

# Adversarial Training Negatively Affects Fairness

Korn Sooksatra \* and Pablo Rivas <sup>†</sup>, *Senior, IEEE*  
*School of Engineering and Computer Science*  
*Department of Computer Science, Baylor University*  
Email: {\*Korn\_Sooksatra1,<sup>†</sup>Pablo\_Rivas}@Baylor.edu

**Abstract**—With the increasing presence of deep learning models, many applications have had significant improvements; however, they face a new vulnerability known as adversarial examples. Adversarial examples can mislead deep learning models to predict the wrong classes without human actors noticing. Recently, many works have tried to improve adversarial examples to make them stronger and more effective. However, although some researchers have invented mechanisms to defend deep learning models against adversarial examples, those mechanisms may negatively affect different measures of fairness, which are critical in practice. This work mathematically defines four fairness scores to show that training adversarially robust models can harm fairness scores. Furthermore, we empirically show that adversarial training, one of the most potent defensive mechanisms against adversarial examples, can harm them.

**Index Terms**—adversarial example, fairness, adversarial training, deep learning

## I. INTRODUCTION

In these years, many applications (e.g., autonomous cars [10], language translation [8] and recommendation systems [20]) have been improved due to the rise of deep learning. Nevertheless, deep learning has a crucial weakness found in [9], [21]. Such weakness is that an adversary adds a small perturbation that is not perceptible by humans to a sample to mislead a deep learning model to predict a wrong class. That generated sample is called an adversarial example, and recently, there have been several attempts [5], [6], [9], [14]–[18], [21] to generate adversarial examples that are difficult to be noticed by humans and effective. In the meantime, there have been some defensive mechanisms [7], [9], [14], [19], [23], [24] discovered to defend against those attacks, and a classifier that is strong against adversarial examples is called an adversarially robust classifier.

However, the works in [2], [22] showed that to make a classifier adversarially robust, we need to sacrifice class-wise accuracy fairness defined in Section III. Moreover, the work in [3] also showed that a classifier trained by adversarial training [14] achieved more leave-one-out unfairness than a typically trained classifier. Hence, the main contribution throughout this work is to demonstrate that adversarial training is linked to problems with class-wise accuracy fairness and leave-one-out fairness and evaluate whether some other definitions of fairness are also harmed by adversarial training.

This paper is organized as follows; Section II explains some frequently-used notations, briefly describes adversarial training, and refers to some related work; Section III provides definitions of fairness used in the experiments; Section IV describes how

we performed the experiments and shows and discusses the experimental results; Section V concludes everything we have done in this work and mentions works that we may involve in the future.

## II. BACKGROUND

### A. Notation

A neural network or classifier is a function  $F(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}^m$  where  $\mathbf{x}$  is an input,  $d$  is a number of attributes of  $x$  and  $m$  is a number of classes. We denote  $\mathcal{X}$  as a set of input and  $\mathcal{Y}$  as a set of classes; thus,  $\mathbf{x} \in \mathcal{X}$ , and  $|\mathcal{Y}| = m$ . Further, a set of data in dataset  $\mathcal{D}$  belonging to class  $y$  is denoted by  $\mathcal{D}_y$  where  $\mathcal{D} = \{(\mathbf{x}, y) | \mathbf{x} \in \mathcal{X}, y \in \mathcal{Y}\}$ .

### B. Adversarial training

This method was first proposed in [9] and shown that it is a very effective defensive mechanism against adversarial examples in [14]. This idea started by an attempt to solve this problem, given classifier  $F$  and dataset  $\mathcal{D}$ ,

$$E_{(\mathbf{x}, y) \sim \mathcal{D}} [\min_F \max_{L^P(\boldsymbol{\delta}) \leq \epsilon} l(F, \mathbf{x} + \boldsymbol{\delta}, y)]$$

where  $\mathbf{x}$  is a clean input,  $y \in \mathcal{Y}$ ,  $l(F, \mathbf{x}, y)$  is a loss function with respect to  $F(\mathbf{x})$  and  $y$ ,  $L^P(\boldsymbol{\delta}) = \sqrt[P]{\sum_{i=0}^{d-1} \delta_i^P}$ ,  $\delta_i$  is element  $i$  of  $\boldsymbol{\delta}$ ,  $d$  is the number of dimensions of  $\mathbf{x}$  and  $\epsilon$  is a bound of added perturbation. Throughout this work, we use  $L^\infty$  for generating adversarial examples.

