






A Differentiable Distance Approximation for Fairer Image Classification

Nicholas Rosa¹(✉)() , Tom Drummond^{1,2}() , and Mehrtash Harandi¹()

¹ Monash University, Clayton, Australia

nicholas.rosa@monash.edu

² The University of Melbourne, Melbourne, Australia

Abstract. Naïvely trained AI models can be heavily biased. This can be particularly problematic when the biases involve legally or morally protected attributes such as ethnic background, age or gender. Existing solutions to this problem come at the cost of extra computation, unstable adversarial optimisation or have losses on the feature space structure that are disconnected from fairness measures and only loosely generalise to fairness. In this work we propose a differentiable approximation of the variance of demographics, a metric that can be used to measure the bias, or unfairness, in an AI model. Our approximation can be optimised alongside the regular training objective which eliminates the need for any extra models during training and directly improves the fairness of the regularised models. We demonstrate that our approach improves the fairness of AI models in varied task and dataset scenarios, whilst still maintaining a high level of classification accuracy. Code is available at https://bitbucket.org/nelliottrosa/base_fairness.

1 Introduction

In recent times, the use of Artificial Intelligence (AI) has permeated many processes that are used to make important decisions, such as filtering applicants for jobs, deciding if an applicant should receive credit and recognizing people in images [15, 25]. Given this, it is essential to ensure that AI-driven models are not exhibiting behaviour which is morally or legally undesirable. In AI, data is a collection of attributes, which can either be explicit (*e.g.* labels) or implicit (*e.g.* information from an image). Some of these attributes are referred to as *protected* attributes as they should not be used to discriminate (*e.g.* gender, race or age). However, it has been shown numerous times that AI models which are naïvely trained are biased against one or more of these protected attributes, as they exhibit lower accuracy for some demographics [4, 11, 19]. This behaviour is discriminatory against these demographics and is *morally or legally undesirable*, or simply unfair. There are two common sources of unfair behaviour that can present itself in AI systems. The first source is biases that are present in

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-26351-4_14.

the data used for training AI models. Biases in the data with respect to protected attributes can cause an AI model trained upon that data to discriminate against the protected attribute [2]. For example, if a dataset used to train a facial recognition model for unlocking doors, only contains images of men (*i.e.* bias with respect to gender) then the learned model will not accurately recognise and admit women (*i.e.* unfair behaviour). The second source of bias is due to some values or demographics of a protected attribute being inherently harder for AI to recognize than others. For example, it has been shown that even when training with a balanced dataset, faces with a darker skin tone are harder to recognize for facial recognition algorithms [28].

Various solutions to the fairness problem have been proposed. We focus on algorithmic in-processing methods for reducing the bias [5, 6, 8–10, 14, 29, 33]. In-processing aims to address the bias of a model by applying an extra objective during training which makes the model bias-aware and consequentially learns a fairer model. In-processing has proven to be quite effective at reducing the unfair behaviour of AI. However, in-processing methods often include extra models which can increase training cost and complexity [16]; use adversarial training [6, 8, 9, 33] which has proven to be notoriously unstable [24] or make assumptions about the representation space of the model which may not hold in all cases [10, 21, 32]. Creating fair AI models is particularly difficult in the computer vision domain as any problems with extra computational cost and complexity are exacerbated by the large models utilized. Additionally the high dimensionality of images means they can contain many implicit attributes, which are often highly correlated to each other and to protected attributes. Disentangling the implicit factors is extra challenging in these cases.

In this paper, we introduce **B**ias **A**ccuracy **S**tandard deviation **E**stimation or **BASE**, a novel fairness algorithm, which optimizes a differentiable approximation of the fairness metric *standard deviation of accuracy across demographics* (σ_{acc}) to learn an AI model which is fair with respect to *equalized odds* (EO). Models that exhibit a low standard deviation of accuracy across demographics or variance of demographics have the property of equal performance on a target task regardless of the demographic of the protected attribute. For example, a facial recognition model which has low variance of demographics for ethnicity, is equally likely to correctly recognize the identity of a person from an image regardless of their ethnicity. Reducing the variance of demographics of a model makes it fairer w.r.t. EO. However, for an AI model that is trained with gradient based optimization the variance of demographics is difficult to use. This is due to the accuracy of a single sample - an integral part of the variance of demographics (Sect. 2.3) - having an undefined gradient at 0 and being 0 everywhere else, which leads to zero influence on the model parameters. **BASE** overcomes this difficulty by instead using a sigmoid based approximation of accuracy which we call *soft-accuracy* inside the variance of demographics metric. This approach has multiple advantages. Firstly computational efficiency, for example, training a classifier on images with **BASE** incurs only the extra computation of calculating the variance of demographics. Compare this to training a classifier with knowledge distillation [16] or adversarial debiasing [33], where additional models are used which incur extra memory usage for the model parameters and gradi-

ents, alongside with extra computation for the forward pass of the additional model. Secondly, BASE makes no assumptions about the representation space. The model will automatically learn the representation space structure required to reduce the variance of demographics. Furthermore, due to its simplicity BASE can be combined with other solutions.

To summarize the main contributions of our work are:

- Provide a novel method for improving the fairness on AI models trained with gradient based optimization, that increases algorithmic simplicity and does not rely on training additional models (Sect. 3.1).
- Show that our method is competitive with and in some cases outperforms current state-of-the-art fair image classifiers when using either a biased dataset or an unbiased dataset (Sect. 4.4, Sect. 4.4).
- Show that our method increasingly outperforms the fairness of a naive classifier when exposed to increasingly biased training sets in which target and protected attributes are strongly correlated. Our method also achieves higher over-all accuracy on heavily biased datasets (Sect. 4.4).

