# Estimating and Improving Fairness with Adversarial Learning

Xiaoxiao Li* Ziteng Cui† Yifan Wu‡ Lin Gu§ Tatsuya Harada ¶

**Abstract**

Fairness and accountability are two essential pillars for trustworthy Artificial Intelligence (AI) in healthcare. However, the existing AI model may be biased in its decision marking. To tackle this issue, we propose an adversarial multi-task training strategy to simultaneously **mitigate** and **detect** bias in the deep learning-based medical image analysis system. Specifically, we propose to add a discrimination module against bias and a critical module that predicts unfairness within the base classification model. We further impose an orthogonality regularization to force the two modules to be independent during training. Hence, we can keep these deep learning tasks distinct from one another, and avoid collapsing them into a singular point on the manifold. Through this adversarial training method, the data from the under-privileged group, which is vulnerable to bias because of attributes such as sex and skin tone, are transferred into a domain that is neutral relative to these attributes. Furthermore, the critical module can predict fairness scores for the data with unknown sensitive attributes. We evaluate our framework on a large-scale public-available skin lesion dataset under various fairness evaluation metrics. The experiments demonstrate the effectiveness of our proposed method for estimating and improving fairness in the deep learning-based medical image analysis system.

*xl32@princeton.edu. Princeton University.
†cuiziteng@sjtu.edu.cn . Shanghai Jiaotong University.
‡yfwu@seas.upenn.edu . University of Pennsylvania.
§lin.gu@riken.jp. RIKEN AIP, University of Tokyo.
¶harada@mi.t.u-tokyo.ac.jp. University of Tokyo, RIKEN AIP

# 1 Introduction

The success of deep learning can be partially attributed to data-driven methodologies that automatically recognize patterns in large amounts of data. However, this end-to-end paradigm leaves the deep learning model vulnerable to biases in the model itself and to biases in the data: such as, user groups with *sensitive (protected) attributes* (*i.e.* age or sex) that are over or under represented in the deep learning model [11, 9]. As a result, the deep learning models tend to replicate and even amplify this data bias. For example, a recent study revealed statistically significant differences in performance on medical imaging-based diagnoses using gender imbalanced datasets [17]. Existing chest X-ray classifiers were found exhibiting disparities in performance across subgroups distinguished by different phenotypes [22, 22]. Similar to forms of discrimination linked to face recognition [27], darker skinned patients may be under-presented [16] in existing dermatology datasets: ISIC 2018 Challenge dataset [25, 8]. A further study [1] demonstrated that the skin lesion algorithm could show variable performance relative to other under-represented subgroups. If the decision-making is (partially) based on the values of sensitive attributes, the consequences can be irreversible or even fatal, especially in medical applications.

Unfairness mitigations mainly fall in the following three areas: 1) adversarial learning [6, 26, 29, 28]; 2) calibration [30, 11]; 3) incorporating priors into feature attribution [13, 18]. The adversarial learning approach is considered the most generalizable approach to different bias inducements. Although there are bias mitigation methods proposed in text and natural image analysis, how to mitigate the biases in medical imaging has not been well explored. Additionally, although there are over 70 fairness metrics [4], they all require the explicit marking of sensitive protected attributes and the explicit labelling of classification tasks, neither of which may be explicitly identified in existing data sets in real healthcare applications.

To enhance the accountability and fairness of artificial intelligence, we propose an adversarial training-based framework. Unlike the existing bias mitigation framework, our proposed pipeline is the first to simultaneously **mitigate** and **detect** biases in medical image analysis. As illustrated in Fig.1, the *discriminator* is comprised of two tasks: 1. the **bias discrimination module** distinguishes between privileged and unprivileged samples, and 2. the **fairness critical module** predicts the fairness scores of given data. Through adversarial training with the discriminator block, the **feature generator** can generate discrimination-free features that serve as the inputs for the **feature classifier**.

**Our contributions** are summarized in four folds: 1) This work addresses the neglected bias problem in medical image analysis and provides bias mitigation solutions via the adversarial training of a discrimination module; 2) To ensure the accountability of trustworthy AI, we propose to co-train a critical module that is able to estimate bias for the inference dataset without knowing the its labels and sensitive attributes; 3) To force the critical module to make fairness estimations independently of bias discrimination, we introduce a novel orthogonality regulation; 4) As an early effort towards fair AI design for medical image analysis, our framework is demonstrated to be effective in predicting and mitigating bias in medical imaging tasks.

