TARA: Training and Representation Alteration for AI Fairness and Domain Generalization

William Paul[1], **Armin Hadzic**[1], **Neil Joshi** [1], **Fady Alajaji**[2], **Philippe Burlina**[1, 3]

[1]The Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Rd, Laurel, MD 20723, USA
firstname.lastname@jhuapl.edu
[2]Department of Mathematics and Statistics
Queens University, ON K7L 3N6, Canada
fa@queensu.ca
[3]Department of Computer Science
Johns Hopkins University
3400 N. Charles Street Baltimore, MD 21218

Keywords: Fairness in AI, Generative Models, Retinal Imagery, Debiasing Metrics

Abstract

We propose a novel method for enforcing AI fairness with respect to protected or sensitive factors. This method uses a dual strategy performing training and representation alteration (TARA) for the mitigation of prominent causes of AI bias by including: a) the use of representation learning alteration via adversarial independence to suppress the bias-inducing dependence of the data representation from protected factors; and b) training set alteration via intelligent augmentation to address bias-causing data imbalance, by using generative models that allow the fine control of sensitive factors related to underrepresented populations. When testing our methods on image analytics, experiments demonstrate that TARA significantly or fully debiases baseline models while outperforming competing debiasing methods, e.g., with (% overall accuracy, % accuracy gap) = (78.75, 0.5) vs. the baseline method's score of (71.75, 10.5) for EyePACS, and (73.71, 11.82) vs. (69.08, 21.65) for CelebA. Furthermore, recognizing certain limitations in current metrics used for assessing debiasing performance, we propose novel conjunctive debiasing metrics. Our experiments also demonstrate the ability of these novel metrics in assessing the Pareto efficiency of the proposed methods.

1 Introduction

Recent advances in AI, via deep learning (DL), for tasks such as object detection (Redmon and Farhadi (2018)), retinal semantic segmentation (Pekala et al. (2019)), or skin diagnostics (P. M. Burlina, Joshi, Ng, et al. (2019)) have led to performance exceeding that of classical machine learning (P. Burlina, Freund, Dupas, and Bressler (2011)), even reaching human-level performance. However, this success is tempered by challenges such as private information leakage (Shokri, Stronati, Song, and Shmatikov (2017)), adversarial attacks (Carlini and Wagner (2017)), low shot learning (P. Burlina, Paul, et al. (2020); Ravi and Larochelle (2016)), or bias with regard to sensitive factors and protected subpopulations (P. Burlina, Joshi, Paul, Pacheco, and Bressler (2020)). These challenges threaten to derail AI deployment in many areas including healthcare, autonomy, or smart cities. In this work, we focus on addressing AI bias.

Two of the dominant sources of AI bias include: (a) data disparity or imbalance with respect to protected subpopulation(s) and (b) conditional dependence of model predictions on protected factor(s). We report on novel approaches for addressing these sources of bias, tackling source (a) via generative methods that synthesize more data for underrepresented populations, while allowing for control of specific semantic attributes of images (called intelligent augmentation or IA). We address source (b) via adversarial two player models that aim to minimize conditional dependence of the model prediction on protected factors (called adversarial independence or AD). Since models may be affected by both sources of bias, we investigate a novel method jointly exploiting the above two strategies consisting of *training and representation alteration* (termed TARA). Finally, the problem of fairness is closely related to generalization Hu, Niu, Sato, and Sugiyama (2018) and Sagawa, Koh, Hashimoto, and Liang (2019). In our study, rather than typical settings that use the natural proportions in the dataset, we focus on an extreme case of data bias akin to domain generalization, where the minority subpopulation, defined as the combination of protected and target classes with the fewest data points, is entirely excluded.

2 Prior Work

Adversarial Approaches: A number of recent studies have investigated AI fairness, for instance Zemel, Wu, Swersky, Pitassi, and Dwork (2013) studies demographic parity, Bolukbasi, Chang, Zou, Saligrama, and Kalai (2016) tackles debiasing of word embeddings, Prost et al. (2019) and Zemel et al. (2013) explore distribution matching to maximize fairness, and Kinyanjui et al. (2020) presents a method for medical image debiasing for skin segmentation. Several studies addressing the *conditional dependence* source of bias employ adversarial methods (Goodfellow et al. (2014)). Enforcing fairness in domain adaption in Ganin et al. (2016) uses this strategy. Beutel, Chen, Zhao, and Chi (2017) uses a separate adversarial network in a natural language processing task to predict the protected factor, and modifies the word embeddings to reduce the adversary's performance. Similarly, Alvi, Zisserman, and Nellåker (2018) employs multiple network heads and a cross entropy loss comparing the predicted distribution to a uniform distribution to reduce bias in embeddings across multiple protected factors in the image domain. The concurrent studies Wadsworth, Vera, and Piech (2018) and Zhang, Lemoine, and Mitchell (2018), the primary inspiration for our methods, use a similar approach and expand to tabular data targeting various forms of fairness. Also, Song, Kalluri, Grover, Zhao, and Ermon (2019) uses an information theoretic approach to mitigate bias through fair controllable representations of data and sensitive factors on tabular data. Translating this to the image domain, Quadrianto, Sharmanska, and Thomas (2019) demonstrates a method of bias reduction by exploiting a highly unconstrained mapping to latent space using residual statistics and enforcing equality of outcome in visual features of faces. The method from Wang, Zhao, Yatskar, Chang, and Ordonez (2019) shows that even balanced datasets can exhibit bias and that using adversarial methods to mask out markers of protected factors (gender) directly in the image domain can provide benefits in some cases. Edwards and Storkey (2015)'s approach uses an adversarial independence approach to maximize fairness of an autoencoder internal representation, which is somewhat similar to our own method, but diverges from it by maximizing the less strict demographic parity instead of equality of odds. These findings motivate our approach. However, we depart from these studies in several important ways: our adversarial network feeds off the internal representation to further reduce protected factor information leakage to the adversarial network, and we combine this with a novel augmentation that allows for selective image marker alteration. Additionally, we consider more extreme cases of bias where the minority population is not represented at all in the dataset.

Augmentation and Generative Methods: To address the other important cause of bias, i.e., data imbalance, our strategy for debiasing also leverages intelligent augmentation, which uses generative models that allow fine control of image attributes for underrepresented factors. Generative approaches applicable to generating more synthetic data include: generative models such as GANs (Grover et al. (2019); Karras, Laine, and Aila (2019)), autoencoders(Madras, Creager, Pitassi, and Zemel (2018)), variational autoencoders (VAEs) (Kingma and Welling (2013); Louizos, Swersky, Li, Welling, and Zemel (2015)) and generative autoregressive models, invertible flow-based latent vector models, or a hybrid of such models Zhao, Song, and Ermon (2017). Such methods have limitations for addressing data imbalance and bias: while they generate realistic images, they do not allow for controlling images with specific attributes (e.g., P. M. Burlina, Joshi, Pacheco, Liu, and Bressler (2019) or Karras et al. (2019)), which would correspond to an underrepresented population (e.g., images of dark skin individuals with Lyme disease, P. M. Burlina et al. (2020)). This motivates the need for methods that allow fine control of individual semantic factors. Along with this control, there is also the matter of ensuring those other uncontrolled attributes remain invariant, which arguably requires *disentanglement*. Control and disentanglement are related (but distinct) concepts formally defined via information theoretic measures (Paul, Wang, Alajaji, and Burlina (2021)). Paul et al. (2021) shows that optimizing the control of semantic attributes also promotes (theoretically and empirically) *disentanglement* among latent factors. Disentanglement appears to have an incidence in promoting fairness as shown empirically in Locatello et al. (2019); however this work does not provide a method to achieve this.

Likewise, generative methods such as Paul et al. (2021) that allow semantic control are interesting for intelligent augmentation but are in practice, not practical for debiasing, due to the fact that such methods hinge on the discovery of factors that are aligned with protected factors (e.g., sex) and also because of the residual entanglement that would remain in the controlling latent space codes Paul et al. (2021).

In sum, while much has been done in past generative modeling, the aforementioned limitations have motivated the development of our novel debiasing approach, which, is able to perform fine semantic attribute control for intelligent augmentation by using latent space manipulation methods, and also importantly enables control of such attributes while keeping other factors of variations fixed, thereby addressing entanglement.

3 Novel Contributions

Our novel contributions are therefore as follows:

1. *Approach:* We introduce a new debiasing strategy, that is able to perform intelligent augmentation by exploiting a novel latent space manipulation method which can finely control data attributes, and also adds a newly formulated adversarial two player approach for enforcing conditional independence working off the prelogits layer of the classification network. We call this overall approach TARA. This method is able to address, for the first time, dual sources of bias (imbalance and dependency) via the combined alteration of training data and data representation. This method is shown to significantly outperform competing debiasing strategies.

2. *Metrics:* We identify and address certain shortcomings of current fairness metrics by proposing novel metrics and demonstrating their utility.

3. *Generalization:* We demonstrate the ability to debias in scenarios of extreme data imbalance entailing domain generalization, which to our knowledge, has not been widely addressed in the AI fairness literature, where models completely lacked training data for specific subpopulations (e.g., dark skin individuals with retinal diseases). This is particularly important for several reasons: subpopulations with combinations of factors (e.g., race/age/gender) yield partitions with little or no training data. Furthermore, it addresses domain transfer when a training dataset was curated for a certain population (e.g., a diabetic retinopathy detector developed for the US population) and is deployed to a new domain (e.g., Singapore) now including a new ethnicity (e.g., Malais) not contained in the initial population.