2 Related Work and Preliminaries

Fair AI has received increasing attention in the past few years and a varied range of solutions has been proposed. Algorithmic methods for reducing the bias can be broken down into three main categories based upon when they apply their fairness constraint. *Pre*-processing methods aim to change the distribution of the data used for training such that a fairer model is produced. These methods include re-sampling, which changes the sampling rate of data during training to ensure each protected class is equally represented [1, 23, 26] and augmentation methods which add synthetic data to the dataset [3, 22, 31, 34] to balance the protected classes. The second class of methods, *post*-processing methods, aim to adjust the prediction after the fact to compensate for the bias [30]. *Pre*-processing and *post*-processing have some major drawbacks. *Pre*-processing only addresses the bias in the dataset and the inherent difficulties of some demographics can still cause a biased model [28, 29]. On the other hand *post*-processing methods require that protected attribute labels to be known at inference time or assume that the target and protected attribute are independent [30]. Our method is related to the final category of *in*-processing, which is discussed further below. *In*-processing methods typically run under a constrained optimization scheme where a loss penalty or a special construction of the AI model is used to reduce the bias during optimization.

2.1 In-processing for Fair Classification

Like many machine learning tasks, the fairness problem is difficult to optimize directly and adversarial training became a common method to create fair representations and predictors [6, 8, 9, 29, 33]. These methods use an adversarial model,

or adversary, whose purpose is to learn the relationship between the predictor and the protected attribute. The output of the adversary is then used to enforce a fairness constraint upon the predictor. This is achieved either by gradient reversal of the adversary or by maximising the entropy of the adversaries predictions. If a strong adversary is unable to determine a relationship between the predictor and the protected attribute then fairness of the predictor can be guaranteed [33]

Other constrained optimization methods have been proposed and their approaches vary greatly. Gong *et al.* [10] minimize the variance of sample density across different demographics within the representation space. Cho *et al.* [5] use a kernel density estimate to approximate the conditional distributions used for measuring fairness in a differentiable manner. Hwang *et al.* [14] reduce the Wasserstein distance between protected groups within the representation space. Finally, in a work most similar to our own Shen *et al.* [27] use cross-entropy loss as a proxy for probability during training to optimise for fairness. Our method differs in two main aspects; our objective directly considers the two elements of the models output vector responsible for determining accuracy and we evaluate our work in the computer vision domain.

2.2 Problem Definition

The ultimate goal of fair machine learning is to create predictors which contain no bias. There is, however, many different forms of bias that can present themselves and as a consequence there are multiple different definitions of fairness. The three most common definitions are Demographic parity [33], Equalized Odds [12] and Equalized Opportunity [12]. In the following section A , \hat{Y} and Y are random variables which represent the protected attribute, the output of a predictor and the true value of the target attribute respectively.

Demographic Parity. Demographic parity is the simplest form of fairness since it only considers the output of the predictor and the protected attribute. A predictor satisfies demographic parity when its output is independent of the protected attribute. That is $\forall a \in \mathcal{A}; \Pr(\hat{Y} = \hat{y} | A = a) = \Pr(\hat{Y} = \hat{y})$. However, this definition does not always allow for perfect classification [12]. If there is any correlation between the protected attribute and the target task then maintaining independence forces a reduction in performance. For example, if we learned a predictor for university admittance with age as a protected attribute, then achieving demographic parity would require our predictor to admit young children with the same probability as those who had just finished high school, regardless of each individuals suitability.

Equalized Odds. Equalized Odds is another definition of fairness that is more commonly applied for computer vision tasks. A predictor satisfies equalized odds when its output is conditionally independent of the protected attribute for all classes of the target class. That is $\forall y \in \mathcal{Y}, \forall a, a' \in \mathcal{A}, \Pr(\hat{Y} = y | A = a, Y = y) = \Pr(\hat{Y} = y | A = a', Y = y)$. This definition allows us to maintain performance

as it is satisfied when a predictor achieves the same level of accuracy for each demographic of the protected attribute.

Equalized Opportunity. Equalized Opportunity is a special case of equalized odds for which there is a class of the target task $y_+ \in \mathcal{Y}$ that confers advantage, *e.g.*, to receive a loan or be hired for a job. It is a relaxation of equalized odds that is satisfied when the output of the predictor is conditionally independent of the protected attribute for only the advantageous class. That is $\forall a, a' \in \mathcal{A}, \Pr(\hat{Y} = y_+ | A = a, Y = y_+) = \Pr(\hat{Y} = y_+ | A = a', Y = y_+)$

Equalized odds and equalized opportunity are more practical definitions of fairness when applied to a computer vision problems because they still allow full predictive capability [12]. Further, since equalized opportunity is a relaxation of equalized odds, if equalized odds is achieved then equalized opportunity is also achieved. Therefore, in this work we aim to create predictors that satisfy equalized odds.