# 2 Problem Setup

## 2.1 Problem Statement

Given a pair of data $\{x, y\}$, where $x \in \mathbb{R}^d$ is the input image and $y \in \mathcal{C}$, where $\mathcal{C} = \{1, 2, \ldots, k\}$ and $k$ is the number of classes representing diagnose label, our framework aims to map input $x$ into an intermediate space $h = G(x) \in \mathbb{R}^{d'}$. The base classification algorithm $C(\cdot)$ then map the $h$ to the

final prediction $\widehat{y} = C(G(x))$. Using threshold, we can map output $\widehat{y}$ to class $c$, namely $f(x) \to c$, for $f = C(G(\cdot))$ and $c \in \mathcal{C}$. The intermediate representation $G(\cdot)$ should ensure the prediction $\widehat{y}$ to suffer least biases with respect to the protected attribute $z$. We denote the data set $X$, sensitive feature set $Z$ and label set $Y$. In our setting, $z \in \mathcal{Z} = \{0, 1\}$ is a binary sensitive feature that may induce bias. $z = 0$ indicates privileged samples, vice versa, $z = 1$ indicates under-privileged samples.

### 2.2 Notation of Fairness

We treat the samples with sensitive features that benefit classification as privileged samples. On the contrary, the sample with sensitive features against classification are viewed as under-privileged samples.

**Definition 2.1** (Fair classifier)**.** *We define a classifier $f$ with respect to data distribution $Pr$ on $\{(X, Z, Y) : \mathbb{R}^d \times \mathcal{Z} \times \mathcal{C}\}$, where $\mathcal{C} = \{1, 2, \ldots, k\}$ and $k$ is the number of classes. The classifier $f$ is fair if*

$$Pr(f(x, z) \in \mathcal{C} | z = 0) = Pr(f(x, z) \in \mathcal{C} | z = 1)$$

In other words, if the distance between the two distribution of outputs given the two conditions, $z = 0$ or $z = 1$, of the function is zero, then $f$ is denoted as fair classifier. We can have the shortened notation for fairness as [7]:

$$\mathcal{U}(f, \mathcal{C}) = |Pr(f(x, z) \in \mathcal{C} | z = 0) - Pr(f(x, z) \in \mathcal{C} | z = 1)|. \tag{1}$$

Instead of directly optimizing Eq. 1, we achieve the fairness while maintaining data utility through adversarial training (illustrated in Section 3).

#### 2.2.1 Fairness measurements

Given $x_i, y_i, z_i$ to be the input, label, bias indicator for the $i$th sample in dataset $\mathcal{D} = \{X, Y, Z\}$. Let denote $\widehat{y}_i$ as the predicted label of sample $i$. $Pr$ is the classification probability. The true positive and false positive rates are:

$$TPR_z = \frac{|\{i | \widehat{y}_i = y_i, z_i = z\}|}{|\{i | \widehat{y}_i = y_i\}|} = Pr_{(x_i, y_i, z_i) \in \mathcal{D}}(\widehat{y}_i = y_i | z_i = z)$$
$$FPR_z = \frac{|\{i | \widehat{y}_i \neq y_i, z_i = z\}|}{|\{i | \widehat{y}_i \neq y_i\}|} = Pr_{(x_i, y_i, z_i) \in \mathcal{D}}(\widehat{y}_i \neq y_i | z_i = z).$$

Following [5, 10, 11], we quantify the discrimination or the bias of the model using the fairness metrics *Statistical Parity Difference (SPD), Equal opportunity difference (EOD), and Average Odds Difference (AOD)*:

- *Statistical Parity Difference (SPD).* SPD measures the difference in the probability of positive outcome between the privileged and under-privileged groups:

$$SPD = Pr_{(x_i, y_i, z_i) \in \mathcal{D}}(\widehat{y}_i = y_i | z_i = 0) - Pr_{(x_i, y_i, z_i) \in \mathcal{D}}(\widehat{y}_i = y_i | z_i = 1).$$

- *Equal opportunity difference (EOD).* EOD measures the difference in TPR for the privileged and under-privileged groups:

$$EOD = TPR_{z=0} - TPR_{z=1}.$$

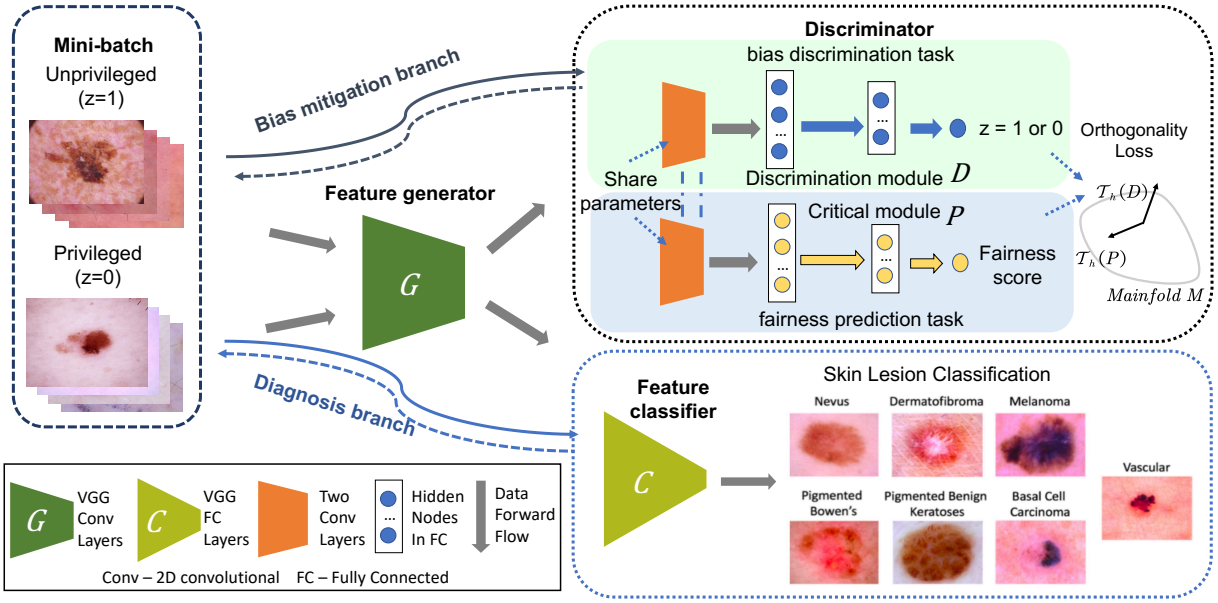



Figure 1: Overview of our proposed bias mitigation and fairness prediction pipeline.

- - *Average Odds Difference (AOD).* AOD is defined as the average of the difference in the false positive rates and true positive rates for under-privileged and privileged groups:

$$AOD = \frac{1}{2}\left[(FPR_{z=0} - FPR_{z=1}) + (TPR_{z=0} - TPR_{z=1})\right].$$

# 3 Methods

## 3.1 Overview

Our pipeline is depicted in Fig. 1. The network architecture is composed of four modules. The feature generator $G$ and feature classifier $C$ are drawn from the vanilla classification model, forming the diagnosis branch. To mitigate bias and predict fairness for inference data, we introduce two modules, the bias discrimination module $D$ for bias detection, and critical module $P$ for fairness prediction. These two modules are on the side of the discriminator and adversarially trained with $G$ , forming the bias mitigation branch. To reduce the total number of parameters, $D$ and $P$ share the first few layers, then split into two branches with the same network architectures but with different parameters.

## 3.2 Mitigate Bias via Adversarial Training

Suppose the training samples can be divided in to privileged data $X^p$ (whose $z = 0$) and under-privileged data $X^u$ (whose $z = 1$). For fair representation learning, we first initialize an adversarial training pipeline that we train a local feature extractor $G$ that takes images as inputs. $G$'s outputs are fed into disease classifier $C$ to predict $y$ and bias discrimination branch $D$ to predict $z$.

Model optimization is divided into three independent steps. First, $G$ and $C$ are trained for the target task using cross entropy $\mathcal{L}_{ce}$. Second, $G$ and $C$ are fixed, $D$ is trained to identify the sensitive attribute associated with the input (such as whether the input a male or femal). The

**Algorithm 1** Adversarial Training Pipeline

**Input:** 1. $(X^p, Y^p)$, privileged samples; 2. $(X^u, Y^u)$, under-privileged samples; 3. $G$, feature generators; $C$, disease classifier; 4. $D$, bias discrimination module; 5. $P$, critic prediction module to predict fairness scores; 6. $K$, number of optimization iterations; 7. $B$, number of mini-batch.

Initialize parameters $\{\theta_G, \theta_C, \theta_D, \theta_P\}$
**for** $k = 1$ to $K$ **do**
  **for** $b = 1$ to $B$ **do**
    Sample mini-batch $\{(X_b, Y_b)\}$ from $\{(X^p, Y^p)\}$ and $\{(X^u, Y^u)\}$
    **Train disease classifier:**
    Update $\theta_C^{(k)}, \theta_G^{(k)}$ loss $\mathcal{L}_{ce}$ to update $\theta_C^{(k)}$ and $\theta_G^{(k)}$ ▷ Cross-entropy loss.
    **Adversarial training:**
    Update $\theta_D^{(k)}, \theta_P^{(k)}$ with $\mathcal{L}_{advD} + \mathcal{L}_P + \mathcal{L}_{orth}$ ▷ Based on Eq. (2) (4) (5).
    Update $\theta_G^{(k)}$ with $\mathcal{L}_{advG}$ ▷ Based on Eq. (3).
  **end for**
**end for**
**Return: trained modules** $\{C, G, P, D\}$

objective function $L_{advD}$ for discriminator module $D$ is defined as:

$$\mathcal{L}_{advD}(X^p, X^u, D) = -\mathbb{E}_{x^p \sim X^p}[\log D(G(x^p))] - \mathbb{E}_{x^u \sim X^u}[\log(1 - D(G(x^u)))]. \tag{2}$$