However, the inner maximization is nontrivial to solve; therefore, the method in [14] generated adversarial examples by using Projected gradient descent (PGD) with small bound  $\epsilon$  from a training dataset and retrained a trained learning model with these adversarial examples.

### C. Related works

In 2021, Benz *et. al* [2] showed that robustness could ruin class-wise accuracy fairness by demonstrating their empirical results. Also, Xu *et al.* [22] had pointed the same problem and proposed a solution that used a fairness regularization technique to find the tradeoff between robustness and class-wise accuracy fairness.

According to the previous works, to the best of our knowledge, there has been no work focusing on the tradeoff between robustness and several definitions of fairness; hence, our work is the first one.

### III. FAIRNESS DEFINITIONS

This section describes some definitions of fairness used in our experiments for indicating the tradeoff between these definitions of fairness and robustness achieved by adversarial training. For simplicity to formally give definitions of fairness, we provide definitions of unfairness that contrasts with fairness. The first definition of unfairness that we would like to introduce is class-wise accuracy unfairness formally defined in Definition 1. This unfairness indicates that a particular classifier is unfairly accurate and was shown to be worse in a robust model than in a standard model in [2], [22]. If this unfairness is very high, classifier  $F$  is not class-wise accuracy fair at all. In the contrary, if this unfairness is very low, this classifier is very class-wise accuracy fair.

**Definition 1** (Class-wise accuracy unfairness). Given classifier  $F$  and dataset  $\mathcal{D}$ , class-wise accuracy unfairness can be measure as

$$\text{CAU}(F, \mathcal{D}) = \max_{y, y' \in \mathcal{Y}} |\text{Acc}(F, \mathcal{D}_y) - \text{Acc}(F, \mathcal{D}_{y'})|$$

where  $\text{Acc}(F, \mathcal{D}) \in [0, 1]$  denotes the accuracy score of  $F$  with respect to dataset  $\mathcal{D}$ .

Next, in addition to CAU, we desire to check if a robust model would unfairly improve the robustness of data belonging to some classes. Then, we give a formal definition of this kind of unfairness in Definition 2. If this unfairness is very high, classifier  $F$  is not class-wise robustly fair at all. On the other hand, if this value is low, this classifier is class-wise robustly fair.

**Definition 2** (Class-wise robustness unfairness). Given classifier  $F$  and dataset  $\mathcal{D}$ , class-wise robustness unfairness can be measure as

$$\text{CRU}_\epsilon(F, \mathcal{D}) = \max_{y, y' \in \mathcal{Y}} |\text{LE}(F, \mathcal{D}_y) - \text{LE}(F, \mathcal{D}_{y'})|$$

where  $\text{LE}(F, \mathcal{D})$  denotes the average least required  $\epsilon$ , which is a bound of added perturbation, to find adversarial examples for dataset  $\mathcal{D}$  to mislead classifier  $F$ . Also, we can measure it in term of success rate of generating adversarial examples as

$$\text{CRU}_s(F, \mathcal{D}, \epsilon) = \max_{y, y' \in \mathcal{Y}} |\text{SR}(F, \mathcal{D}_y, \epsilon) - \text{SR}(F, \mathcal{D}_{y'}, \epsilon)|$$

where  $\text{SR}(F, \mathcal{D}, \epsilon)$  is a success rate of adversarial attack with bound  $\epsilon$  on classifier  $F$  with respect to dataset  $\mathcal{D}$ .