2.3 Distance Measures for Equalized Odds

Though the goal is to achieve true equalized odds, current methods are unable to achieve it [5, 16, 33]. Therefore, we need to use metrics to quantify how far a predictor is from true equalized odds. In this work we use three different metrics to measure the level of fairness of a predictor. The first two metrics use the difference in predictor output between different demographics of a protected attribute. This difference is called the *difference of equalized odds* (DEO).

$$\text{DEO}(a, a', y) \triangleq |\Pr(\hat{Y} = y | A = a, Y = y) - \Pr(\hat{Y} = y | A = a', Y = y)|. \quad (1)$$

DEO can be directly used when the protected attribute is binary and can easily be extended for more demographics by aggregating DEO across the different target and protected attribute values. The methods used to aggregate DEO differ between various works in the literature. We use the aggregation methods from Jung *et al.* [16] who propose two different methods of aggregation, DEO_{\max} and DEO_{avg} which are shown in Eqs.(2) and (3), respectively. DEO_{\max} can be used to understand the peak bias of an AI model and DEO_{avg} can be used to understand the bias of a model in the majority of cases.

$$\text{DEO}_{\max} \triangleq \max_y (\max_{a, a'} (\text{DEO}(a, a', y))). \quad (2)$$

$$\text{DEO}_{\text{avg}} \triangleq \frac{1}{|\mathcal{Y}|} \sum_y (\max_{a, a'} (\text{DEO}(a, a', y))). \quad (3)$$

Another fairness metric that is commonly reported, often in the Fair face recognition literature, is the standard deviation of accuracy across the demographics of the protected attribute, denoted by σ_{Acc} . This metric is shown in Eq. (5), where μ is the average accuracy across all the demographics. Note that

$\Pr(\hat{Y} = y|A = a)$ is equivalent to the accuracy of the predictor \hat{Y} in the domain of demographic a .

$$\mu = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} [\Pr(\hat{Y} = y|A = a)] \quad (4)$$

$$\sigma_{\text{Acc}} \triangleq \sqrt{\frac{1}{|\mathcal{A}|} \sum_a [\Pr(\hat{Y} = y|A = a) - \mu]^2} \quad (5)$$

All these metrics represent a distance from true equalized odds. In all cases this means that lower values indicate a fairer classifier.

3 Method

3.1 A Differentiable Approximation for Distance from Equalized Odds

The strategy used to train an AI model for classification uses a distance measure between the models output distribution and the true data distribution, referred to as the *loss* or *objective* function. Then a gradient optimization method is used to update the parameters of the model to reduce the distance measure. This is a simple but incredibly effective strategy. We aim to use the same strategy to increase the fairness of an AI model. We use σ_{Acc} as an objective function to reduce the distance from true EO.

In what follows, we use boldface fonts to denote vectors, *e.g.*, $\hat{\mathbf{y}} \in \hat{\mathcal{Y}}$ denotes the output vector of the model. We use \hat{y}_t to show the element corresponding to the ground truth label y in $\hat{\mathbf{y}}$. Furthermore, $\hat{y}_m = \max(\hat{\mathbf{y}} \setminus \{\hat{y}_t\})$ represents the largest non ground truth element of $\hat{\mathbf{y}}$ and $\hat{\mathcal{Y}}_a$ represents the domain of demographic a for the protected attribute. Accuracy of a single sample $\hat{\mathbf{y}}$ is defined in Eq. (6).

$$\text{Acc}(\hat{y}_t, \hat{y}_m) \triangleq \begin{cases} 1 & \hat{y}_t > \hat{y}_m \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

In essence, if the element \hat{y}_t is greater than all other elements, the model has correctly predicted the outcome for this sample and therefore, has an accuracy of one.

Since $\mathbb{E}_{\hat{\mathbf{y}} \sim \hat{\mathcal{Y}}_a} [\text{Acc}(\hat{y}_t, \hat{y}_m)] = \Pr(\hat{Y} = y|A = a)$, we substitute the expectation into Eqs. (4) and (5), which gives us Eqs. (7) and (8).

$$\mu = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \mathbb{E}_{\hat{\mathbf{y}} \sim \hat{\mathcal{Y}}_a} [\text{Acc}(\hat{y}_t, \hat{y}_m)] \quad (7)$$

$$\sigma_{\text{Acc}} = \sqrt{\frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \left[\mathbb{E}_{\hat{\mathbf{y}} \sim \hat{\mathcal{Y}}_a} [\text{Acc}(\hat{y}_t, \hat{y}_m)] - \mu \right]^2} \quad (8)$$

This is the objective we would like to optimize. However to be used for gradient based optimization that AI models are trained with an objective needs to be differentiable, which σ_{Acc} is not due to the undefined gradient at $\hat{y}_t = \hat{y}_m$ of $\text{Acc}(\hat{y}_t, \hat{y}_m)$. Instead we approximate the accuracy using a sigmoid based soft accuracy function, shown in Eq. (9), which is a differentiable approximation of accuracy. The soft accuracy is characterised by κ , which is a hyper-parameter that describes the sharpness of the function. A higher value of κ leads to a closer approximation of accuracy with $\lim_{\kappa \rightarrow \infty} \text{Acc}_{\text{soft}}(\hat{y}_t, \hat{y}_m) = \text{Acc}(\hat{y}_t, \hat{y}_m)$, however this is paired with an increased sparsity of the gradient.

$$\text{Acc}_{\text{soft}}(\hat{y}_t, \hat{y}_m) \triangleq \frac{1}{1 + e^{-\kappa(\hat{y}_t - \hat{y}_m)}} \quad (9)$$

We then substitute soft accuracy into σ_{Acc} for accuracy. This gives us the objective shown in Eq. (11).

$$\mu_{\text{soft}} = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \mathbb{E}_{\hat{y} \sim \hat{y}_a} [\text{Acc}_{\text{soft}}(\hat{y}_t, \hat{y}_m)] \quad (10)$$

$$\sigma_{\text{Acc}_{\text{soft}}} \triangleq \sqrt{\frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \left[\mathbb{E}_{\hat{y} \sim \hat{y}_a} [\text{Acc}_{\text{soft}}(\hat{y}_t, \hat{y}_m)] - \mu_{\text{soft}} \right]^2} \quad (11)$$

This is the differentiable objective that we can optimize to obtain a fair predictor.