Then, $D$ is fixed. $G$ is further updated to confuse $D$ using $L_{advG}$ by encouraging the outputs of $D$ for arbitrary input to be 0 — no bias in the feature space:

$$\mathcal{L}_{advG}(X^p, X^u, G) = -\mathbb{E}_{x^p \sim X^p}[\log(1 - D(G(x^p)))] - \mathbb{E}_{x^u \sim X^u}[\log(1 - D(G(x^u)))]. \tag{3}$$

### 3.3 Co-training Discriminator on Conditional and Critic Modules

Further more, we are motivated that fairness may not be guaranteed by using the adversarial strategy proposed above if the testing data distribution is not the same as the training data. To avoid providing biased decision, we want to know if the model is fair to the testing data during inference stage, when data's labels and sensitive attributes are not available. Hence, we propose to co-train a critical module $P$, which predict the fairness scores for the given trained model and the testing data. As a valid assumption in training, each mini-batch contains samples from both privileged and under-privileged samples.[i] The label of training $P$ can be any fairness measurements described in Section 2.2.1. Given a training batch $X_b$, we calculate its fairness measurement $SPD_b$ using $X_b$'s true and predicted labels at the previous epoch as the prediction label for $P$. We choose SPD score as the label because SPD is less sensitive than the alternatives [21]. Thus, with the fixed feature generator $G$, the loss function for predictor $P$ is expressed as:

$$\mathcal{L}_P(X_b, P) = \frac{1}{2}\mathbb{E}_{x \sim X_b}\|P(G(x)) - SPD_b\|^2. \tag{4}$$

During training, as the feature generator $G$ gradually adjusts the fairness in the embedded space, thus the critical module $P$ can learn various levels of fairness.

[i] This can be achieve by stratified sampling and random shuffling.

#### 3.3.1 Orthogonality Regularization

Unlike the classic adversarial training, (*i.e.*, DCGAN [20]), we propose a model equipped with an independent predictor (critic module $P$) taking the feature $h = G(x)$ output form generator for fairness prediction as the auxiliary task.

Two bottlenecks of multi-task adversarial training are the collapse of the hidden space and mutually curbed interactions of different tasks [24, 12, 19, 2]. Because the bias discrimination module $D$ and critical module $P$ are co-training using the same loss, the model behavior can be dominated by one module. The extreme scenario could be $D$ directly copies $P$ if they have the same architecture and vice versa. Motivated by [24, 19, 2] that avoid the coupling effect of multi-task training by regularizing the gradients, we design our orthogonality regularization as follows. We can describe the tangent space $\mathcal{T}_h$ for the gradients on the hidden space $h = G(x)$.

To prevent the collapse of the tangent space (namely zero gradients for all the coordinates) and ensure the independence of training discriminator module $D$ and critical module $P$, we impose a regularity condition that the two tangent vectors $\mathcal{T}_h(D) = \frac{\partial D}{\partial h} \in \mathbb{R}^{1\times d'}$ and $\mathcal{T}_h(P) = \frac{\partial P}{\partial h} \in \mathbb{R}^{1\times d'}$ for the bias discrimination task and the fairness prediction task to be orthogonal. Namely we want to achieve $\mathcal{T}_h(D) \perp \mathcal{T}_h(P)$.

We denote $\mathbf{J} = [\mathcal{T}_h(D); \mathcal{T}_h(P)]$ and define the orthogonality loss as

$$\mathcal{L}_{orth}(X, D, P) = \mathbb{E}_{x\sim X} \left\| \mathbf{J}_x^\top \mathbf{J}_x - \mathbf{I}_2 \right\|. \tag{5}$$

$\mathcal{L}_{orth}$ is simultaneously optimized with $\mathcal{L}_{advD}$ (Eq. (3)) and $\mathcal{L}_{advP}$ (Eq. (4)). Putting each module together, the training scheme is described in Algorithm 1.

## 4 Experiment

We evaluate our framework on skin lesion detection against sensitive attribute: *age*, *sex* and *skin tone*.

### 4.1 Datasets and Overall Setup

Our data is extracted from the open skin lesion analysis dataset, Skin ISIC 2018 [25, 8]. The dataset consists of 10015 images: 327 actinic keratosis (AKIEC), 514 basal cell carcinoma (BCC), 115 dermatofibroma (DF), 1113 melanoma (MEL), 6705 nevus (NV), 1099 pigmented benign keratosis (BKL), 142 vascular lesions (VASC). Each example is a RGB color image with size $450 \times 600$. The privileged and under-privileged groups of ISIC dataset are separated by *age* ( $\geq$60 and <60), *sex* (male and female) and *skin tone* (dark and light skin). [ii]

We randomly split the dataset into a training set (80%) and a test set (20%) to evaluate our algorithms. All the experiments are repeated five trials with different random seeds for splitting. Batch size is set to 64. We use Adam optimizer [15] and learning rate (lr) for classification training $(G + C)$ is 1e-3; while lr for discriminator $(D + P)$ is 1e-4.