Furthermore, Black and Fendrickson [3] proposed a new definition of unfairness called leave-one-out unfairness formally defined in Definition 3. The main idea of this unfairness is that when removing one random sample from the training dataset, after training a classifier, its output for an arbitrary input should not be much different from the one trained with the entire training dataset. In other words, if this unfairness is high, some samples in the training dataset tremendously influence the classifier.

**Definition 3** (Leave-one-out unfairness). Given classifier  $F$ , dataset  $\mathcal{D}$  and sample  $\mathbf{x}$ , leave-one-out unfairness can be measure as

$$\text{LUF}(F, \mathcal{D}, \mathbf{x}) = \max_{(i, \cdot) \in \mathcal{D}^L, y \in \mathcal{Y}} |F_{\mathcal{D}}(\mathbf{x})_y - F_{\mathcal{D} \setminus \{i\}}(\mathbf{x})_y|$$

where  $\mathcal{D}^L$  is a leave-out set sampled from dataset  $\mathcal{D}$ ,  $F_{\mathcal{D}}(\mathbf{x})$  is classifier  $F$  trained with dataset  $\mathcal{D}$ ,  $\mathcal{D} \setminus \{i\}$  is dataset  $\mathcal{D}$  excluding sample  $i$  and  $F(\mathbf{x})_y$  is the confidence of input  $\mathbf{x}$  to belong to class  $y$ . Further, we can compute the expected value of LUF over dataset  $\mathcal{D}$  as

$$\overline{\text{LUF}}(F, \mathcal{D}) = \mathbb{E}_{(\mathbf{x}, \cdot) \in \mathcal{D}} [\text{LUF}(F, \mathcal{D}, \mathbf{x})].$$

At last, we would like to describe equalized odd unfairness introduced in [11] and formally defined in Definition 4. This unfairness implies that an unprotected attribute has much influence on the prediction of a classifier while it is irrelevant. For example, for a job classifier, ethnicity should be irrelevant to the output, and this attribute is called an unprotected attribute. However, if the classifier considers ethnicity to predict a job, it is unfair in this definition of unfairness.

**Definition 4** (Equalized odd unfairness). Given classifier  $F$  and dataset  $\mathcal{D}$ , this unfairness can be measured as

$$\text{EOU}(F, \mathcal{D}, a, y) = \max_{i, j \in V_a} |\mathbb{E}_{\mathbf{x} \in \mathcal{D}^{a=i}} [F(\mathbf{x})_y] - \mathbb{E}_{\mathbf{x} \in \mathcal{D}^{a=j}} [F(\mathbf{x})_y]|$$

where  $a$  is an protected attribute,  $y \in \mathcal{Y}$ ,  $V_a$  is a set of values of attribute  $a$ ,  $\mathcal{D}^{a=i}$  is dataset  $\mathcal{D}$  in which values of attribute  $a$  of all samples are  $i$ . In the other words, EOU is the maximum of the differences between the expected values of confidence of class  $y$  of classifier  $F$  with respect to dataset  $\mathcal{D}$  with different values of protected attribute  $a$ .

### IV. EXPERIMENTS AND RESULTS

#### A. Setup

We experimented on fairness and robustness classifiers with Python version 3 and Tensorflow module [1] on Google Colab. We used two datasets on these experiments and also implemented classifiers for them. Those datasets and classifiers are described as follows:

- 1) **CIFAR-10 dataset [13]**. This dataset contains 32x32 color images belonging to 10 classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship and truck. It has balanced 50000 training samples and balanced 10000 test samples. Then, we implemented a CIFAR-10 classifier consisting of two convolutional layers followed by max poolings and *ReLU* activation functions, two hidden dense layers followed by *ReLU* activation functions and one output layer followed by a *softmax* activation function for computing confidences of all classes. We trained this classifier with the training samples by using Adam optimizer [12], 250 batch size and 20 epochs. As a result, the classifier achieved 71.96% accuracy on the test samples. Further, we created adversarial examples on the test samples by using PGD [14] with  $\epsilon = 0.01$ , and the classifier achieved 4.24% accuracy. We call accuracy resulting from clean dataset "standard accuracy" and call accuracy resulting from adversarial examples "robust accuracy". After that, we merged the training samples with the adversarial examples generated from the training samples and trained another classifier with those merged

samples by using adversarial training [14] to obtain a robust classifier. Consequently, the robust classifier can achieve 72.43% standard accuracy and 83.12% robust accuracy.