3.2 Training Objective

By itself the soft accuracy fairness objective does not learn to classify. In fact the easiest solution for a model to achieve equalized odds is to randomly classify each sample. Since it is important that the model still achieves high utility we combine the soft accuracy fairness objective with a cross entropy classification objective. This gives us the full objective which is shown in Eq. (12).

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \gamma \sigma_{\text{Acc}_{\text{soft}}} \quad (12)$$

The two losses, \mathcal{L}_{ce} and $\sigma_{\text{Acc}_{\text{soft}}}$, aim to achieve different objectives, which are classification performance and fairness respectively. Applying too much weight to one objective can harm the other. We use γ as a hyper-parameter to balance the utility of the model with the fairness. A higher value for γ will result in a fairer classifier, however at this can often come at the cost of classification performance. We experimentally determined the optimal value of γ for each dataset by performing a grid search. However, we observe an extensive search is not required and finding the correct order of magnitude results in good performance.

3.3 Balancing the Training Dataset

When calculating $\sigma_{\text{Acc}_{\text{soft}}}$ on a mini-batch the number of samples used to estimate the soft accuracy for each protected demographic is highly important. If the number of samples for a particular demographic is too low then the variance of the soft accuracy estimation will increase. Differences in variance between the different demographics lead to instability of training gradients which has a negative impact on performance. To counter this effect we simply oversample the training dataset set such that each protected, target attribute pair is evenly sampled. This is achieved by randomly duplicating samples from the undersampled pairs until all protected, target attribute pairs contain the same number of samples. There exist more sophisticated methods [22,31] which could be used to augment the training dataset and their use may lead to gains in performance. However, we leave this investigation to future work.

4 Experiments

In the following section we thoroughly investigate and validate the capability of our soft accuracy fairness objective.

4.1 Baselines

We compare our algorithm with four different baselines. The first is a naïve classifier that is not aware of fairness in any regard. This baseline represents the worst case scenario for fairness. Since one source of bias is an unbalanced dataset we also include a naïve classifier which is trained by oversampling the dataset such that it is balanced. We refer to this baseline as *Naïve Balanced*. The third baseline is Adversarial Debiasing (AD) [33] which is used as a common benchmark method and the final baseline is the state-of-the-art in-processing method, MFD [16]. The original MFD paper only provided results for the age task with the UTKFace dataset. Additionally, the original MFD paper only implemented a simple data augmentation scheme. We employed further data augmentations which allowed our naïve classifier to achieve a much higher accuracy (74.7% vs 83.1%). In the spirit of fair comparison, we apply their method code with our datasets and augmentation scheme, this allows MFD to achieve comparable accuracy. Where applicable, results from the original paper are reported as MFD^o. Similarly, AD was originally implemented on non computer vision tasks, we re-implement AD for evaluation on CV tasks. For both re-implementations we perform a sweep of the bias loss hyper-parameter, discard hyper-parameter choices that lead to a large reduction in accuracy and report the best results.

4.2 Datasets

We use three datasets for our experiments. UTKFace [17], CelebA [20] and Fairface [18]. UTKFace and CelebA are face image datasets commonly used to benchmark fairness. UTKFACE contains 20k samples with annotations of age, gender

and ethnicity. CelebA contains 200k images which are labelled with 40 binary attributes. The images from UTKFace and CelebA cover a large variation in position, facial expression, illumination, occlusion and resolution. Buolamwini and Gebre [4] note that collecting a balanced dataset should be the first step in a fairness solution. Therefore it is important that we also evaluate our method under these conditions, for which we use the Fairface dataset. Fairface is also a face image dataset. It contains 98k images with annotations of age, gender and ethnicity. Fairface was created in an effort to reduce racial bias in existing datasets and had a strong focus on reducing the imbalance of races in the dataset during its creation. As shown in Fig. 1, compared to UTKFace, the race labels in Fairface are much more balanced. Using these UTKFace, CelebA and Fairface we evaluate two scenarios. Where a task is trained with a balanced dataset and where the task is trained with a biased dataset. UTKFace provides the age labels as integers, instead of learning a regression problem we group ages together into classes. To allow comparison we follow the division used by Jung *et al.* [16] where ages are divided into three classes, less than 20, 20–39 and greater than 40. Fairface provides age labels in classes already, however they are heavily imbalanced with far fewer samples in the extreme young and old classes. To maintain Fairface as a balanced test set we divide the ages in four new classes to balance them. These four classes are 0–19, 20–29, 30–39 and 40+.

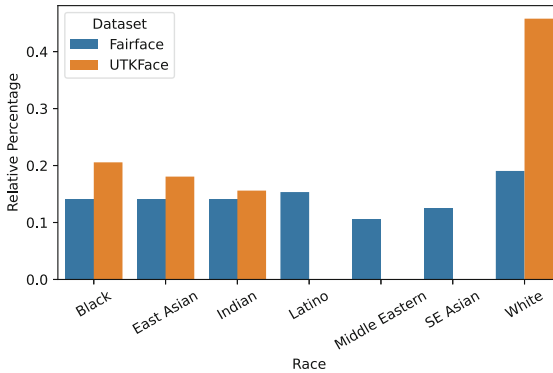


Fig. 1. The relative distribution of different races in the UTKFace dataset and Fairface dataset. UTKFace labels both East Asian and South East Asian faces together so these are shown under East Asian.