4. *Proxy Measure for Sensitive Attributes and its Robustness:* We start examining the robustness of a debiasing approach to mismatch in sensitive attributes. Specifically, we propose for the first time the use of a proxy factor (i.e., the Individual Typology angle or ITA) for a sensitive attribute and demonstrate transfer of debiasing faculty from a proxy protected factor to a target factor.

   Finally, as a contribution to medical imaging, we demonstrate, for the first time, the use of ITA as a relevant protected factor for retinal images, allowing for debiasing without requiring costly and error-prone manual clinical image annotation.

4 Methods

Nomenclature and Definitions of Fairness: Henceforth as nomenclature we denote protected factor(s) by the random variable S, the classifier's prediction by $\hat{Y}$, and the underlying true label by Y. Developing unbiased AI systems requires a clear understanding of what constitutes fairness, which is ideally expressed in formal mathematical terms. Common definitions of fairness (Hardt, Price, and Srebro (2016); Mehrabi, Morstatter, Saxena, Lerman, and Galstyan (2019)) include *demographic parity, equality of odds, and equality of opportunity* (see Appendix for mathematical formulations). All these formal definitions entail some form of conditional independence of the prediction $\hat{Y}$ from the sensitive protected factor S. Equality of odds, in particular, states that a predictive model must produce predictions that are conditionally independent of protected factors given the true outcome:

$$P(\hat{Y} = \hat{y} | S = s, Y = y) = P(\hat{Y} = \hat{y} | Y = y), \forall s, y, \hat{y}. \tag{1}$$

This motivates a method of debiasing that directly learns a data representation that exhibits conditional independence, an ingredient we use in the adversarial independence and TARA methods. Our study adopts the stricter goal of equality of odds, which yields equality of performance (accuracy). This leads us to measure debiasing performance using commonly adopted metrics such as accuracy and area under the receiver operating characteristic curve (AUC). We also propose novel metrics that are consistent with this goal, but address certain limitations of accuracy, which are detailed later.

We begin by introducing the adversarial independence method for minimizing bias by maximizing conditional independence. Next, we present intelligent augmentation, a method addressing data generation for underrepresented classes, which we then combine with adversarial independence to form the TARA method.

Adversarial Independence: A strategy for achieving resilience to bias is to learn a data representation where information about a protected factor can be suppressed. We use a method for debiasing neural network models using adversarial training derived from Zhang et al. (2018). During training, we simultaneously train a prediction network F with input X and parameters Θ_F on a classification task represented via

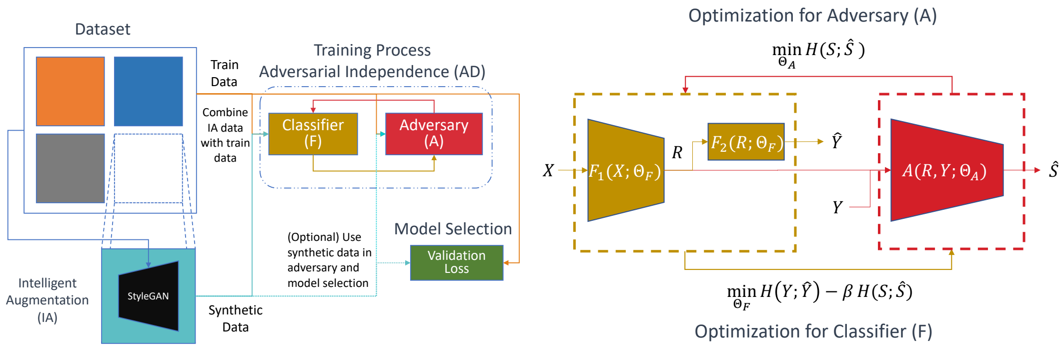
$$F(X; \Theta_F) = F_2(F_1(X; \Theta_F); \Theta_F) = F_2(R; \Theta_F) = \hat{Y}$$

and an adversarial network A to predict the protected factor S, as illustrated in Figure 1. In the process of computing $\hat{Y}$, the prediction network calculates some internal representation $R = F_1(X; \Theta_F)$ that should be fair (chosen based on F and F_2). We consider two cases for the internal representation R:

- R corresponds to the logits with $F_2(R; \Theta_F) = \mathit{softmax}(R)$.
- R is given by activations prior to the final linear layer with $F_2(R; \Theta_F)$ being a linear layer followed by a softmax layer.

Simultaneously, the adversarial network with parameters Θ_A ingests this internal representation R and the true Y to produce $\hat{S}$ (the prediction for S):

$$A(R, Y; \Theta_A) = \hat{S}.$$



(a) TARA Process

(b) Adversarial Independence

Figure 1: (a) The TARA process performs intelligent augmentation to generate samples for a subpopulation with no representation in the dataset. The resulting data can be used to train the adversarial module (A), described in (b), for adversarial independence training and classifier F model selection. Represented by the teal dashed lines, the generated samples can be optionally filtered out of the input to the adversary in all phases of training and out of the model selection process (validation loss). (b) Bias is reduced by maximizing the prediction module's performance while minimizing the adversarial module's (A) ability to predict the protected factor (S) using the internal representation R and known label (Y).

The cross entropy $H(\cdot;\cdot)$ is[1] then applied to each of the two predictions and combined to compute the total loss which is optimized as follows:

$$\min_{\Theta_F} \max_{\Theta_A} H(Y; F(X; \Theta_F)) - \beta H(S; A(R, Y; \Theta_A)) \tag{2}$$

where the $\beta > 0$ is a hyperparameter used to balance the impact of the adversarial loss contribution. In Equation 2, we use back propagation to optimize the prediction network Θ_F parameters with the total loss and then the adversarial network's Θ_A parameters using $H(S; \hat{S})$. Combining the impact of the two loss terms ensures that the prediction network will be penalized for producing an R that can be used to reproduce the protected factor. The resulting prediction network should then be more resilient to bias with respect to the protected factor.

Intelligent Augmentation: An alternative approach to improving model fairness is by generating more data for underrepresented populations to reduce dataset imbalance. For example, consider the retinal image analysis use case where the goal is to generate retinal images for underrepresented populations (e.g., dark skin individuals with referable diabetic retinopathy (DR)). These underrepresented retinal images could be generated from subpopulations whose data is more abundant (e.g., healthy dark skin individuals or DR-referable light skin individuals), while holding other image characteristics invariant (importantly, disease markers, but also vasculature and possibly other more subtle markers like gender markers Poplin et al. (2018)). Our method does this through a combination of data generation using multiscale GANs (StyleGAN) and altering a desired

[1] Recall that, given two random variables U and V with distributions P_U and P_V, respectively, then the cross-entropy between U and V is denoted by $H(U; V)$ and is given by $H(U; V) = \mathbb{E}_{P_U}[-\log P_V(U)]$, where $\mathbb{E}_{P_U}[\cdot]$ denotes expectation under the P_U distribution.

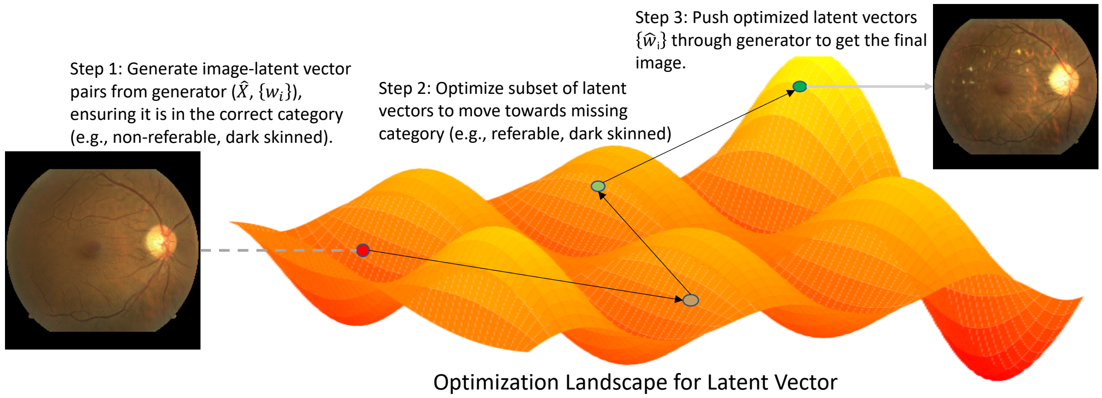


Figure 2: Conceptual depiction of intelligent augmentation which manipulates generated images to impart desired factors of underrepresented populations while keeping other factors invariant, shown on retinal data. Step 2 uses Equation 3 to update the latent vectors.

factor by controlling the direction of change in latent space via gradient descent that maximally alters a factor of choice. Once the GAN is trained on the available data, the desired transformation is obtained in three steps, as illustrated in Figure 2: (1) sample from the generator to obtain image and style space vectors pairs ($\hat{X}$, $\{w_i\}$), where i denotes the resolution scale; (2) map the vectors $\{w_i\}$ into $\{\hat{w}_i\}$, imparting the desired factor value change in latent space; (3) map the latent space vector $\{\hat{w}_i\}$ to get the final image with the factor changed.

A classifier C_1 is first trained on the factor of interest on real images that are then used to label the vector-image pairs ($\hat{X}$, $\{w_i\}$), generated unconditionally from StyleGAN. To make the optimization process easier, we then bootstrap a second classifier C_2 that takes the $\hat{X}$'s corresponding style vectors $\{w_i\}$ as input and replicates the prediction of C_1 on $\hat{X}$. As a result, C_2's gradient can then be used to control $\{w_i\}$ directly, as gradient descent yields a non-rectilinear trajectory in latent space that is maximally modifying with respect to (w.r.t.) the selected factor.