- 2) **Adult dataset [4].** This dataset has been used to predict whether the income of a person exceeds \$50K per year by considering 14 attributes. These attributes consist of 6 continuous attributes and 8 discrete attributes. Moreover, the dataset has 30162 training samples and 15060 test samples. Then, we implemented a classifier for this dataset composed of four hidden dense layers followed by *tanh* activation functions and an output layer followed by a *softmax* activation function. Next, because the number of class-1 samples (i.e., samples that have income exceeding \$50K per year) is much less than the number of class-0 samples (i.e., samples that have income less than \$50K per year), we added the same class-1 samples to the training dataset. Then, we trained the classifier with this augmented training dataset using the same setting as for the CIFAR-10 dataset, and it achieved 79.53% standard accuracy on the test dataset. After that, we generated adversarial examples from the test samples by PGD [14] with  $\epsilon = 10$ , and the classifier can achieve 14.3% robust accuracy on these adversarial examples. Further, we created adversarial examples on the training samples and merged them with the training samples. Then, we trained another classifier with the merged samples by using adversarial training [14], and this robust classifier can achieve 79.37% standard accuracy and 96.37% robust accuracy.

Our experiments were constructed according to the definitions of unfairness explained in Section III. First, we start with the experiment of class-wise accuracy unfairness on CIFAR-10 and Adult classifiers and move to the experiment of class-wise robustness unfairness on CIFAR-10 and Adult classifiers. Then, we use CIFAR-10 classifiers for the experiment of leave-one-out unfairness and Adult classifiers for equalized odd unfairness.

### B. Class-wise accuracy unfairness

We evaluated the standard CIFAR-10 classifier (i.e., a classifier trained with clean CIFAR-10 samples) and the robust one (i.e., a classifier trained by adversarial training) on the CIFAR-10 test samples. The result of this evaluation is shown in Fig. 1, and noticeably, the robust classifier had less accuracy than the standard classifier for samples of hard classes (e.g., cat and dog). On the contrary, the robust classifier achieved higher accuracy than the standard one for samples of easy classes (e.g., airplane, horse and ship). This phenomenon led the robust classifier to have a larger gap between the maximum class-wise accuracy and the minimum class-wise accuracy than the standard one; therefore, the robust classifier achieved higher CAU than the standard classifier. Specifically, the robust classifier had 0.344 CAU while the standard one had 0.286 CAU. Thus, the robust CIFAR-10 classifier trained by adversarial training was more class-wise accuracy unfair than the standard CIFAR-10 classifier.

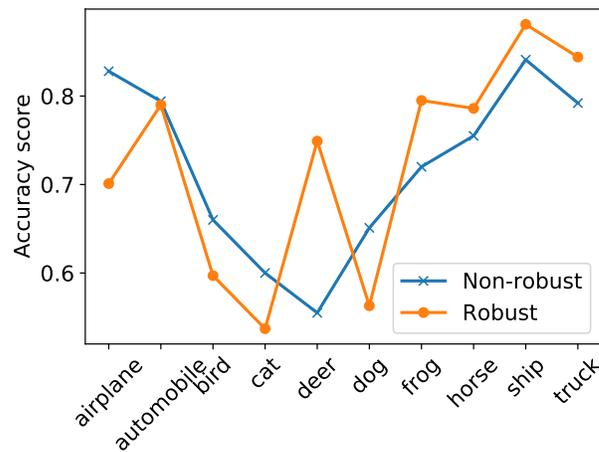


Fig. 1: Accuracy score of standard and robust classifiers with samples belonging to each class on the CIFAR-10 dataset.