Skewed Fairface. Since it is imperative to understand how the performance of a fairness algorithm is related to the bias of a dataset, we present a protocol for controlling the bias within a dataset. We apply this protocol to Fairface to create a dataset which we name Fairface Skewed (FairfaceS). FairfaceS is characterised by a skew parameter (s) which can range from 0 to 1. The skew parameter describes the relative distribution of (target, protected) attribute pairs where

a higher skew parameter leads to a dataset with a higher correlation between the target attribute and the protected attribute. The relative distribution is calculated by arranging the classes into a 2D array. Two diagonal corners are assigned the value of 1 and the other two diagonal corners are assigned the value of $1 - s$. Bilinear interpolation is then used to calculate the remaining values of the matrix. An example of the relative distribution for different skews is shown in Fig. 2. Fairface is then under-sampled such that the relative distribution of each (target, protected) pair matches that in the matrix. This protocol imposes an order on the class however, in the absence of a rigorous similarity metric between separate demographics and target attribute values we simply order the classes alphabetically.

| | | | | | | |
|-----------------|-----------|------|-----------|------|-----------|------|
| Black | 1 | 0.9 | 1 | 0.5 | 1 | 0.1 |
| East Asian | 0.98 | 0.92 | 0.92 | 0.58 | 0.85 | 0.25 |
| Indian | 0.97 | 0.93 | 0.83 | 0.67 | 0.7 | 0.4 |
| Latino | 0.95 | 0.95 | 0.75 | 0.75 | 0.55 | 0.55 |
| Middle Eastern | 0.93 | 0.97 | 0.67 | 0.83 | 0.4 | 0.7 |
| Southeast Asian | 0.92 | 0.98 | 0.58 | 0.92 | 0.25 | 0.85 |
| White | 0.9 | 1 | 0.5 | 1 | 0.1 | 1 |
| | Female | Male | Female | Male | Female | Male |
| | Skew: 0.1 | | Skew: 0.5 | | Skew: 0.9 | |

Fig. 2. The relative distribution of (protected, target) pairs in the FairfaceS dataset for different skew values.

As the skew value increases the mutual information between the protected attribute and the target attribute increases leading to an increase in the bias. Using FairfaceS allows us to evaluate how a fairness algorithm performs under varying degrees of dataset bias. Additionally, because FairfaceS uses genuine attributes that can be linked in complicated manners, rather than creating a bias with respect to augmentations such as grayscaling an image, it allows for greater understanding of how a system may behave in a real-world scenario.

Balanced Test Set and Triplicate Experiments. When evaluating the fairness of a model it is important that the test set have a uniform distribution of (protected, target) pairs. If a particular pair is undersampled or oversampled it will have a disproportionate impact on the results, e.g. if the White, Male pair is more prevalent in the test set then the accuracy on this pair will affect the average accuracy more. To ensure that our results do not include any bias toward a particular label pair we select samples for the test set such that each protected, target label pair is included in equal numbers. We also observed that whilst the target classification performance is stable over different training and

test splits, the fairness varies by a large degree. To ensure robust results we perform our experiments on three different train test splits and report the mean and standard deviation. The exception is our experiments with CelebA, for which we use the official train, validation, and test sets as this allows us to compare to previous work. The results for CelebA are reported over three different random initializations.

4.3 Implementation Details

For all experiments we use a Resnet18 [13]. For experiments on UTKFace, Fairface and FairfaceS models are initialised from weights that were pretrained on Imagenet-1k [7]. Models in the CelebA experiments are randomly initialised. More details about the exact training procedure can be found in the supplementary material. MFD and AD are implemented according to their original papers. However, we follow Jung *et al.* [16] and remove the gradient projection from the original work to increase stability of training.

4.4 Classification Tasks

In this section we investigate the performance of our method on two tasks, age and gender classification.

Unbalanced Data. First we test the scenario in which the training data from the task is not balanced. This is the case for the majority of AI tasks unless special care has been taken during the creation of the dataset. For this experiment, we use the UTKFace and CelebA datasets. For UTKFace we use *race* as the protected attribute and test both age and gender as target attributes. For CelebA we use the *Male* attribute as the protected attribute and *Attractive* as the target attribute. The results are shown in Tables 1, 2 and 3, respectively. In both UTKFace scenarios all fairness methods improve fairness over a naïve classifier. The age classification task is harder than gender and is also much less fair, with the naïve classifier only achieving a σ_{Acc} of 8.5 compared to 3.0 for the gender task. In the highly unfair scenario with age as the target attribute, we observe that BASE achieves the best fairness for σ_{Acc} and DEO_{avg} , whilst achieving the highest over-all accuracy. It is only outperformed on DEO_{max} by MFD^o which does so with at a significantly lower over-all accuracy. Whilst the data for the gender task is still unbalanced, we observe that the naïve classifier can already achieve a better level of fairness leading us to believe this is a fairer task. For this task, BASE is competitive and achieves the second best accuracy and fairness. MFD achieves the greater fairness, however, this comes at the expense of a lower over-all accuracy. In the CelebA scenario BASE achieves the highest performance in all metrics. Again the fairness of the naï classifier is low for this scenario, showing that the CelebA task is unfair. These experiments show that BASE works best in an environment that is particularly unfair.