### 4.2 Biases in the Initial Model Condition

First, we present the bias existed in the **vanilla model** which only contains feature extractor $G$ and feature classifier $C$ module. We use VGG11[23] to classify skin lesions among the 7 classes (AKIEC, BCC, DF, MEL, NV, BKL, and VASC). The convolutional layers are treated as $G$ and

[ii] The processing details for skin color labeling are described in Appendix B.

Table 1: Model performance in mean(std). Higher scores indicate better performance.

| | **Sex** | Precision | Recall | F1-score | **Age** | Precision | Recall | F1-score | **Skin** | Precision | Recall | F1-score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Vanilla | Male | 0.898 | 0.869 | 0.885 | $\geq 60$ | 0.812 | 0.798 | 0.802 | Dark | 0.852 | 0.789 | 0.810 |
| | | (0.007) | (0.005) | (0.005) | | (0.010) | (0.009) | (0.010) | | (0.015) | (0.031) | (0.026) |
| | Female | 0.934 | 0.918 | 0.901 | $< 60$ | 0.948 | 0.922 | 0.931 | Light | 0.928 | 0.896 | 0.908 |
| | | (0.018) | (0.016) | (0.015) | | (0.003) | (0.006) | (0.004) | | (0.011) | (0.019) | (0.015) |
| Ours w/o $\mathcal{L}_{orth}$ | Male | 0.908 | 0.900 | 0.903 | $\geq 60$ | 0.836 | 0.827 | 0.829 | Dark | 0.852 | 0.862 | 0.836 |
| | | (0.017) | (0.014) | (0.014) | | (0.025) | (0.025) | (0.026) | | (0.022) | (0.024) | (0.028) |
| | Female | 0.931 | 0.889 | 0.903 | $< 60$ | 0.953 | 0.941 | 0.946 | Light | 0.933 | 0.908 | 0.919 |
| | | (0.012) | (0.010) | (0.009) | | (0.008) | (0.009) | (0.009) | | (0.013) | (0.019) | (0.015) |
| Ours w $\mathcal{L}_{orth}$ | Male | 0.898 | 0.885 | 0.892 | $\geq 60$ | 0.824 | 0.835 | 0.827 | Dark | 0.845 | 0.866 | 0.839 |
| | | (0.010) | (0.013) | (0.010) | | (0.006) | (0.010) | (0.009) | | (0.004) | (0.007) | (0.011) |
| | Female | 0.890 | 0.900 | 0.895 | $< 60$ | 0.924 | 0.929 | 0.912 | Light | 0.923 | 0.910 | 0.914 |
| | | (0.016) | (0.006) | (0.005) | | (0.011) | (0.017) | (0.006) | | (0.005) | (0.010) | (0.004) |

Table 2: Fairness scores and correlations (Corr) of predictions in mean(std). Lower SPD, EOD and AOD indicate less bias. Hight Corr indicate better fairness prediction.

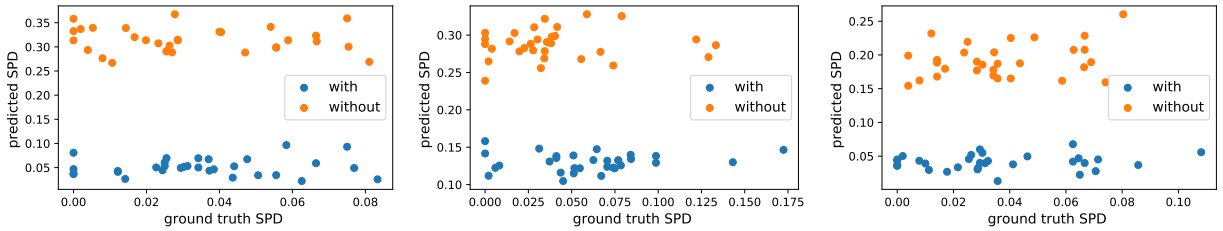
| | **Sex** ($\times 10^{-2}$) | | | | **Age** ($\times 10^{-2}$) | | | | **Skin Tone** ($\times 10^{-2}$) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SPD | EOD | AOD | Corr | SPD | EOD | AOD | Corr | SPD | EOD | AOD | Corr |
| Vanilla | 8.3 | 8.8 | 10.2 | - | 12.8 | 9.6 | 7.8 | - | 10.2 | 11.3 | 8.6 | - |
| | (2.9) | (1.6) | (2.9) | - | (3.2) | (3.0) | (1.7) | - | (3.5) | (2.9) | (2.9) | - |
| Ours w/o $\mathcal{L}_{orth}$ | 8.0 | 6.6 | **6.1** | 13.1 | 11.2 | 9.8 | 9.3 | 24.4 | 7.6 | 13.9 | 12.4 | 15.3 |
| | (4.0) | (3.8) | (4.1) | (24.8) | (2.4) | (5.3) | (4.2) | (12.2) | (4.8) | (2.2) | (5.6) | (12.4) |
| Ours w/ $\mathcal{L}_{orth}$ | **2.4** | **6.2** | 6.3 | **38.8** | **6.9** | **7.9** | **5.7** | **32.6** | **6.8** | **6.9** | **7.3** | **44.3** |
| | (0.6) | (2.1) | (1.8) | (6.0) | (0.4) | (1.1) | (1.3) | (6.6) | (0.8) | (1.9) | (1.4) | (11.3) |