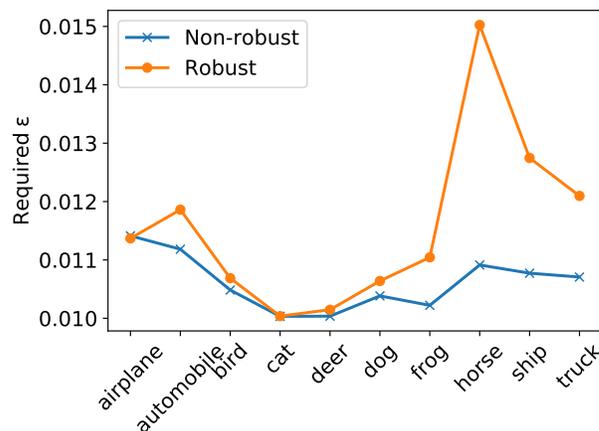


Fig. 2: Average required  $\epsilon$  of standard and robust classifiers with samples belonging to each class on CIFAR-10 dataset to generate adversarial examples.

TABLE I:  $CRU_s$  of the standard and robust CIFAR-10 classifiers with different values of  $\epsilon$ .

Model	$\epsilon = 0.01$	$\epsilon = 0.012$	$\epsilon = 0.014$
Standard	13.43	8.45	4.71
Robust	25.85	16.27	10.56

Furthermore, we have performed the same experiment on the Adult classifiers, and since there are only two classes in the Adult dataset, it is easy to compute CAU for the classifiers. As a result, the robust Adult classifier achieved 0.0205 CAU, higher than 0.0195 CAU of the standard Adult classifier. Therefore, the robust classifier is more class-wise accuracy unfair than the standard classifier. We can summarize all these experiments that applying adversarial training to create a robust classifier can ruin class-wise accuracy fairness.

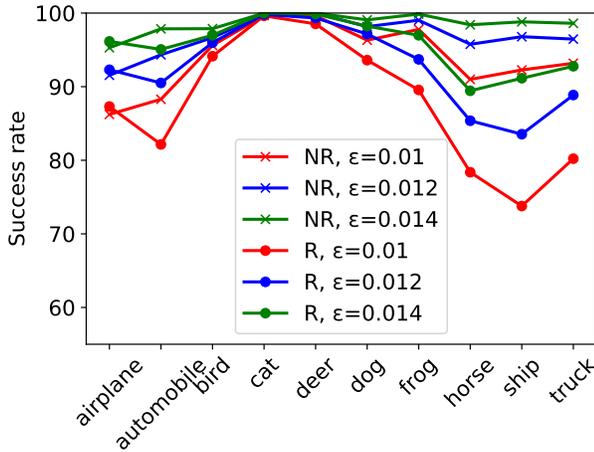


Fig. 3: Success rate of generating adversarial examples with different  $\epsilon$  of standard and robust classifiers with samples belonging to each class on CIFAR-10 dataset where NR is the standard classifier and R is the robust classifier.

TABLE II:  $CRU_s$  of the standard and robust Adult classifiers with different values of  $\epsilon$ .

Model	$\epsilon = 10$	$\epsilon = 15$	$\epsilon = 20$
Standard	6.11	10.81	10.7
Robust	18.66	23.38	22.72