Table 1. Comparison of methods on UTKFace dataset with age as the target variable. Best results are **bold** and second best are underline. Results marked \diamond are reported directly from [16].

| Method | Acc. \uparrow | σ_{Acc} \downarrow | DEO _{max} \downarrow | DEO _{avg} \downarrow |
|---------------------------|----------------------------------|------------------------------------|----------------------------------|----------------------------------|
| Naïve Classifier | 82.5 \pm 1.5 | 8.8 \pm 1.2 | 45.3 \pm 2.5 | 25.8 \pm 1.3 |
| Naïve Classifier Balanced | 83.3 \pm 1.1 | 8.7 \pm 1.3 | 43.4 \pm 3.8 | 21.9 \pm 2.2 |
| MFD $^\diamond$ [16] | 74.7 \pm 0.7 | - | 28.5 \pm 1.8 | <u>17.8 \pm 1.4</u> |
| MFD [16] | 83.4 \pm 0.5 | <u>6.6 \pm 1.3</u> | 32.3 \pm 3.5 | 18.3 \pm 2.0 |
| AD [33] | <u>83.6 \pm 1.4</u> | 7.3 \pm 1.4 | 41.0 \pm 5.6 | 21.2 \pm 3.9 |
| BASE <i>ours</i> | 83.8 \pm 0.6 | 5.6 \pm 0.7 | <u>29.0 \pm 2.6</u> | 16.0 \pm 1.3 |

Table 2. Comparison of methods on UTKFace dataset with gender as the target variable. Best results are **bold** and second best are underline.

| Method | Acc. \uparrow | σ_{Acc} \downarrow | DEO _{max} \downarrow | DEO _{avg} \downarrow |
|---------------------------|----------------------------------|------------------------------------|---------------------------------|---------------------------------|
| Naïve Classifier | 93.1 \pm 0.7 | 2.8 \pm 0.3 | 10.3 \pm 3.2 | 7.2 \pm 0.8 |
| Naïve Classifier Balanced | <u>93.5 \pm 0.9</u> | 2.8 \pm 0.6 | 10.0 \pm 1.5 | 7.2 \pm 1.3 |
| MFD [16] | 92.4 \pm 0.4 | 2.1 \pm 0.4 | 8.0 \pm 1.0 | 5.7 \pm 0.6 |
| AD [33] | 93.9 \pm 1.1 | 2.6 \pm 0.9 | 8.0 \pm 2.6 | <u>6.2 \pm 2.0</u> |
| BASE <i>ours</i> | 93.4 \pm 0.3 | <u>2.3 \pm 0.9</u> | <u>9.0 \pm 1.0</u> | <u>6.2 \pm 0.8</u> |

Table 3. Comparison of methods on CelebA dataset with attractive as the target variable. Best results are **bold** and second best are underline. Results marked \diamond are reported directly from [21].

| Method | Acc. \uparrow | σ_{Acc} \downarrow | DEO _{max} \downarrow | DEO _{avg} \downarrow |
|---------------------------|----------------------------------|------------------------------------|---------------------------------|---------------------------------|
| Naïve Classifier | 79.6 \pm 0.2 | 3.0 \pm 3.5 | 26.3 \pm 0.4 | 25.9 \pm 0.6 |
| Naïve Classifier Balanced | <u>80.1 \pm 0.2</u> | <u>1.1 \pm 0.3</u> | <u>3.5 \pm 0.4</u> | <u>2.1 \pm 0.4</u> |
| MFD $^\diamond$ [16] | 78 \pm 0.3 | - | - | 7.4 \pm 0.3 |
| FSCL $^\diamond$ [21] | 79.1 \pm 0.4 | - | - | 11.5 \pm 0.3 |
| FSCL+ $^\diamond$ [21] | 79.1 \pm 0.4 | - | - | 6.5 \pm 0.4 |
| BASE <i>ours</i> | 80.7 \pm 0.1 | 0.8 \pm 0.2 | 3.0 \pm 0.7 | 1.9 \pm 0.5 |

Balanced Data. Next, we test the scenario in which the training data for the task has been collected with a focus on ensuring that it is balanced with respect to the protected attribute. For this experiment, we use the Fairface dataset and *race* as the protected attribute. For the classification target attribute we test both age and gender. The results are shown in Tables 4 and 5, respectively.

In these two scenarios, we observe that the fairness of the naïve classifier is already high due to the balanced nature of the data. For both target attributes, the naïve classifier with balanced sampling achieves the best fairness for two

Table 4. Comparison of methods on Fairface dataset with age as the target variable. Best results are **bold** and second best are underline.

| Method | Acc. \uparrow | $\sigma_{\text{Acc.}}$ \downarrow | DEO _{max} \downarrow | DEO _{avg} \downarrow |
|---------------------------|----------------------------------|-------------------------------------|----------------------------------|----------------------------------|
| Naïve Classifier | 67.4 \pm 0.2 | 2.1 \pm 0.5 | 23.6 \pm 4.3 | 16.2 \pm 0.4 |
| Naïve Classifier Balanced | 66.2 \pm 0.4 | 1.9 \pm 0.4 | 12.4 \pm 0.8 | 10.3 \pm 0.2 |
| MFD [16] | <u>68.3 \pm 0.3</u> | 1.9 \pm 0.2 | <u>14.6 \pm 0.9</u> | <u>10.6 \pm 1.4</u> |
| AD [33] | 68.4 \pm 0.2 | 2.2 \pm 0.2 | 21.3 \pm 0.5 | 15.7 \pm 0.9 |
| BASE <i>ours</i> | 68.4 \pm 0.4 | <u>2.0 \pm 0.04</u> | 14.8 \pm 0.5 | 10.8 \pm 1.2 |

Table 5. Comparison of methods on Fairface dataset with gender as the target variable. Best results are **bold** and second best are underline.