Consider, for example, a linear discriminator that is used to separate images in latent space W w.r.t. the selected factor. This linear discriminator determines the hyperplane that separates training vectors into two classes: those that have vs. those that do not have the selected factor. Then the resulting direction normal to this hyperplane is maximally changing w.r.t. the selected factor.

However, instead of a simple linear classifier we use a fully connected network for C_2. The loss function of this classifier is used to perform gradient descent in W so as to arrive at $\hat{W}$, that has the desired *softmax* value, thereby allowing fine control of the degree to which the factor is expressed. Mathematically, this is expressed as:

$$w_{i,j+1} = w_{i,j} - \gamma \nabla_{w_{i,j}} \log(C_2(\{w_{i,j}\}_i)[1]), \hat{w}_i = w_{i,n} \tag{3}$$

where i is the resolution scale, n is the total number of steps, j is the current step to approach the desired factor, and $C_2(\{w_{i,j}\}_i)[1]$ is the probability value that C_2 outputs for $Y = 1$.

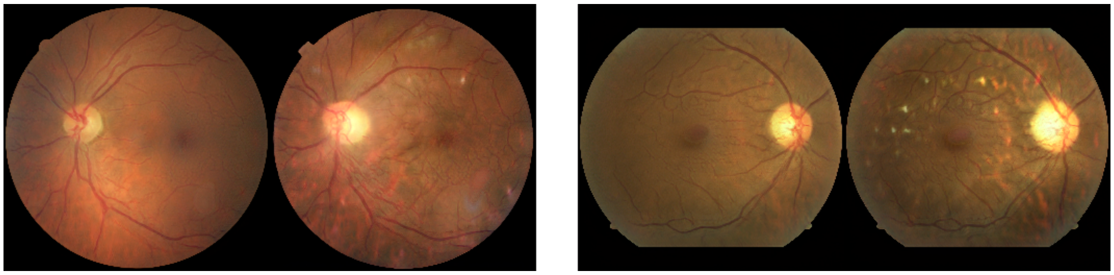


Figure 3: Examples of generating new data for underrepresented populations (here, dark skin individuals with referable diabetic retinopathy (DR)) via our intelligent augmentation. Left are the original generated samples, and right are the transformed versions, characterized by DR lesions (bright spots) and variations in fundus background, see P. Burlina, Joshi, et al. (2020).

Although we have a classifier that can tell the sensitive attribute from an image, this does not mean that other attributes are not affected. As StyleGAN encourages disentanglement between different resolution scales of the style vectors so that certain attributes are represented only in certain resolution scales, we only employ this update for specific scales to get $\{\hat{w}_i\}$, typically the finer styles as the coarse styles overly affect the image. We can then generate the final image by passing $\{\hat{w}_i\}$ through the generator. See Figures 3 and 4 for examples of this transformation.

Alternatively, the conditional StyleGAN Oeldorf and Spanakis (2019) could have been used in place of the unconditional StyleGAN we selected. While the conditional StyleGAN has the potential benefit of conditioning the generator on the protected factor, it would limit our understanding of how well the missing factor was supervised, which is why we chose to use the unconditional StyleGAN. There is a possibility for the conditional StyleGAN to produce unrealistic examples; should this be the case, we would have little recourse to correct the generator. However, the unconditional StyleGAN allows for superior control over the protected factors using the style vector; should these results be unrealistic, we can make corrections. The unconditional StyleGAN's generator also has the added benefit of being reusable to control multiple sensitive attributes.

Training and Representation Alteration (TARA): The intelligent augmentation method supplements underrepresented classes with additional samples, whereas the adversarial independence method modifies the training procedures such that protected factors cannot be accurately predicted from learned representations produced by a classifier. Consequently, as these two methods affect complimentary domains, we combine the two methods into what we call the TARA method. Last, as the augmentation process approximates the true category that is missing in the dataset, any discrepancies between our synthetic category and images in the true category might be magnified by the application of adversarial independence. Consequently, excluding synthetic images in the training of adversarial independence can be beneficial. We denote this system by TARA+Filter and depict it in Figure 1.(a) via the deactivation of the teal dashed lines.

5 Metrics

In this section we recall existing metrics and propose novel metrics to characterize debiasing performance:

Overall Accuracy and Accuracy Gap: The accuracy gap acc_{gap} for a given model is measured as the difference in accuracy between the populations that have the maximum and the minimum accuracy. Reducing acc_{gap} is a prime objective in conjunction with maintaining overall accuracy acc for the debiased algorithm. The possible trade-off between both objectives of acc and acc_{gap} justifies the need for a single metric to assess a debiased models' performance.

Minimum Accuracy (acc_{min}): A potential single performance metric acc_{min} is based on the minimum accuracy across all protected subpopulations.

Conjunctive Accuracy Improvement (CAI_α): We propose two novel single performance measures as possible indicators of success. The first is a weighted linear combination of two differential terms including the (signed) accuracy gap decrease and the (signed) overall accuracy improvement, where both terms are computed with respect to a baseline and candidate algorithm:

$$CAI_\alpha = \alpha(acc^b_{gap} - acc^d_{gap}) + (1-\alpha)(acc^d - acc^b) \tag{4}$$

where α is a weight coefficient and acc^b and acc^d denote the accuracy of the baseline and debiased models, respectively. Similarly, gap^b and gap^d represent the accuracy gap of the baseline and debiased models, respectively. We call this metric the Conjunctive Accuracy Improvement (CAI_α). Deciding how to weigh the respective importance of the two metrics is a matter beyond engineering which also should involve ethicists and policy makers (see Section 8 for details). The cases of $\alpha = 0.5$ and 0.75 are reported here for illustrative purposes to motivate future discussions.

Generalizing to other metrics ($CAUCI_\alpha$): The second proposed metric extends CAI_α to AUC, which we call the Compound AUC Improvement (abbreviated henceforth as $CAUCI_\alpha$). We also use AUC gap and minimum AUC for consistency. These ideas can similarly be extended to other metrics such as F1-score (not pursued here).

6 Datasets Used and Individual Typology Angle

6.1 OSMI Mental Health

First, to understand the effect of the proposed novel metrics, we used adversarial independence (without intelligent augmentation) on OSMI (OSMI Mental Health in Tech Survey 2016 LTD (2016)). This tabular records dataset released on Kaggle to encourage evaluation of the state of mental health across the technology industry. We investigated gender and age bias when predicting whether a person sought treatment for mental illness. We omitted the *other* class from the three possible gender classes (*male, female, other*) due to ambiguous sample quantity and quality. For age debiasing we simplified the problem to a binary class: younger (≤ 40 years old) and older (> 40 years old). Eight tabular features were used to train the binary prediction network.

(a) Young to Old

(b) Light to Dark

Figure 4: Examples of new data generation for underrepresented populations (older females in (a), and older individuals with darker skin in (b)) via our intelligent augmentation on faces. Left are the original generated samples, and right are the transformed versions.

6.2 Individual Typology Angle

Next we describe image datasets used with all (adversarial independence, intelligent augmentation, and TARA) debiasing methods. Before doing so, we introduce a method used as a proxy for race and skin tone[2] protected factor, the Individual Topology Angle (ITA) Wilkes, Wright, du Plessis, and Reeder (2015). The ITA was found to correlate with the Fitzpatrick Skin Type typically used in dermatology for characterizing the skin color of an individual. To compute ITA, an image is first converted to the CIELab color space, which was designed to match perceptual differences with differences in numerical values L of lightness, scale a between red and green, and scale b between blue and yellow. The computed per pixel ITA is:

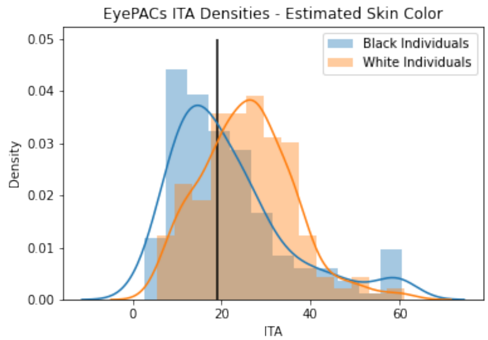
$$ITA = \frac{180}{\pi} \arctan\left(\frac{L-50}{b}\right). \tag{5}$$

The ITA for a given image was then computed by averaging pixel values over some masked area, which is determined for each dataset (see Appendix).

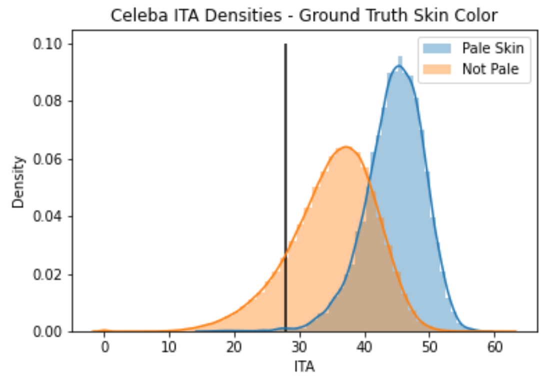
6.3 EyePACs

Sourced from the Kaggle Diabetic Retinopathy challenge, EyePACS (EyePACS (2015)) includes retinal fundus images of individuals potentially affected by diabetic retinopathy (DR). The original labels, ranging from 0 (not afflicted) to 4 (most severe), were

[2] Our usage of the terms *race*, *ethnicity* and *skin tone* are consistent with P. Burlina, Joshi, et al. (2020) and Christiansen et al. (2020).