### C. Class-wise robustness unfairness

Note that the adversarial attack used in this experiment was PGD [14]. For each sample of the CIFAR-10 test dataset, we increased  $\epsilon$  until we found an adversarial example of that sample, and we computed the average required  $\epsilon$  over samples of each class as seen in Fig. 2. Intuitively, samples in classes that require higher  $\epsilon$  to generate adversarial examples are more robust than those in classes requiring less  $\epsilon$ . Thus, the robustness of samples in hard classes (e.g., cat and dog) was slightly or not improved at all by adversarial training. In contrast, the samples in easy classes (e.g., horse and ship) were significantly improved by adversarial training. Implicitly, the robust CIFAR-10 classifier trained by adversarial training achieved higher  $CRU_\epsilon$ , which was 0.005, than the one achieved by the standard CIFAR-10 classifier, which was 0.0014. In addition, we also experimented on a success rate of generating adversarial examples on samples in each class with different  $\epsilon$  as seen in Fig. 3. The implication was the same as what we explained for Fig. 2, and  $CRU_s$  for each setting is shown in Table I. Moreover, Fig. 3 and Table I implied that the difference between  $CRU_s$  obtained from the robust classifier and the one obtained from the standard classifier was lower when  $\epsilon$  was higher. This implication was because the adversarial attack was more effective when  $\epsilon$  was higher. Then, samples in easy classes were much more vulnerable while samples in hard classes were slightly more vulnerable. Therefore, the gap between  $CRU_s$  of the classifiers was smaller.

In addition to CIFAR-10 classifiers, we experimented on a

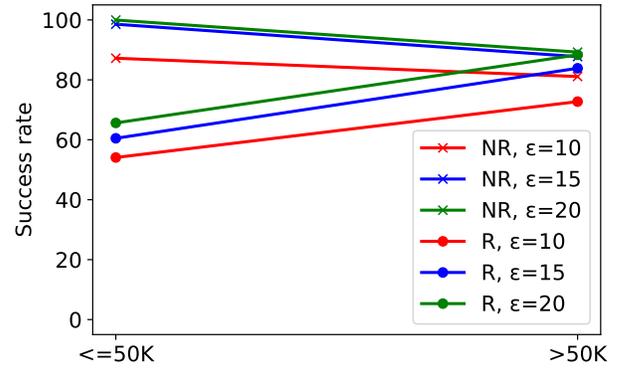


Fig. 4: Success rate of generating adversarial examples with different  $\epsilon$  of standard and robust classifiers with samples belonging to each class on Adult dataset where NR is the standard classifier and R is the robust classifier.

TABLE III:  $\overline{LUF}$ ,  $\overline{LUF}_t$  and  $\overline{LUF}_l$  of the standard and robust CIFAR-10 classifiers.

Model	$\overline{LUF}$	$\overline{LUF}_t$	$\overline{LUF}_l$
Standard	0.7694	29.26	2926.1
Robust	0.7945	29.35	2934.89

success rate of generating adversarial examples on samples in each class of Adult test samples by using the Adult classifiers as shown in Fig. 4. We found that the robust classifier achieved more  $CRU_s$  than the one achieved by the standard classifier as shown in Fig. 4 and Table II. Nonetheless, we did not see any pattern of  $CRU_s$  in different settings of  $\epsilon$  as we found for CIFAR-10 classifiers. The reason could be that the training and test samples were unbalanced; hence, when we applied adversarial training on the dataset, a class with much more samples could be much more robust than a class with much fewer samples. In the end, we summarized that a robust classifier trained by adversarial training could be more class-wise robustly unfair than a standard classifier.

### D. Leave-one-out unfairness

In this experiment, we used CIFAR-10 dataset and the particular classifiers and picked 100 samples from the training samples as leave-out samples. Hence,  $|\mathcal{D}^L| = 100$ , and for a classifier, we had additional 100 leave-out classifiers each of which was trained with the training samples excluding one different leave-out sample. From the 100 leave-out classifiers, we could compute  $\overline{LUF}$  of the original classifier. Further, for test samples, we also computed the average number of leave-out classifiers that resulted in different classes to the original one and could formally define it as

$$\overline{LUF}_t(F, \mathcal{D}) = \mathbb{E}_{(\mathbf{x}, \cdot) \in \mathcal{D}} \sum_{(i, \cdot) \in \mathcal{D}^L} 1\{C_{\mathcal{D}}(\mathbf{x}) \neq C_{\mathcal{D} \setminus \{i\}}(\mathbf{x})\}$$

where  $1\{T\}$  is 1 if  $T$  is true and 0 otherwise, and

$$C_{\mathcal{D}}(\mathbf{x}) = \operatorname{argmax}_{y \in \mathcal{Y}} F_{\mathcal{D}}(\mathbf{x})_y.$$

TABLE IV: Average confidences of the Adult classifiers to predict that incomes exceed \$50K per year where *gender* attribute is protected.

