improving fairness based on Race, but not so much on Gender. This is different from the paper. Secondly, there is no clear pattern among the curves of different colors (which denote different groups). This implies that ARL does not give underrepresented groups a larger weighting. Although this is consistent with the findings

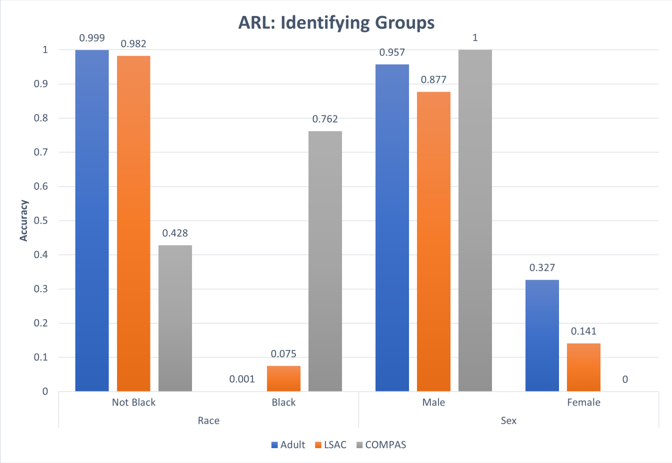

Figure 3: The Adversary is focusing on the majority class. For the minority class, the accuracy gets as little as zero.

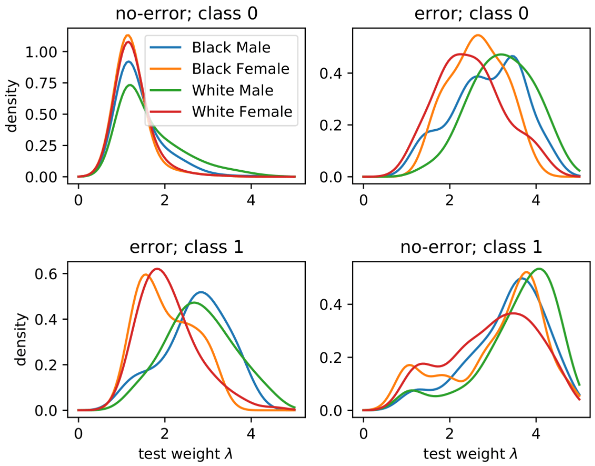

Figure 4: Example weights learnt by ARL on the Adult test set

of the paper, it contradicts authors' presentation of ARL's mechanism.

4 DISCUSSION

I was unable to recreate the advantage of ARL over the comparative models, as indicated in the previous section. This is not to say that ARL did worse in my replication. Actually, I produced better numbers for the ARL. At the same time, the other models in my replication did significantly better than what was reported in the paper. As a result, the advantage of ARL over alternative approaches was nullified.

The authors employed a fixed number of training steps, but I used early stopping. It is possible that the variations between the authors' and my findings are due to different comparison points. I tried a new grid search with a maximum number of 10k training steps instead of 5k because some training runs did not end owing to early stopping but rather due to reaching the maximum number of training steps. Due to resource constraints (the program ran for about thirty hours before failing), I could not finish that grid search. For this reason, I could not rule out that training for too few training steps skewed my results.

A potential reason why ARL does not outperform a simple risk-minimizing baseline is that ARL apparently does not upweight minorities, as discussed in section 3.5 (Figure 4). This is entirely consistent with the figures in the original paper but contradicts with the motivation for ARL. Perhaps the linear adversary is too weak to be effective. The identifiability results in Table 7 suggest that it mostly learns to exploit the class imbalance. I also want to point out that the datasets chosen by the authors might be too simple to differentiate between the various methods, because the baseline model is already quite fair to subgroups, leaving little room for ARL to show any distinct advantage. In fact, the advantage of ARL reported in the original paper is not all that significant. This supports the idea that the discrepancy between my results and the original results is due to minor differences in training procedure, such as the use of early stopping.

Finally, there were some concerns with dataset preparation and minor inconsistencies between the paper's description of the algorithms and their actual implementation in the code, but none of these issues posed a serious barrier to replicating the paper.

4.1 Completeness

I mentioned the following tasks in my project proposal:

(1) Preprocessing datasets following the steps mentioned in the paper - Completed (section 2.2.2).
(2) Implementing the proposed ARL approach - Completed (section 2.1.1).
(3) Implementing two baseline and one state-of-the-art approach to compare against ARL - Completed (section 2.1) with one exception. The authors used two variants of DRO, i) with η tuned as a hyperparameter, ii) with η tuned as detailed in the original paper. For my replication, I have only done the former, using the η I got from my grid-search.
(4) Evaluating and comparing ARL and the other approaches on the datasets - Completed (section 3.2, 3.3, 3.4).
(5) Understanding ARL, presenting insights into the inner working of ARL- Completed (section 3.5, 3.6)

5 CONCLUSION

In this replication study, I found no advantage of ARL over baseline models. Because the results of the original research favour ARL just marginally and that the adversary does not appear to upweight minorities as much, I conclude that ARL in its current form cannot increase fairness on these datasets. However, because the theoretical motivation appears to be sound, it could be promising to use ARL to datasets with more severe fairness issues, or to further investigate whether the adversary's capability can be changed to increase performance.

5.1 Source Code

All the source code for this replication study is available at https://github.com/golamSaroar/replication_fairness_without_demographics

REFERENCES

[1] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, Vol. 81. PMLR.

[2] Dheeru Dua and Casey Graff. 2017. UCI Adult Data. https://archive.ics.uci.edu/ml/datasets/adult.

[3] Romana Hasnain-Wynia, David W. Baker, David Nerenz, Joe Feinglass, Anne C. Beal, Mary Beth Landrum, Raj Behal, and Joel S. Weissman. 2007. Disparities in health care are driven by where minority patients seek care: Examination of the hospital quality alliance measures. *Archives of internal medicine (Chicago, Ill. : 1908)* 167 (2007).

[4] Preethi Lahoti. 2020. Tensorflow Implementation by The Authors. https://github.com/google-research/google-research/tree/master/group_agnostic_fairness.

[5] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi. 2020. Fairness without Demographics through Adversarially Reweighted Learning. *CoRR* abs/2006.13114 (2020).

[6] ProPublica. 2016. COMPAS Recidivism Risk Score Data and Analysis. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[7] Thashim. 2018. DRO Implementation based on the original paper. https://worksheets.codalab.org/worksheets/0x17a501d37bbe49279b0c70ae10813f4c/.

[8] Linda F. Wightman. 1998. LSAC National Longitudinal Bar Passage Study. http://www.seaphe.org/databases.php.