(a) EyePACs ITA Density



(b) CelebA ITA Density

Figure 5: We use the ITA as a proxy for skin color. The ITA distribution is shown for retinal (EyePACS) and face (CelebA) datasets. Black lines denote cutoff used to create a binary variable as a sensitive factor. Both figures show consistency with other labels, estimated skin color of individuals for EyePACs and pale skin for CelebA.

binarized to denote a status greater than (mild DR). In P. Burlina, Joshi, et al. (2020), the dataset was also annotated by a clinician with an additional label that reflects an estimation of the individual's skin color in relation to their race: a binary factor reflecting image markers related to race such as darker pigmentation in the fundus, thicker blood vessels, larger cup to optical disk sizes usually associated with Black individuals rather than White individuals.

Although we have annotations for this sensitive attribute, the data collection was expensive, time consuming, and incomplete, covering only a small portion of the dataset that was then extrapolated to the rest of the dataset. There are issues with uncertainty vis-a-vis self-reporting of race. Consequently, rather than training using this attribute as S, we instead investigate using another attribute as S that requires no manual annotation, namely the ITA, and evaluate the resulting systems on both types of sensitive attributes, using ITA as a proxy for the estimated skin color.

To calculate the ITA, we added another binary variable denoting fundi with ITA ≤ 19, which is taken to mean dark skinned. This cutoff is chosen to match established categories in Kinyanjui et al. (2020) and to mimic the previous label for the annotated race, as the fundus color should be a major component of that label. From Figure 5a, we see that the cutoff closely separates the two distributions induced by the race label, but not exactly due to potential existence of other criteria in the labeled race (optical disk size) as well as retinal artifacts affecting computing the ITA.

6.4 CelebA

CelebA consists over 200,000 celebrity faces with various descriptive factors, including gender, considered as a protected factor; and age, as a prediction target. As with EyePACs, ITA is used as a protected factor (as proxy for skin color), and annotated for each image in a manner similar to Merler, Ratha, Feris, and Smith (2019). Figure 5b plots the distribution of ITA computed for CelebA vs. ground truth labels for the existence of pale skin. This figure demonstrates separation between the two modes for those with pale skin vs. those without pale skin, with some overlap between ITA of 40 and 50. One potential factor that may hurt separation is a significant number of celebrities having bright lights shining directly on them, overly lightening their skin relative to their actual skin color. To focus on extreme cases of data bias, we use a threshold of 28, which matches category thresholds from Kinyanjui et al. (2020) and also excludes almost all individuals denoted as Pale Skin.

7 Experiments

Domain Generalization: We address the use case of pronounced data imbalance, where we constructed training partitions in a manner where a certain category of targeted and protected factors, namely the ones that contain the fewest members, were excluded entirely. This tested how well the proposed method generalized to unseen categories, a bias problem akin to *domain generalization*. Alternative and more benign cases were reported in the Appendix, in addition to extensive datasets details.

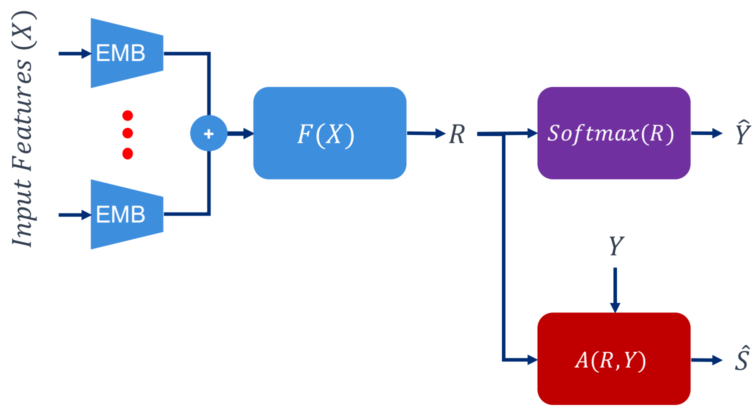


Figure 6: Adversarial Independence Neural Network ($EMB + AD$) data using an embedding-based prediction module (F), for a given tabular input feature (X), to produce an internal representation (R) and corresponding prediction ($\hat{Y}$).

Implementation details for AD: In implementing adversarial independence on OSMI, we provided the logits to the adversary due to simplicity. For image data, since adversarial independence using the logits performed close to the baseline in most respects and did not affect the model, we used the activations prior to the final linear layer as the input to the adversary. We also used two settings for $\beta = 0.5, 1.0$ and reported the best overall accuracy.

OSMI Experiments: We developed a series of experiments in order to evaluate the impact of adversarial independence on improving fairness, with respect to gender and age, in an extreme case of domain generalization when predicting whether a given person sought mental health treatment using the OSMI dataset, and assessing the novel performance metrics before moving to image data. Due to the tabular nature of the dataset we did not perform any intelligent augmentation and instead explored adversarial independence more extensively. Our proposed adversarial independence debiasing approach (*EMB+AD*, depicted in Figure 6) included a prediction network with an

Metrics	Baseline (EMB)	EMB+Noise	EMB+AD
$acc(\%)$	63.22 (5.07)	67.82 (4.91)	**82.18 (4.02)**
$acc_{gap}(\%)$	20.69 (0.87)	24.14 (1.42)	**4.59 (0.58)**
$acc_{min}(\%)$ (subpop.)	52.87 (F)	55.75 (F)	**79.89 (M)**
$CAI_{0.5}(\%)$	-	0.58	**17.53**
$CAI_{0.75}(\%)$	-	-1.44	**16.82**
AUC	0.7213 (0.0471)	0.8030 (0.0418)	**0.8592 (0.0365)**
AUC_{gap}	**0.0606 (0.0087)**	0.0610 (0.0102)	0.0805 (0.0129)
AUC_{min} (subpop.)	0.8171 (M)	**0.8430 (M)**	0.8236 (M)
$CAUCI_{0.5}$	-	0.0407	**0.0590**
$CAUCI_{0.75}$	-	**0.0201**	0.0196

Table 1: Performance metrics for debiasing methods on OSMI predicting Y= sought mental health treatment, trained on partitioning with respect to S = Gender, and evaluated on a test set balanced across treatment status and gender (M/F). Methods include: embeddings prediction network (EMB), noise debias (Noise), adversarial debias (AD). 95% confidence intervals are shown in parentheses.

Metrics	Baseline (EMB)	EMB+Noise	EMB+AD
acc	68.36 (4.29)	75.00 (3.99)	**80.75 (3.63)**
acc_{gap}	21.68 (1.15)	18.14 (1.41)	**1.33 (0.14)**
acc_{min} (subpop.)	57.52 (Older)	65.93 (Older)	**80.09 (Older)**
$CAI_{0.5}$	-	5.09	**16.37**
$CAI_{0.75}$	-	4.32	**18.36**
AUC	0.7658 (0.0444)	0.8211 (0.0353)	**0.8787 (0.0301)**
AUC_{gap}	0.0014 (0.0002)	**0.0003 (0.0001)**	0.0139 (0.0021)
AUC_{min} (subpop.)	0.8648 (Older)	**0.8780 (Older)**	0.8761 (Younger)
$CAUCI_{0.5}$	-	0.0282	**0.0502**
$CAUCI_{0.75}$	-	0.0147	**0.0189**

Table 2: Performance metrics for debiasing methods on OSMI predicting Y= sought mental health treatment, trained on partitioning with respect to $S = $ Age, and evaluated on a test set balanced across treatment status and age (Older/Younger).

embedding for each of input features concatenated and forwarded to a fully connected layer (*EMB*), while the adversarial debiasing network was a single fully connected layer (*AD*). The gender debiasing results shown in Table 1 reflect a 18.96% increase in overall accuracy and 16.1% reduction in accuracy gap and between male and female when using *EMB+AD* compared with the baseline method *EMB*. The CAI_α mirrored both of these accuracy improvements with $CAI_{0.5}$ and $CAI_{0.75}$ being the highest for *EMB+AD*. Similar to gender debiasing, the age debiasing results in Table 2 suggest *EMB+AD* had superior performance over the other two methods. Moreover, applying Gaussian noise (*Noise*) to the prediction module logits as regularization was not found to be as beneficial as using *AD* to explicitly reduce bias towards the protected factor. Results also illustrated the effectiveness of the conjunctive metrics in reflecting the best overall fairness performance in a compact manner.

EyePACS Experiments: We tested bias induced by domain generalization, by excluding DR-referable fundi from dark skin individuals from training, and kept the training dataset otherwise balanced across the DR label. Testing used two cases. The first used a test set equally balanced between the disease status (DR) and the skin color (ITA), with 600 examples per category. The second case tested generalization to changing the protected factor used in testing from ITA to race. This was done by matching the test set conditions in P. Burlina, Joshi, et al. (2020) that were equally balanced across disease status and estimated skin color by the clinician, with 100 examples per category. The two test sets were disjoint from each other. Table 3 shows that TARA+Filter performed best in terms of $CAI_{0.5}$, with accuracy gap reduced to almost zero while overall accuracy increased by 7.5%. For the second test case in Table 4, TARA+Filter performed best, and reduced the accuracy gap to near zero. In both cases, TARA outperformed adversarial independence which in turn outperformed intelligent augmentation, with all methods beating the baseline.