the fully-connected layers serve as $C$. The results were listed in the 'Vanilla' row of Table 1, where we used different weighted metrics (precision, recall, F1-score) to evaluate the testing classification performance on the subgroups with different sensitive attributes in the format of mean(std) for five trials. We noticed the performance gaps between the subgroups, which reveals the biases. We denoted the group with better classification performance (such as female, age $< 60$, light skin) as privileged group ($z = 0$) and the opposite group (such as male, age $\geq 60$, dark skin) as under-privileged group with bias ($z = 1$).

### 4.3 Bias Mitigation and Prediction

**Bias mitigation strategies.** Now we present our main experimental study by adding our bias mitigation strategies with the bias discrimination module $D$ and critical module $P$. As shown in Fig. 1, the shared network of bias $D$ and $P$ is a two layer convolutional neural network, with channel numbers of 16 and 32, and kernel size of 3. The split branches of $D$ and $P$ are both two-layer FC networks (128-32-1), but will be updated independently. For the bias discrimination module $D$, we apply a sigmoid function to the outputs. For the critical module $P$, we average the output in the minibatch as the predicted fairness score. We followed the training strategy described in Algorithm 1.

**Fairness and classification performance.** We quantitatively measure the fairness scores (SPD, EOD, AOD) on the testing set for our models and the vanilla model, as shown in Table 2. Lower fairness scores indicate the model is less biased. All results are reported in the mean(std) format for five trials. Also, we conduct **ablation study** on orthogonality regularization $\mathcal{L}_{orth}$ (Eq 5). We name our adversarial learning scheme with and without orthogonality as 'Ours w/ $\mathcal{L}_{orth}$' and 'Ours w/o $\mathcal{L}_{orth}$' correspondingly. Compared with the vanilla model, 'Ours w/ $\mathcal{L}_{orth}$' achieves consistently




(a) Sex (b) Age (c) Skin tone

Figure 2: Fairness score (SPD) prediction for the batches of testing dataset (randomly selecting the same splitting for all) with models debiased on different sensitive attributes. Note that x-axis and y-axis have different scales and intervals.

lower fairness scores on all the above three fairness metrics, which suggests our model is less biased. Whereas, 'Ours w/o $\mathcal{L}_{orth}$' perform worse in bias mitigation than 'Ours w/ $\mathcal{L}_{orth}$.' Furthermore, as shown in Table 1, compared with the vanilla model the vanilla model, our schemes preserve the utility of the data and even boost model classification performance. Overall, 'Ours w/o $\mathcal{L}_{orth}$' slightly but not significantly outperforms 'Ours w/ $\mathcal{L}_{orth}$' on classification, which may imply a trade of adding orthogonality regularization.

**Fairness prediction** Last, we show the effectiveness of the critical module $P$ for fairness prediction on the testing dataset batches. Each batch contains 64 images drawn from both privileged and under-privileged groups. The *predicted SPD scores* are generated from $P$. We calculate the *ground truth SPD scores* based on classifier $C$ and ground truth classification labels for evaluation.[iii] The Pearson correlation between the ground truth SPD and predicted SPD are presented in the 'Corr' column of Table 2. Our method with $\mathcal{L}_{orth}$ achieves significant **more correlated** fairness prediction than without $\mathcal{L}_{orth}$. Furthermore, we show predicted and true SPD with respect to different sensitive attributes in Fig. 2. Our method with $\mathcal{L}_{orth}$ predict the **closer** fairness scores to ground truth than those without $\mathcal{L}_{orth}$. For example, in the task of predicting model fairness with respect to sex, predicted SPD scores with $\mathcal{L}_{orth}$ and ground truth **both** fall in the range [0, 0.1], while scores predicted without $\mathcal{L}_{orth}$ fall in range [0.25, 0.38].

**Remark** Compared to the vanilla model, our scheme can reduce biases in classifications and predict fairness scores for inference data, while keeping comparable classification performance. The orthogonality regularization improves bias mitigation and fairness prediction in multi-task adversarial training.