| Method | Acc. \uparrow | $\sigma_{\text{Acc.}}$ \downarrow | DEO _{max} \downarrow | DEO _{avg} \downarrow |
|---------------------------|-----------------------------------|-------------------------------------|---------------------------------|---------------------------------|
| Naïve Classifier | 93.4 \pm 0.05 | <u>2.0 \pm 1.6</u> | 7.5 \pm 1.7 | 6.3 \pm 0.8 |
| Naïve Classifier Balanced | 93.6 \pm 0.05 | 1.9 \pm 0.2 | 6.9 \pm 0.3 | 6.4 \pm 0.3 |
| MFD [16] | 93.4 \pm 0.1 | 2.2 \pm 0.05 | 7.7 \pm 1.1 | 7.0 \pm 0.4 |
| AD [33] | 93.6 \pm 0.2 | 2.1 \pm 0.1 | 7.3 \pm 0.6 | 6.7 \pm 0.4 |
| BASE <i>ours</i> | <u>93.5 \pm 0.02</u> | <u>2.0 \pm 0.1</u> | <u>7.0 \pm 0.6</u> | <u>6.4 \pm 0.4</u> |

of the three metrics. However, this comes at the cost of accuracy for the age task. For both tasks BASE achieves the second best results for $\sigma_{\text{Acc.}}$, with equal highest overall accuracy in the age task and the second best overall accuracy for the gender task.

Biased Data. Finally, we investigate how our method performs with an increasingly biased dataset. For this experiment we use the FairfaceS dataset (Sect. 4.2) with gender as the target variable. We evaluate a naïve classifier and BASE over a range of different skew parameters and observe the effect on accuracy and fairness. The results are shown in Fig. 3.

We observe that, as one would expect, as the skew increases and consequently the bias in the dataset increases both accuracy and fairness decay for both methods. Additionally, at low levels of bias, whilst BASE is able to increase the fairness of the classifier in all metrics, this comes at the cost of overall accuracy compared to the naïve classifier. However, as the skew increases the accuracy of the naïve classifier decays at a greater rate than BASE. At extreme skew levels, BASE is even able to achieve a higher degree of overall accuracy. The same results can be seen with the fairness metrics. With the performance of the naïve classifier decaying at a higher rate than BASE. Even though BASE produces a more fair predictor at low skew levels, the performance gap only increases as the skew increases.

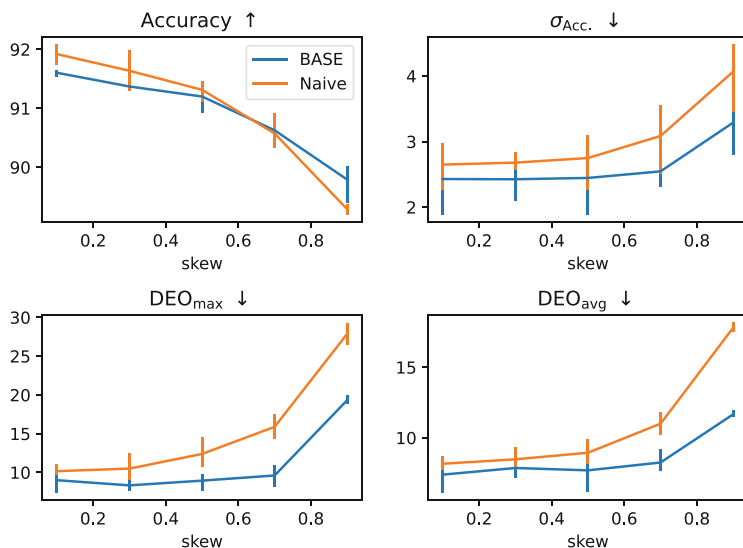


Fig. 3. The accuracy and fairness of a Naïve classifier and BASE over different skew parameters of the FairfaceS dataset. Error bars are the 95% confidence interval over 3-fold cross-validation.

5 Conclusion

In this work, we introduce a new fairness objective based upon optimising the standard deviation of soft accuracy across demographics of a protected attribute. Experimental results on UTKFace, CelebA, Fairface and FairfaceS show that our system is able to produce fairer AI models for computer vision tasks under widely varying conditions whilst being particularly effective for more unfair scenarios and can even improve the overall accuracy compared to a naive model in heavily biased data-sets.

References

1. Amini, A., Soleimany, A.P., Schwarting, W., Bhatia, S.N., Rus, D.: Uncovering and mitigating algorithmic bias through learned latent structure. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. AIES 2019, Honolulu, HI, USA, pp. 289–295. Association for Computing Machinery, New York (2019). <https://doi.org/10.1145/3306618.3314243>
2. Beutel, A., Chi, E.H., Chen, J., Zhao, Z.: Data decisions and theoretical implications when adversarially learning fair representations. In: FAT/ML (2017). <https://arxiv.org/pdf/1707.00075.pdf>