Metrics	Baseline (β = 0)	AD (β=0.5)	IA (β=0)	TARA (β=0.5)	TARA+Filter (β=1.0)
acc	70.0 (1.83)	76.12 (1.71)	71.5 (1.81)	**78.0 (1.66)**	77.5 (1.67)
acc_{gap}	3.5 (3.66)	2.41 (3.41)	1.5 (3.61)	2.34 (3.32)	**0.16 (3.34)**
acc_{min} (subpop.)	68.25 (Dark)	74.92 (Light)	70.75 (Dark)	76.83 (Light)	**77.42 (Dark)**
$CAI_{0.5}$	-	3.61	1.75	4.58	**5.42**
$CAI_{0.75}$	-	2.35	1.875	2.87	**4.38**
AUC	0.7861 (0.0164)	0.835 (0.0149)	0.7733 (0.0168)	0.851 (0.0142)	**0.8547 (0.0141)**
AUC_{gap}	0.0318 (0.0314)	0.0299 (0.0288)	**0.0051 (0.032)**	0.0185 (0.0280)	0.0305 (0.0275)
AUC_{min} (subpop.)	0.7941 (Light)	0.8316 (Light)	0.7969 (Light)	0.8477 (Light)	**0.8473 (Light)**
$CAUCI_{0.5}$	-	0.0254	0.0070	0.0391	**0.0654**
$CAUCI_{0.75}$	-	0.0136	0.0168	0.0262	**0.0639**

Table 3: Performance metrics for debiasing methods on EyePACs predicting Y= DR Status, trained on partitioning with respect to S = ITA, and evaluated on a test set balanced across DR status and ITA. Methods include baseline, IA: intelligent augmentation, AD: adversarial independence, TARA: TARA with or without filtering.

CelebA Experiments: For CelebA, we conducted two experiments of partitioning to predict age, one for each protected factor (gender and skin color). For each protected factor we proceeded similarly as with EyePACS, in that we excluded the smallest subpopulation from our training dataset, older females for gender partitioning and older dark skinned individuals for skin color partitioning. The test set used for both experiments was balanced across age, gender, and skin color with 1000 examples per category.

Metrics	Baseline (β=0)	AD (β=0.5)	IA (β=0)	TARA (β=0.5)	TARA+Filter (β=1.0)	(Prior) Baseline	(Prior) IA
acc	71.75 (4.41)	76.00 (4.19)	73.25 (4.34)	76.5 (4.16)	**78.75 (4.01)**	66.75 (4.62)	74.75 (4.29)
acc_{gap}	10.5 (8.76)	5.0 (8.46)	7.5 (8.64)	4.0 (8.30)	**0.5 (8.01)**	12.5 (9.15)	7.5 (8.54)
acc_{min} (subpop.)	66.5 (Black)	73.5 (Black)	69.5 (Black)	74.5 (Black)	**78.5 (Black)**	60.5 (Black)	71.0 (Black)
$CAI_{0.5}$	-	4.88	2.25	5.63	**8.5**	-	-
$CAI_{0.75}$	-	5.19	2.625	6.06	**9.25**	-	-
AUC	0.7707 (0.0412)	0.8369 (0.0362)	0.7936 (0.0397)	0.8608 (0.0339)	**0.8703 (0.0329)**	-	-
AUC_{gap}	0.1229 (0.0813)	0.0447 (0.0725)	0.0545 (0.0784)	0.0299 (0.0678)	**0.0111 (0.0656)**	-	-
AUC_{min} (subpop.)	0.7107 (Black)	0.8134 (Black)	0.7701 (Black)	0.8455 (Black)	**0.8661 (Black)**	-	-
$CAUCI_{0.5}$	-	0.0722	0.0457	0.0916	**0.1057**	-	-
$CAUCI_{0.75}$	-	0.0752	0.0570	0.0923	**0.1087**	-	-

Table 4: Performance metrics for debiasing methods on EyePACs predicting Y= DR Status, trained on partitioning with respect to S = ITA, and evaluated on a test set balanced across DR status and estimated skin color/race. The last two columns compare with methods in P. Burlina, Joshi, et al. (2020) which were trained with the partitioning with s = race.

Metrics	Baseline (β=0)	AD (β=1.0)	IA (β=0)	TARA (β=0.5)	TARA+Filter (β=1.0)
acc	69.08 (1.01)	73.31 (0.97)	69.9 (1.01)	72.09 (0.98)	**73.71 (0.96)**
acc_{gap}	21.65 (1.96)	14.02 (1.91)	13.1 (1.98)	18.37 (1.93)	**11.82 (1.91)**
acc_{min} (subpop.)	58.25 (Female)	66.3 (Female)	63.35 (Female)	62.9 (Female)	**67.8 (Female)**
$CAI_{0.5}$	-	5.93	4.685	3.15	**7.34**
$CAI_{0.75}$	-	6.78	6.618	3.21	**8.53**
AUC	0.7506 (0.0095)	**0.8206 (0.0084)**	0.766 (0.0093)	0.8042 (0.0087)	0.8164 (0.0085)
AUC_{gap}	0.1242 (0.016)	0.1040 (0.0154)	0.1233 (0.016)	0.1284 (0.0159)	**0.1008 (0.0164)**
AUC_{min} (subpop.)	0.7717 (Female)	**0.7987 (Female)**	0.7631 (Female)	0.7733 (Female)	0.7749 (Female)
$CAUCI_{0.5}$	-	**0.0451**	0.0082	0.0247	0.0446
$CAUCI_{0.75}$	-	0.0327	0.0045	0.0103	**0.0340**

Table 5: Performance metrics for debiasing methods on CelebA predicting Y= Age, trained on partitioning with respect to S = Gender, and evaluated on a test set balanced across age and gender.

Metrics	Baseline (β=0)	AD (β=0.5)	IA (β=0)	TARA (β=0.5)	TARA+Filter (β=1.0)
acc	74.44 (0.96)	**76.45 (0.93)**	75.29 (0.95)	69.58 (1.01)	75.05 (0.95)
acc_{gap}	13.93 (1.89)	9.6 (1.85)	9.18 (1.88)	12.15 (2.00)	**7.25** (1.89)
acc_{min} (subpop.)	67.47 (Dark)	**71.65 (Dark)**	70.7 (Dark)	63.5 (Dark)	71.43 (Dark)
$CAI_{0.5}$	-	3.17	2.8	-1.54	**3.64**
$CAI_{0.75}$	-	3.75	3.78	0.12	**5.16**
AUC	0.8184 (0.0084)	**0.8614 (0.0076)**	0.8278 (0.0083)	0.8059 (0.0087)	0.8448 (0.0079)
AUC_{gap}	0.1003 (0.0153)	**0.0564 (0.0145)**	0.0714 (0.0153)	0.0683 (0.0167)	0.0661 (0.0158)
AUC_{min} (subpop.)	0.8038 (Dark)	**0.8462 (Dark)**	0.8201 (Dark)	0.7882 (Dark)	0.8132 (Dark)
$CAUCI_{0.5}$	-	**0.0435**	0.0191	0.0097	0.0303
$CAUCI_{0.75}$	-	**0.0437**	0.0240	0.0209	0.0323

Table 6: Performance metrics for debiasing methods on CelebA predicting Y= Age, trained on partitioning with respect to S = Skin Color, and evaluated on a test set balanced across age and skin color.

For results, we see a similar story to EyePACs in that TARA+Filter performs the best, though TARA was behind adversarial independence followed by intelligent augmentation and the baseline. Notably, with regards to skin color TARA performed worse than the baseline on CelebA in terms of overall accuracy, with only a slight decrease in the accuracy gap, which resulted in worse $CAI_{0.5}$. Results also echo the results for OSMI and EyePACS in demonstrating the usefulness of the conjunctive metrics in pointing out the debiasing methods that exhibit the best overall fairness performance (in this case achieved by TARA).

8 Discussion

Experimental Results: All methods tested, except the TARA with no filtering in Table 6, successfully show improvements in both overall accuracy and accuracy gap compared to the baseline as summarized by $CAI_{0.5}$. TARA with filtering was the best over almost all metrics for EyePACs, with TARA with no filtering coming in second and adversarial independence in third. For both CelebA experiments, adversarial independence and TARA with filtering were competitive with each other as the best of all methods, with the former method having better overall accuracy for skin color partitioning and better AUC metrics for gender partitioning. The latter had a better accuracy gap for both partitions and a better accuracy overall for the gender.

The test sets for each of the datasets were carefully balanced with respect to the given labels Y and protected factors S, while also including a completely excluded subpopulation. The training sets, on the other hand, had natural selection bias inherent in the dataset and contained no samples from the excluded subpopulation. As a result, the distributions of the train and test sets differed with respect to Y and S. However, for each of these datasets Y should have been independent of S, which suggested that enforcing fairness across S should have *increased* classification accuracy because independence was being enforced Wick, panda, and Tristan (2019). This was further reinforced by the empirical results we presented on each of the datasets, which showed accuracy improvements for methods that enforced independence between Y and S, such as adversarial independence and TARA.

For all experiments, intelligent augmentation was outperformed by adversarial independence and at least one variant of TARA, though it still had improvements in the accuracy gap even when compared to adversarial independence. However, adversarial independence was sensitive to β, as using $\beta = 1.0$ for both adversarial independence and the non-filtering TARA resulted in overall accuracies of approximately 60% with similar accuracy gaps to the current implementation. Another consideration was that we chose to use a different representation, the activations before the last linear layer, than the original implementation using the *SoftMax* values, which did not affect the overall accuracy or accuracy gap significantly, possibly due to the more complex model used. In contrast, as intelligent augmentation only affects the dataset, training was less complicated, and could extend to other domains, such as segmentation, an avenue we leave for future work. A potential alternative to intelligent augmentation, oversampling, was not considered as a viable option due to the proclivity of models overfitting to subpopulations with limited number of examples, particularly in small datasets as are common in fields such as medical imaging. To improve model generalization on underrepresented subpopulations our approach addresses the overfitting problem by artificially creating new samples of said subpopulations instead of oversampling its small pool of samples.