Model	$\mathcal{D}_0$		$\mathcal{D}_1$	
	Male	Female	Male	Female
Standard	0.272	0.198	0.717	0.626
Robust	0.248	0.168	0.762	0.668

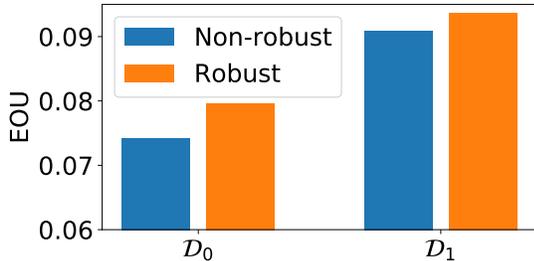


Fig. 5: EOU achieved by the standard and robust Adult classifiers on  $\mathcal{D}_0$  and  $\mathcal{D}_1$  where *gender* attribute is protected.

Additionally, for the leave-out classifiers, we computed the average number of the test samples on which the leave-out classifiers resulted in different classes to the original classifier, and this could be formally defined as

$$\overline{\text{LUF}}_l(F, \mathcal{D}) = \mathbb{E}_{(\mathbf{i}, \cdot) \in \mathcal{D}^L} \sum_{(\mathbf{x}, \cdot) \in \mathcal{D}} 1\{C_{\mathcal{D}}(\mathbf{x}) \neq C_{\mathcal{D}(\setminus \mathbf{i})}(\mathbf{x})\}.$$

We showed those values above of our standard and robust CIFAR-10 classifiers in Table III. Explicitly, the robust classifier achieved higher values of  $\overline{\text{LUF}}$ ,  $\overline{\text{LUF}}_t$  and  $\overline{\text{LUF}}_l$  than the standard classifier. Thus, a robust classifier trained by adversarial training is more leave-one-out unfair than the standard one.

#### E. Equalized odd unfairness

We only used the Adult dataset and its classifiers in this experiment since the CIFAR-10 dataset did not have any protected attribute. First, we considered *gender* attribute as a protected attribute because people’s genders were not supposed to be considered when predicting whether those people had incomes exceeding \$50K per year. This attribute had only two values: male and female. We divided the test samples into two sets: a set of samples with incomes less than \$50K per year ( $\mathcal{D}_0$ ) and a set of samples with incomes exceeding \$50K per year ( $\mathcal{D}_1$ ). The average confidences of the Adult classifiers to predict that incomes exceeded \$50K per year in  $\mathcal{D}_0$  and  $\mathcal{D}_1$  are shown in Table IV, and the confidences resulted from female samples were less than the ones resulted from male samples. Implicitly, the classifiers were biased. Plus, we computed EOU from the information in Table IV and summarized the results in Fig. 5. Noticeably, EOU in  $\mathcal{D}_1$  was higher than EOU in  $\mathcal{D}_0$  since samples in  $\mathcal{D}_1$  were more biased than samples in  $\mathcal{D}_0$ . Furthermore, EOU achieved by the robust classifier was slightly higher than EOU achieved by the standard classifier in both  $\mathcal{D}_0$  and  $\mathcal{D}_1$ .