## 5 Discussion and Conclusion

In this work, we take the first step toward reducing bias in deep learning algorithms with clinical applications. Specifically, we propose an adversarial training strategy that co-trains a bias discrimination module with the vanilla classifier. Although our adversarial learning strategy shows effectiveness in bias mitigation in comparison with various fairness measurements, we cannot conclude that the model is fair for arbitrary testing cohorts. In compliance with these limitations,

[iii] Note that our prediction $P(x)$ does not require knowing inference data's labels and sensitive attributes in practice if we do not evaluate prediction correctness.

we propose a novel fairness prediction scheme (critical module) to estimate the model fairness for a group of inference data. Therefore, model users can evaluate if the deployed model meets their fairness requirements. To this end, our work provides a promising avenue for future research seeking to thoroughly mitigate bias and evaluate the fairness of the model for incoming inference data.

## References


[1] Abbasi-Sureshjani, S., Raumanns, R., Michels, B.E.J., Schouten, G., Cheplygina, V.: Risk of training diagnostic algorithms on data with demographic bias. In: Cardoso, J., Van Nguyen, H., Heller, N., Henriques Abreu, P., Isgum, I., Silva, W., Cruz, R., Pereira Amorim, J., Patel, V., Roysam, B., Zhou, K., Jiang, S., Le, N., Luu, K., Sznitman, R., Cheplygina, V., Mateus, D., Trucco, E., Abbasi, S. (eds.) Interpretable and Annotation-Efficient Learning for Medical Image Computing. pp. 183–192. Springer International Publishing, Cham (2020)

[2] Bansal, N., Chen, X., Wang, Z.: Can we gain more from orthogonality regularizations in training deep cnns? arXiv preprint arXiv:1810.09102 (2018)

[3] Barron, J.T.: A generalization of otsu's method and minimum error thresholding. ECCV (2020)

[4] Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y.: AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias (Oct 2018), https://arxiv.org/abs/1810.01943

[5] Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., et al.: Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943 (2018)

[6] Beutel, A., Chen, J., Zhao, Z., Chi, E.H.: Data decisions and theoretical implications when adversarially learning fair representations. arXiv preprint arXiv:1707.00075 (2017)

[7] Chzhen, E., Denis, C., Hebiri, M., Oneto, L., Pontil, M.: Fair regression via plug-in estimator and recalibration with statistical guarantees (2020)

[8] Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 168–172. IEEE (2018)

[9] Du, M., Yang, F., Zou, N., Hu, X.: Fairness in deep learning: A computational perspective. IEEE Intelligent Systems pp. 1–1 (2020)

[10] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. pp. 214–226 (2012)

[11] Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29**, 3315–3323 (2016)

[12] He, G., Huo, Y., He, M., Zhang, H., Fan, J.: A novel orthogonality loss for deep hierarchical multi-task learning. IEEE Access **8**, 67735–67744 (2020)

[13] Hendricks, L.A., Burns, K., Saenko, K., Darrell, T., Rohrbach, A.: Women also snowboard: Overcoming bias in captioning models. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 771–787 (2018)

[14] Huo, J.y., Chang, Y.l., Wang, J., Wei, X.x.: Robust automatic white balance algorithm using gray color points in images. IEEE Transactions on Consumer Electronics **52**(2), 541–546 (2006)

[15] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[16] Kinyanjui, N.M., Odonga, T., Cintas, C., Codella, N.C.F., Panda, R., Sattigeri, P., Varshney, K.R.: Fairness of classifiers across skin tones in dermatology. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. pp. 320–329. Springer International Publishing, Cham (2020)

[17] Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., Ferrante, E.: Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proceedings of the National Academy of Sciences **117**(23), 12592–12594 (2020)

[18] Liu, F., Avci, B.: Incorporating priors with feature attribution on text classification. arXiv preprint arXiv:1906.08286 (2019)

[19] Qi, G.J., Zhang, L., Hu, H., Edraki, M., Wang, J., Hua, X.S.: Global versus localized generative adversarial nets. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1517–1525 (2018)

[20] Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)

[21] Savani, Y., White, C., Govindarajulu, N.S.: Intra-processing methods for debiasing neural networks. Advances in Neural Information Processing Systems **33** (2020)

[22] Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I.Y., Ghassemi, M.: Chexclusion: Fairness gaps in deep chest x-ray classifiers (2020)

[23] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

[24] Suteu, M., Guo, Y.: Regularizing deep multi-task networks using orthogonal gradients. arXiv preprint arXiv:1912.06844 (2019)

[25] Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**, 180161 (2018)

[26] Wang, T., Zhao, J., Yatskar, M., Chang, K.W., Ordonez, V.: Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5310–5319 (2019)

[27] Wu, Y., Yang, F., Xu, Y., Ling, H.: Privacy-protective-gan for privacy preserving face de-identification. Journal of Computer Science and Technology (2019)

[28] Xu, D., Yuan, S., Zhang, L., Wu, X.: Fairgan: Fairness-aware generative adversarial networks. In: 2018 IEEE International Conference on Big Data (Big Data). pp. 570–575. IEEE (2018)

[29] Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pp. 335–340 (2018)

[30] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Men also like shopping: Reducing gender bias amplification using corpus-level constraints. arXiv preprint arXiv:1707.09457 (2017)

# Appendix

## Roadmap

We list the notations table in Section A. The details of data preprocessing are in Section B.