When we debias with respect to ITA as a sensitive factor, we note from Table 3 that the resulting system succeeds in achieving parity for ITA. Furthermore, we note that this system succeeds at achieving parity with regard to another sensitive factor (i.e., race) for which ITA was a proxy; see Table 4.

This suggests that the fundus pigmentation, out of all the factors that comprise the presumed race, is the principal factor causing bias, and that ITA, which did not require a specialist to acquire, can be used as a surrogate for the difficult to acquire demographic information. In sum, we posit that ITA is a robust proxy measure for the debiasing method with race as a sensitive factor for EyesPACS. Future work could include a closer examination of robustness of debiasing methods to mismatch in sensitive factors.

Note, our construction for domain generalization is not held in Table 4 as some Black referable individuals were present in the training set. This may explain why the baseline accuracy was higher than the baseline from P. Burlina, Joshi, et al. (2020), though it was a different training dataset. On the other hand, the factor omitted did have varying effects, as to what occurs between the two CelebA experiments, with the omission of older females having a larger impact on the baseline for the accuracy gap and overall accuracy compared to excluding older darker individuals.

For the tabular data, we performed a set of experiments on the OSMI dataset to understand the network configuration that maximized fairness under our generalization experiment. The adversarial independence method (*EMB+AD*) was found to maximize overall accuracy, $CAI_{0.5}$, $CAUCI_{0.5}$, and minimum accuracy of the subpopulations that were under represented for both protected factors (female and old). *EMB+AD* was also shown to target debiasing of the specific factors instead of acting as general regularization like the *Noise* debiasing approach. The overall results for gender and age debiasing suggest that using adversarial learning to reduce the prediction network's bias towards a protected class can reduce bias and encourage it to identify more optimal network weights. A more extensive ablation study can be found in the Appendix.

Metrics: We introduced novel metrics, the conjunctive accuracy improvement CAI_{α}, a combination of overall accuracy and accuracy gap, and the compound AUC improvement, a combination of AUC and AUC gap, or $CAUCI_{\alpha}$. As these series of experiments on tabular and image data had demonstrated, these novel metrics were generally useful in reflecting wholistic improvements in fairness. For situations where each individual metric was the best, our new metrics reflect this case of Pareto optimality. For more ambiguous cases where one was superior while another was not, such as the accuracy metrics in Table 6, these examples show the need for a single metric that can help assess best overall performance but also reflect the desired policy objective and ethical imperatives. Table 1 shows a situation where the specific value of α changed the ranking of the methods, indicating the importance of what α is set to. Regarding AUC metrics, showing the effect of methods on the AUC of individual subpopulations appears to be an important consideration for certain scenarios. Unlike accuracy, AUC is not decomposed directly into the AUC on subpopulations, as the AUC represents the capacity of the classifier to choose a specific true positive rate and false positive rate. Consequently, the AUC gap (measuring the disparity in tuning each subpopulation) and minimum AUC (indicating how restricted the worse-off calibration was) were suited to cases where protected factor ground-truth exists in order to calibrate each subpopulation individually. While CAI_{α} and $CAUCI_{\alpha}$ may indeed be useful, guidance for how to set the parameter α, or how it relates to legal concepts such as Disparate Impact is deferred to future work.

Future Work: Future work will explore the fact that adversarial independence may still produce biased results with regard to a protected factor that was not tested against or has yet to be considered protected. As intelligent augmentation is independent of the downstream task, expanding to other domains, such as segmentation, is also of interest.

9 Conclusion

This study proposed TARA, a novel approach to debiasing using joint alteration of data representation and training, aiming to address both sources of bias, conditional dependence and data imbalance. We showed that it outperformed competing methods. We introduced novel fairness metrics addressing some issues in current bias metrics, as a basis for future investigations and discussions between AI scientists, ethicists and policy makers regarding how to best compare and assess debiasing.

Appendix

Approach	Goal	Methods	Application
Adversarial Independence	Ensure conditional independence from a protected factor	Adversarial learning	Tabular records and Images
Intelligent Augmentation	Mitigate data imbalance for protected subpopulations	Generative methods + latent space manipulation	Images
TARA	Mitigate data imbalance and ensure conditional independence	Generative methods and Adversarial learning	Images

Table 7: Methods Summary

Additional details on fairness definitions; nomenclature; methods; datasets; preprocessing and implementation; shared code and data; and supplemental discussion items are described below.

Methods Summary:

A summary of our methods is described in Table 7.

Nomenclature and Definitions: The following includes some of the most commonly used formal definitions of fairness (for more, see also Mehrabi et al. (2019)).

We denote the protected factor(s) as $S = s$, the classifier's decision for the outcome as $\hat{Y} = \hat{y}$, and the true outcome, or the underlying true label, depending on context, as $Y = y$. We focus on the following definitions of fairness Hardt et al. (2016):

Demographic Parity: Demographic parity states that all subpopulations should have a positive decision (e.g. credit approval) at equal rates. Mathematically, demographic parity states that:

$$P(\hat{Y} = \hat{y}|S = s) = P(\hat{Y} = \hat{y}), \forall s, \hat{y}. \tag{6}$$

Demographic parity may not be appropriate in situations where a fundamental correlation exists between Y and S: consider, for example, a health condition that is predominant in certain age groups, e.g. age-related macular degeneration P. Burlina et al. (2011).

Equality of Odds: On the other hand, equality of odds expresses that a predictive model must produce predictions that are conditionally independent of protected factors given the true outcome, where:

$$P(\hat{Y} = \hat{y}|S = s, Y = y) = P(\hat{Y} = \hat{y}|Y = y), \forall s, y, \hat{y}. \tag{7}$$

Unlike demographic parity, the conditional independence ensures that, when Y has a causal relationship with S, the performance of the prediction being correct ($\hat{Y} = Y$) is not adversely affected by the strict condition of independence.

Equality of Opportunity: Equality of opportunity further relaxes the equality constraints in Eq. 7 by dictating that a model must produce predictions that are independent of a protected factor, for a specific (and not necessarily all) values of the label y, such that:

$$P(\hat{Y} = \hat{y}|Y = y) = P(\hat{Y} = \hat{y}|S = s, Y = y), \forall s, \hat{y}. \tag{8}$$

Equality of Performance: Conditional independence stated above has corollary implications for performance (error rates) of the classifier. Take the example of a binary classification problem, then Eq. 8, when stated for $y = 1$, ensures an equal *true positive* rate exists across all protected factors s. However, unlike equality of odds, it does not necessarily require an equal *false negative* rate across all s. Equality of Odds however does. Demographic parity, equality of odds, and equality of opportunity have served as the foundation for recent advances in AI bias mitigation. For our study we adopt the stricter goals of equality of odds and measure success using commonly adopted metrics as well as novel proposed metrics that are consistent with this goal.

Next, we provide extended details on each dataset.

OSMI Mental Health Data Details: The OSMI Mental Health in Tech Survey 2014 LTD (2014) was released on Kaggle to encourage evaluation of mental health in technology industry and how mental health relates to job related factors. Rado and Neagu (2019) and Sharma, Anand, Jaiswal, and Goyal (2018) used the OSMI Mental Health Survey 2014 to predict the likelihood a given individual had sought mental health treatment. Approaches evaluated ranged from decision trees to neural networks. Both studies claimed accuracies ranging from 79% to 98%. However, evaluation datasets were not standardized across either work. In 2016, OSMI compiled a new mental health survey which included a more extensive questionnaire and more samples. The 2016 OSMI dataset LTD (2016) included questions asking whether a person has been diagnosed with a mental illness, if so which mental illness, and whether they had sought treatment for a mental illness. Reddy, Thota, and Dharun (2018) developed a series of off-the-shelf machine learning models that used this dataset to try to predict whether an employee had treatment for mental health related disorders in the past.

Our work explores the gender and age bias present in the OSMI Mental Health in Tech Survey 2016 dataset (denoted as OSMI) and whether deep learning models can be trained to mitigate bias using adversarial independence. For the binary classification task of estimating whether a person sought treatment for a mental illness the prediction network F trained using the set of tabular input features, X, listed in Table 8. Unlike many of the employer-specific features from the OSMI dataset, these features were selected because they best corresponded to the task of estimating if a person sought mental health treatment. We did, however, ignore features related to personal or family history of mental illness as they were both overly correlated to the likelihood of a person seeking treatment according to Rado and Neagu (2019) and Sharma et al. (2018).

Mental Health Feature	# Classes
Age	53
Gender	2
Benefits and insurance coverage	2
Care options	2
Anonymous discussions	2
Interference with work performance	2
Medical leave availability	2
Perceived negative impact of discussing mental health	2

Table 8: OSMI Features

ITA Note: As the ITA is computed per pixel, care was needed in determining which areas were used to calculate the ITA, as it might have been adversely be affected by light artifacts and lighting in general.

EyePACs Data Details: The EyePACs dataset contained 88,692 images for 44,346 participants, with two images, the left and right fundus, for each participant. We resized these images to 256x256 pixels after being cropped to the outline of the fundus, and the labels binarized such that 0 and 1 were "not referable" and 2, 3, and 4 were "referable" for the disease. To compute the mask for ITA on each image, we ran a one-class SVM, with a RBF kernel and a upper bound of 80% for the training errors, on the luminance dimension, with each non-background pixel as a data point, to mask any anomalous areas such as light artifacts along with the background. The ITA was then computed per pixel and averaged over the non-masked area.