In addition, we experimented on the Adult dataset when *race* attribute was protected, and the average confidences of

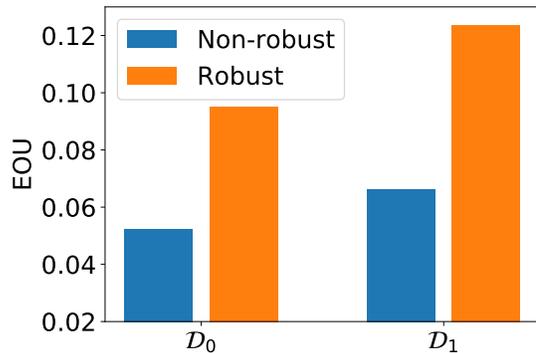


Fig. 6: EOU achieved by the standard and robust Adult classifiers on  $\mathcal{D}_0$  and  $\mathcal{D}_1$  where *race* attribute is protected.

the classifiers to predict that incomes exceeded \$50K per year are shown in Table V. Explicitly, the classifiers were biased, especially on  $\mathcal{D}_1$  as seen in Fig. 6. The implication was the same as we obtained when *gender* attribute was protected; however, we found another exciting implication which was that the robust classifier achieved much higher EOU than the standard classifier. Therefore, this inferred that when the number of protected attribute values was higher, the difference between EOUs of the standard and robust classifier also increased. Finally, we summarize that a robust classifier trained by adversarial training is more equalized odd unfair than a standard classifier.

## V. CONCLUSION AND FUTURE WORK

We have briefly explained definitions of class-wise accuracy, class-wise robustness, leave-one-out, and equalized odd unfairness and formally defined them in mathematical formulas. Further, in our experiments on the CIFAR-10 and Adult dataset, we have shown that training a classifier by adversarial training could expose more unfairness than regular training. Thus, with these findings, people may not try to create an adversarially robust classifier since it can ruin some definitions of fairness. Then, solutions addressing this problem are needed, as in the work of [22] which tried to address the issue between robustness and class-wise accuracy fairness. Furthermore, we still do not have mathematical proof of these findings and need to evaluate further whether other methods [7], [19], [24] that focus on robustness also harm these kinds of fairness.

Future work will focus on a) developing mathematical, theoretical proofs of our current empirical findings; b) checking if there exists any robustness method in particular that harms fairness; and c) addressing these problems using a regularization technique within an optimization problem or solving the optimization problem with fairness-constraints.

## ACKNOWLEDGMENT

The work presented here was supported in part by the Baylor AI lab in Baylor University’s Department of Computer Science.

TABLE V: Average confidences of the Adult classifiers to predict that incomes exceed \$50K per year where *race* attribute is protected.

Model	$\mathcal{D}_0$					$\mathcal{D}_1$				
	White	Asian	Indian american	Black	Other	White	Asian	Indian american	Black	Other
Standard	0.255	0.24	0.21	0.203	0.215	0.706	0.688	0.65	0.639	0.657
Robust	0.233	0.242	0.148	0.159	0.149	0.754	0.763	0.64	0.658	0.642

## REFERENCES

- [1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283, 2016.
- [2] Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. Robustness may be at odds with fairness: An empirical study on class-wise accuracy. In *NeurIPS 2020 Workshop on Pre-registration in Machine Learning*, pages 325–342. PMLR, 2021.
- [3] Emily Black and Matt Fredrikson. Leave-one-out unfairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 285–295, 2021.
- [4] Catherine Blake. Uci repository of machine learning databases. <http://www.ics.uci.edu/~mllearn/MLRepository.html>, 1998.
- [5] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017.
- [6] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017.
- [7] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019.
- [8] Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, 2015.
- [9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [10] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386, 2020.
- [11] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.
- [12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [13] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [14] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
- [15] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.
- [16] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.
- [17] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.
- [18] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016.
- [19] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pages 582–597. IEEE, 2016.
- [20] Ayush Singhal, Pradeep Sinha, and Rakesh Pant. Use of deep learning in modern recommendation system: A summary of recent works. *arXiv preprint arXiv:1712.07525*, 2017.
- [21] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
- [22] Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In *International Conference on Machine Learning*, pages 11492–11501. PMLR, 2021.
- [23] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.
- [24] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pages 7472–7482. PMLR, 2019.