# A Notation Table

Table 3: Notations used in the paper.

| Notations | Descriptions |
|---|---|
| $x, y, z$ | input image $x \in X$, classification label $y \in Y$, binary sensitive feature $z \in Z$ |
| $X^p, X^u$ | privileged images (whose $z = 0$), unprivileged images (whose $z = 1$) |
| $X_b, Y_b$ | mini-batch images and labels |
| $\mathcal{D}$ | data set, where $\mathcal{D} = \{X, Y, Z\}$ |
| $\mathcal{Z}$ | sensitive feature space, where $z \in \mathcal{Z} = \{0, 1\}$ |
| $k$ | the number of classes |
| $\mathcal{C}$ | classification label space, where $\mathcal{C} = \{1, \ldots, k\}$ |
| $G$ | feature generator that takes $x$ as input |
| $C$ | disease classifier that takes $G(x)$ as input |
| $D$ | bias discrimination module that takes $G(X_b)$ as input and outputs bias detection result (0 or 1) |
| $P$ | critical predict module that takes $G(X_b)$ as input and predicts batch-wise fairness score |
| $f$ | classifier $f(x) = \widehat{y}$, with $f = C(G(\cdot))$ and $\widehat{y} \in \mathcal{C}$ |
| $Pr$ | probability |
| $TPR$ | true positive rate |
| $FPR$ | false positive rate |
| $SPD$ | statistical parity difference |
| $SPD_b$ | statistical parity difference score calculated on batch $b$ |
| $EOD$ | equal opportunity difference |
| $AOD$ | averaged odds difference |
| $h$ | hidden feature space, where $h = G(x)$ |
| $\mathcal{T}_h(D), \mathcal{T}_h(P)$ | gradient tangent planes |
| $\mathbf{J}$ | Jacobian matrix |
| $\mathbf{I}_N$ | $N$ by $N$ identity matrix |
| $\mathcal{L}_{ce}$ | cross-entropy loss |
| $\mathcal{L}_{advG}$ | adversarial loss for updating feature generator |
| $\mathcal{L}_{advD}$ | adversarial loss for updating discriminator block |
| $\mathcal{L}_P$ | prediction loss |
| $\mathcal{L}_{orth}$ | orthogonality regularization |

# B Data Prepossessing

We rescale the images to $224 \times 224$ by center cropping and resizing. The privileged and unprivileged groups *skin tone* is based on individual typology angle (ITA) criteria. We first use gray world white balance algorithm [14] to correct the influence of light. Then we utilize the recent Generalized Histogram Thresholding (GHT)[3] method for segmentation, which kept surrounding skin of the lesions only. We followed [16] transformed images from RGB-space to CIELab-space to calculate the ITA value of the non-lesion area. We set skins with ITA less than 45 as the dark skins. We set skins with ITA higher than 45 as the light skins. The description of this process is shown in Fig.3. Based on age, 36.2% of subjects are $\geq 60$, and 63.8% of subjects are $< 60$. Based on sex, 54.3% of subjects are male, and 45.7% of subjects are female. Based on skin tone, 43.2% of subjects are dark skin and 63.8% of subjects are light skin.

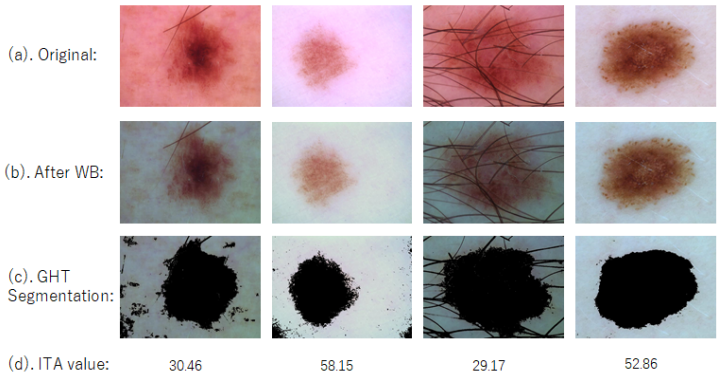



Figure 3: ITA value calculation:(a). the original image, (b). use white balance algorithm to reduce the influence of light, (c). image segmentation (d). calculate the ITA value