CelebA Data Details: These images were 218x178 pixels, and were preprocessed by taking a 128x128 crop with the center at (121, 89). Older dark skinned females were the smallest subpopulation consisting of 1,380 images, where an ITA less than 28 denoted dark skinned, whereas younger light skinned females were the largest at 93,477 images.

We computed the ITA in a similar manner to Merler et al. (2019), where we used a skin segmentation step to filter out invalid pixels, and used a landmark detector to segment out the chin, cheeks and forehead areas. We diverged slightly in that we used Gaussian blur on the image of ITA values (with a kernel size of 11) and we chose the median ITA value per region that we then averaged over to get the final value.

Metrics	Baseline (EMB)	Cat	EMB+Noise	Freeze EMB+ADDP	Cat+ADDP	EMB+ADDP	EMB+AD
acc	63.22 (5.07)	50.57 (5.25)	67.82 (4.91)	76.15 (4.48)	56.03 (5.21)	77.59 (4.38)	**82.18 (4.02)**
acc_{gap}	20.69 (0.87)	**2.29 (0.01)**	24.14 (1.42)	6.32 (0.58)	2.87 (0.05)	3.45 (0.34)	4.59 (0.58)
acc_{min} (subpop.)	52.87 (F)	49.43 (M)	55.75 (F)	72.99 (M)	54.60 (M)	75.86 (M)	**79.89 (M)**
$CAI_{0.5}$	-	2.88	0.58	13.65	5.32	15.81	**17.53**
$CAI_{0.75}$	-	10.64	-1.44	14.01	11.57	16.52	**16.82**
AUC	0.7213 (0.0471)	0.5807 (0.0518)	0.8030 (0.0418)	0.8407 (0.0384)	0.5905 (0.0517)	**0.8633 (0.0361)**	0.8592 (0.0365)
AUC_{gap}	0.0606 (0.0087)	0.0854 (0.0026)	0.0610 (0.0102)	0.1030 (0.0146)	0.0663 (0.0019)	**0.0537 (0.0085)**	0.0805 (0.0129)
AUC_{min} (subpop.)	0.8171 (M)	0.5580 (F)	**0.8430 (M)**	0.7894 (M)	0.5615 (M)	0.8380 (M)	0.8236 (M)
$CAUCI_{0.5}$	-	-0.0827	0.0407	0.0385	-0.0683	**0.0745**	0.0590
$CAUCI_{0.75}$	-	-0.0538	0.0201	-0.0020	-0.0370	**0.0407**	0.0196

Table 9: OSMI Demographic parity results. Performance metrics for debiasing methods on OSMI predicting Y= sought mental health treatment, trained on partitioning with respect to S = Gender, and evaluated on a test set balanced across treatment status and gender (M/F). Methods include: FC network (Cat), FC with embeddings network (EMB), noise debias (Noise), adversarial debias with demographic parity (ADDP), adversarial debias with equality of odds (AD), and freeze training (Freeze).

In addition to what was reported in the main body of the paper, we also performed more experiments on OSMI, which provided additional insights in the workings of adversarial independence. These are detailed next.

Extended OSMI Experiments and Discussion:

We performed an ablation study on the network architecture depicted in Figure 6, which we evaluated on the OSMI dataset to determine the impact of adversarial independence (without intelligent augmentation) on fairness. We experimented with replacing the embedding layers (methods containing *Cat*) and removing the adversarial module altogether (without *AD* or *ADDP*). The *Cat* methods concatenated each of the input features into a single fully connected layer. Our proposed adversarial independence debiasing approach (*EMB+AD*) included a prediction network with an embedding for each of the input features concatenated and forwarded to a fully connected layer, while

Metrics	Baseline (EMB)	Cat	EMB+Noise	Freeze EMB+ADDP	Cat+ADDP	EMB+ADDP	EMB+AD
acc	68.36 (4.29)	63.05 (4.45)	75.00 (3.99)	73.89 (4.05)	57.08 (4.56)	79.42 (3.73)	**80.75 (3.63)**
acc_{gap}	21.68 (1.15)	19.03 (0.68)	18.14 (1.41)	16.81 (1.22)	12.39 (0.24)	3.99 (0.38)	**1.33 (0.14)**
acc_{min} (subpop.)	57.52 (O)	53.54 (O)	65.93 (Y)	65.49 (O)	50.88 (Y)	77.43 (O)	**80.09 (O)**
$CAI_{0.5}$	-	-1.33	5.09	5.20	-1.00	14.38	**16.37**
$CAI_{0.75}$	-	0.66	4.32	5.04	4.15	16.03	**18.36**
AUC	0.7658 (0.0444)	0.7249 (0.0412)	0.8211 (0.0353)	0.8458 (0.0333)	0.6171 (0.0448)	**0.8928 (0.0285)**	0.8787 (0.0301)
AUC_{gap}	0.0014 (0.0002)	0.0360 (0.0030)	**0.0003 (0.0001)**	0.0090 (0.0014)	0.0636 (0.0024)	0.0310 (0.0054)	0.0139 (0.0021)
AUC_{min} (subpop.)	0.8648 (O)	0.7567 (Y)	0.8780 (Y)	0.8754 (Y)	0.6059 (Y)	**0.8826 (Y)**	0.8761 (Y)
$CAUCI_{0.5}$	-	-0.03775	0.0282	0.0362	-0.1055	0.0487	**0.0502**
$CAUCI_{0.75}$	-	-0.0362	0.0147	0.0143	-0.0838	0.0096	**0.0189**

Table 10: OSMI Demographic parity results. Performance metrics for debiasing methods on OSMI predicting Y= sought mental health treatment, trained on partitioning with respect to S = Age, and evaluated on a test set balanced across treatment status and age (O/Y for Older/Younger).

the adversarial debiasing network was a single fully connected layer (*AD*). Moreover, we also evaluated an alternative adversarial module constructed based on demographic parity (*ADDP*) instead of equality of odds, meaning the adversarial module only received the prediction module's logits as an input and did not use the prediction task's target label.

First, we examined the gender debiasing results shown in Table 9 reflecting a 12.65% increase in overall accuracy when using and embedding-based prediction network (*EMB*) compared with the concatenation network without adversarial debiasing (*Cat*) performance. However, the *Cat* network had the smallest accuracy gap among all methods evaluated, but at the cost of near random accuracy. All the *Cat*-based methods suffered from poor accuracy which is likely attributed to generalizing poorly to the unseen class (females seeking treatment) due to the limited network structure (a single fully connected layer), unlike the *EMB*-based methods which also had the contribution of embedding layers. Together these deficiencies in the *Cat*-based methods were visible in both the CAI_α and $CAUCI_\alpha$ being the lowest. These same experiments were also conducted for the case debiasing the age protected factor (shown in Table 10), which also showed the importance of using embeddings for OSMI. Overall, embedded features (*EMB* methods) maximized overall accuracy, AUC, and minimum accuracy of the subpopulations (in this case female) that was under represented.

Next, the impact of using adversarial independence based on demographic parity (*ADDP*) instead of equality of odds (*AD*) was examined. The results suggest that *ADDP* was almost as effective as *AD* in terms of accuracy-based metrics, as best represented in the CAI_α for $alpha = 0.5$ and 0.75. However, for debiasing gender the *ADDP* was superior to the *AD* method in terms of the $CAUCI_\alpha$. Similarly, for the case of debiasing age the *EMB+AD* and *EMB+ADDP* were competitive with one another with respect to the AUC-based metrics. To better understand which method performed better overall in terms of AUC, we examined the compound AUC improvement ($CAUCI_\alpha$) which showed *EMB+AD* was marginally better for $\alpha = 0.5$ and substantially better for $\alpha = 0.75$, which weighed the AUC gap more heavily. The results for gender and age debiasing suggested that using adversarial learning based on equality of odds (*AD*) to reduce the prediction network's bias towards a protected class could reduce bias and encourage it to identify more optimal network weights.

Last, we evaluated the impact of an alternative to training the adversarial debiasing module, where the adversarial network (*ADDP*) was frozen when training the prediction network and then the prediction network was frozen when training the adversarial network (denoted as *Freeze EMB+ADDP*). Alternating freeze training allowed each network to be optimized individually in order to potentially improve individual performance without affecting the other. However, the results in Tables 9 and 10 indicate that freeze training performed worse than non-freeze training in $CAUCI_\alpha$ and CAI_α for debiasing gender and age. As a result, optimizing both the prediction and adversarial networks without freezing parameters was shown to be preferable.

OSMI Experiments Implementation Details: OSMI was partitioned by randomly shuffling and spliting into 70%, 10%, and 20% partitions corresponding to train, validation, and test sets, respectively. As mentioned in Section 6 Domain Generalization, categories that had the fewest members (Gender debiasing = female seeking treatment, Age debiasing = older seeking treatment) were relegated to the test partition only. The test split was constructed in a manner where each member of the protected class had equal representation. Each of the models being evaluated were trained on the respective dataset for up to 100 epochs, with early stopping triggering when the validation loss had not decreased for 10 epochs. After some experimentation, we found that the adversarial loss balancing term β resulted in the best performance when set to 1. The network weights corresponding to the smallest validation loss were retained for evaluation on the test set. We used the Adam optimizer with a learning rate scheduler set to reduce the learning rate by a factor of 0.1 for every 10 epochs the training loss plateaued. All adversarial modules were pre-trained for 100 epochs with the prediction module frozen. The pre-training procedure was designed to help reduce the likelihood of poor initialization of network weights for the debiasing module.