3. van Breugel, B., Kyono, T., Berrevoets, J., van der Schaar, M.: DECAF: generating fair synthetic data using causally-aware generative networks. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W. (eds.) *Advances in Neural Information Processing Systems*, vol. 34, pp. 22221–22233. Curran Associates, Inc. (2021). <https://proceedings.neurips.cc/paper/2021/file/ba9fab001f67381e56e410575874d967-Paper.pdf>
4. Buolamwini, J., Gebru, T.: Gender shades: intersectional accuracy disparities in commercial gender classification. In: Friedler, S.A., Wilson, C. (eds.) *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. *Proceedings of Machine Learning Research*, vol. 81, pp. 77–91. PMLR, February 2018. <https://proceedings.mlr.press/v81/buolamwini18a.html>
5. Cho, J., Hwang, G., Suh, C.: A fair classifier using kernel density estimation. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) *Advances in Neural Information Processing Systems*, vol. 33, pp. 15088–15099. Curran Associates, Inc. (2020). <https://proceedings.neurips.cc/paper/2020/file/ac3870fcad1fc367825cda0101eee62-Paper.pdf>
6. Creager, E., et al.: Flexibly fair representation learning by disentanglement. In: Chaudhuri, K., Salakhutdinov, R. (eds.) *Proceedings of the 36th International Conference on Machine Learning*. *Proceedings of Machine Learning Research*, vol. 97, pp. 1436–1445. PMLR, Long Beach, California, USA, June 2019. <http://proceedings.mlr.press/v97/creager19a.html>
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
8. Dhar, P., Gleason, J., Roy, A., Castillo, C.D., Chellappa, R.: PASS: protected attribute suppression system for mitigating bias in face recognition. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15087–15096, October 2021
9. Gong, S., Liu, X., Jain, A.K.: Jointly de-biasing face recognition and demographic attribute estimation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) *ECCV 2020*. LNCS, vol. 12374, pp. 330–347. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58526-6_20
10. Gong, S., Liu, X., Jain, A.K.: Mitigating face recognition bias via group adaptive classifier. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3413–3423. IEEE, Nashville, June 2021. <https://doi.org/10.1109/CVPR46437.2021.00342>, <https://ieeexplore.ieee.org/document/9577411/>
11. Grother, P., Ngan, M., Hanaoka, K.: Face recognition vendor test part 3: demographic effects, December 2019. <https://doi.org/10.6028/NIST.IR.8280>
12. Hardt, M., Price, E., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, vol. 29. Curran Associates, Inc. (2016). <https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf>
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016
14. Hwang, S., Park, S., Lee, P., Jeon, S., Kim, D., Byun, H.: Exploiting transferable knowledge for fairness-aware image classification. In: Ishikawa, H., Liu, C.-L., Pajdla, T., Shi, J. (eds.) *ACCV 2020*. LNCS, vol. 12625, pp. 19–35. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-69538-5_2

15. Joseph, M., Kearns, M., Morgenstern, J.H., Roth, A.: Fairness in learning: classic and contextual bandits. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, vol. 29, pp. 325–333. Curran Associates, Inc. (2016). <http://papers.nips.cc/paper/6355-fairness-in-learning-classic-and-contextual-bandits.pdf>
16. Jung, S., Lee, D., Park, T., Moon, T.: Fair feature distillation for visual recognition. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12110–12119. IEEE, Nashville, June 2021. <https://doi.org/10.1109/CVPR46437.2021.01194>, <https://ieeexplore.ieee.org/document/9578197/>
17. Kamiran, F., Calders, T.: Classifying without discriminating. In: 2009 2nd International Conference on Computer, Control and Communication, pp. 1–6 (2009). <https://doi.org/10.1109/IC4.2009.4909197>
18. Karkkainen, K., Joo, J.: FairFace: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1548–1558 (2021)
19. Klare, B.F., Burge, M.J., Klontz, J.C., Vorder Bruegge, R.W., Jain, A.K.: Face recognition performance: role of demographic information. *IEEE Trans. Inf. Forensics Secur.* **7**(6), 1789–1801 (2012). <https://doi.org/10.1109/TIFS.2012.2214212>
20. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015
21. Park, S., Lee, J., Lee, P., Hwang, S., Kim, D., Byun, H.: Fair contrastive learning for facial attribute classification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10389–10398, June 2022
22. Ramaswamy, V.V., Kim, S.S.Y., Russakovsky, O.: Fair attribute classification through latent space de-biasing. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9297–9306. IEEE, Nashville, June 2021. <https://doi.org/10.1109/CVPR46437.2021.00918>, <https://ieeexplore.ieee.org/document/9578650/>
23. Roh, Y., Lee, K., Whang, S.E., Suh, C.: FairBatch: batch selection for model fairness. In: *ICLR* (2021). <https://openreview.net/forum?id=YNnpaAKeCfx>
24. Roth, K., Lucchi, A., Nowozin, S., Hofmann, T.: Stabilizing training of generative adversarial networks through regularization. In: Guyon, I., et al. (eds.) *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc. (2017). <https://proceedings.neurips.cc/paper/2017/file/7bccfde7714a1ebadf06c5f4cea752c1-Paper.pdf>
25. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823 (2015). <https://doi.org/10.1109/CVPR.2015.7298682>
26. Shekhar, S., Fields, G., Ghavamzadeh, M., Javidi, T.: Adaptive sampling for minimax fair classification. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W. (eds.) *Advances in Neural Information Processing Systems*, vol. 34, pp. 24535–24544. Curran Associates, Inc. (2021). <https://proceedings.neurips.cc/paper/2021/file/cd7c230fc5deb01ff5f7b1be1acef9cf-Paper.pdf>
27. Shen, A., Han, X., Cohn, T., Baldwin, T., Frermann, L.: Optimising equal opportunity fairness in model training (2022). <https://doi.org/10.48550/ARXIV.2205.02393>, <https://arxiv.org/abs/2205.02393>

28. Wang, M., Deng, W.: Mitigating bias in face recognition using skewness-aware reinforcement learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9319–9328. IEEE, Seattle, June 2020. <https://doi.org/10.1109/CVPR42600.2020.00934>, <https://ieeexplore.ieee.org/document/9156925/>
29. 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 (ICCV), October 2019
30. Wang, Z., et al.: Towards fairness in visual recognition: effective strategies for bias mitigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020
31. 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 (2018). <https://doi.org/10.1109/BigData.2018.8622525>
32. Xu, X., et al.: Consistent instance false positive improves fairness in face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 578–586, June 2021
33. 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, AIES '18, New Orleans, LA, USA, pp. 335–340. Association for Computing Machinery, New York (2018). <https://doi.org/10.1145/3278721.3278779>
34. Zietlow, D., et al.: Leveling down in computer vision: pareto inefficiencies in fair deep classifiers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10410–10421, June 2022