Image-based Experiments Implementation details: For image experiments, we used a ResNet50 classifier pretrained on ImageNet with the final linear layer replaced with a randomly initialized layer with an output dimension of 2. The adversarial network consisted of four fully connected layers of width 512 with LeakyRelu activations and $alpha = 0.01$. We used stopping criterion on the lowest validation loss, with a patience of 5 epochs for all methods. Outside of dataset specific preprocessing, we used Imagenet normalization on the input, and resized to 224x224 using a bicubic interpolation. Additionally, we used SGD with a learning rate of 0.001 and Nesterov momention of 0.9, and AdamW with a learning rate of 0.005 for the adversary.

EyePACs Experiments: Our training dataset was made up of 10,346 referable lighter skin images ($ITA = 0, DR = 1$), 5,173 non-referable lighter skin images ($ITA = 0, DR = 0$) and 5,173 non-referable darker skinned images ($ITA = 1, DR = 0$). One finding of note was that the lighter skin subpopulation transitioned from performing better than the darker skin subpopulation to performing worse when using adversarial independence methods without filtering.

CelebA Experiments: We used a total training dataset size of 48,000 for each partitioning, where we again kept each training dataset balanced across our target factor, age, and our protected factors. For gender, there were 24,000 older male images, 12,000 younger male and 12,000 younger female images. For ITA, there were 24,000 older dark skinned images, 12,000 younger lighter skinned images, and 12,000 younger darker skinned images.

Metrics: Note, in some cases the minimum accuracy performance flipped between subpopulations for the baseline versus the methods with adversarial independence. Given this occurrence, despite the minimum accuracy avoiding selection of a coefficient α, compared to CAI_{alpha}, it nevertheless still requires scrutiny from ethicists and policy makers to ensure the metric does not give false understanding of which subpopulation is underrepresented.

References

Alvi, M., Zisserman, A., & Nellåker, C. (2018). Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops.*
2

Beutel, A., Chen, J., Zhao, Z., & Chi, E. H. (2017). Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*.
2

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing systems* (pp. 4349–4357).
2

Burlina, P., Freund, D. E., Dupas, B., & Bressler, N. (2011). Automatic screening of age-related macular degeneration and retinal abnormalities. In *2011 annual international Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 3962–3966).
2, 18

Burlina, P., Joshi, N., Paul, W., Pacheco, K. D., & Bressler, N. M. (2020). Addressing artificial intelligence bias in retinal disease diagnostics. *arXiv preprint arXiv:2004.13515 and Translational Vision Science and Technology (accepted)*.
2, 8, 10, 11, 14, 16

Burlina, P., Paul, W., Mathew, P., Joshi, N., Pacheco, K. D., & Bressler, N. M. (2020). Low-shot deep learning of diabetic retinopathy with potential applications to address artificial intelligence bias in retinal diagnostics and rare ophthalmic diseases. *JAMA Ophthalmology*.
2

Burlina, P. M., Joshi, N., Pacheco, K. D., Liu, T. A., & Bressler, N. M. (2019). Assessment of deep generative models for high-resolution synthetic retinal image generation of age-related macular degeneration. *JAMA Ophthalmology*, *137*(3), 258–264.
3

Burlina, P. M., Joshi, N. J., Mathew, P. A., Paul, W., Rebman, A. W., & Aucott, J. N. (2020). Ai-based detection of erythema migrans and disambiguation against other skin lesions. *Computers in Biology and Medicine*, *125*, 103977.
3

Burlina, P. M., Joshi, N. J., Ng, E., Billings, S. D., Rebman, A. W., & Aucott, J. N.

(2019). Automated detection of erythema migrans and other confounding skin lesions via deep learning. *Computers in biology and medicine*, *105*, 151–156.
2

Carlini, N., & Wagner, D. (2017). Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security* (pp. 3–14).
2

Christiansen, S. L., Iverson, C., Flanagin, A., Livingston, E. H., Fischer, L., & Manno, C. (2020). *AMA manual of style: A guide for authors and editors*. JAMA Network, Oxford University Press.
10

Edwards, H., & Storkey, A. (2015). Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*.
3

EyePACS. (2015). *Diabetic retinopathy detection.* (data retrieved from Kaggle, https://www.kaggle.com/c/diabetic-retinopathy-detection)
10

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, *17*(1), 2096–2030.
2

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* (pp. 2672–2680).
2

Grover, A., Song, J., Kapoor, A., Tran, K., Agarwal, A., Horvitz, E. J., & Ermon, S. (2019). Bias correction of learned generative models using likelihood-free importance weighting. In *Advances in Neural Information Processing Systems* (pp. 11058–11070).
3

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* (pp. 3315–3323).
5, 18

Hu, W., Niu, G., Sato, I., & Sugiyama, M. (2018). Does distributionally robust supervised learning give robust classifiers? In *International conference on machine learning* (pp. 2029–2037).
2

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4401–4410).
3

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
3

Kinyanjui, N. M., Odonga, T., Cintas, C., Codella, N. C., Panda, R., Sattigeri, P., &

Varshney, K. R. (2020). Fairness of classifiers across skin tones in dermatology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 320–329).
2, 11, 12

Locatello, F., Abbati, G., Rainforth, T., Bauer, S., Schölkopf, B., & Bachem, O. (2019). On the fairness of disentangled representations. In *Neurips.*
3

Louizos, C., Swersky, K., Li, Y., Welling, M., & Zemel, R. (2015). The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*.
3

LTD, O. S. M. I. (2014). *Osmi mental health in tech survey 2014.* (data retrieved from Kaggle, https://www.kaggle.com/osmi/mental-health-in-tech-survey)
19

LTD, O. S. M. I. (2016). *Osmi mental health in tech survey 2016.* (data retrieved from Kaggle, https://www.kaggle.com/osmi/mental-health-in-tech-2016)
9, 19

Madras, D., Creager, E., Pitassi, T., & Zemel, R. (2018). Learning adversarially fair and transferable representations. In *International conference on machine learning* (pp. 3384–3393).
3

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.
5, 18

Merler, M., Ratha, N., Feris, R. S., & Smith, J. R. (2019). Diversity in faces. *arXiv preprint arXiv:1901.10436*.
11, 20

Oeldorf, C., & Spanakis, G. (2019). Loganv2: Conditional style-based logo generation with generative adversarial networks. In *2019 18th IEEE international Conference on Machine Learning and Applications (ICMLA)* (pp. 462–468).
8

Paul, W., Wang, I.-J., Alajaji, F., & Burlina, P. (2021). Unsupervised semantic attribute discovery and control in generative models. *Neural Computation*, *33*(3), 802—826.
3, 4

Pekala, M., Joshi, N., Liu, T. A., Bressler, N. M., DeBuc, D. C., & Burlina, P. (2019). Deep learning based retinal oct segmentation. *Computers in Biology and Medicine*, *114*, 103445.
2

Poplin, R., Varadarajan, A. V., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S., … Webster, D. R. (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, *2*(3), 158.
6

Prost, F., Qian, H., Chen, Q., Chi, E. H., Chen, J., & Beutel, A. (2019). Toward a

better trade-off between performance and fairness with kernel-based distribution matching. *arXiv preprint arXiv:1910.11779*.
2

Quadrianto, N., Sharmanska, V., & Thomas, O. (2019). Discovering fair representations in the data domain. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8227–8236).
3

Rado, O., & Neagu, D. (2019). On selection of optimal classifiers. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence* (pp. 494–499).
19

Ravi, S., & Larochelle, H. (2016). Optimization as a model for few-shot learning. 2

Reddy, U. S., Thota, A. V., & Dharun, A. (2018). Machine learning techniques for stress prediction in working employees. In *2018 IEEE international Conference on Computational Intelligence and Computing Research (ICCIC)* (pp. 1–4).
19

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.
2

Sagawa, S., Koh, P. W., Hashimoto, T. B., & Liang, P. (2019). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.
2

Sharma, S., Anand, S., Jaiswal, A., & Goyal, M. (2018). Predictive analysis using classification techniques in healthcare domain. 19

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)* (pp. 3–18).
2

Song, J., Kalluri, P., Grover, A., Zhao, S., & Ermon, S. (2019). Learning controllable fair representations. In *The 22nd international Conference on Artificial Intelligence and Statistics* (pp. 2164–2173).
3

Wadsworth, C., Vera, F., & Piech, C. (2018). Achieving fairness through adversarial learning: an application to recidivism prediction. In *Conference on Fairness, Accountability, and Transparency in Machine Learning (FATML).*
3

Wang, T., Zhao, J., Yatskar, M., Chang, K.-W., & Ordonez, V. (2019). Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE international Conference on Computer Vision* (pp. 5310–5319).
3

Wick, M., panda, s., & Tristan, J.-B. (2019). Unlocking fairness: a trade-off revisited. In H. Wallach, H. Larochelle, A. Beygelz-

imer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2019/file/373e4c5d8edfa8b74fd4b6791d0cf6dc-Paper.pdf
16

Wilkes, M., Wright, C. Y., du Plessis, J. L., & Reeder, A. (2015). Fitzpatrick skin type, individual typology angle, and melanin index in an african population: steps toward universally applicable skin photosensitivity assessments. *JAMA Dermatology*, *151*(8), 902–903.
10

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In *International Conference on Machine Learning* (pp. 325–333).
2

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 335–340).
3, 5

Zhao, S., Song, J., & Ermon, S. (2017). Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*.
3