Counterfactual Fairness for Facial Expression Recognition

Jiaee Cheong[1,2], Sinan Kalkan[3], and Hatice Gunes[1]

[1] University of Cambridge, Cambridge, UK
jc2208@cam.ac.uk,hatice.gunes@cl.cam.ac.uk
[2] Alan Turing Institute, London, UK
[3] Middle East Technical University, Ankara, Turkey skalkan@metu.edu.tr

Abstract. Given the increasing prevalence of facial analysis technology, the problem of bias in these tools is becoming an even greater source of concern. Causality has been proposed as a method to address the problem of bias, giving rise to the popularity of using counterfactuals as a bias mitigation tool. In this paper, we undertake a systematic investigation of the usage of counterfactuals to achieve both statistical and causal-based fairness in facial expression recognition. We explore bias mitigation strategies with counterfactual data augmentation at the pre-processing, in-processing, and post-processing stages as well as a stacked approach that combines all three methods. At the in-processing stage, we propose using Siamese Networks to suppress the differences between the predictions on the original and the counterfactual images. Our experimental results on RAF-DB with counterfactuals added show that: (1) The in-processing method outperforms at the pre-processing and post-processing stages, in terms of accuracy, F1 score, statistical fairness and counterfactual fairness, and (2) stacking the pre-processing, in-processing and post-processing stages provides the best performance.

Keywords: Bias Mitigation, Counterfactual Fairness, Facial Expression Recognition.

1 Introduction

Given the increasing prevalence and stakes involved in machine learning applications, the problem of bias in such applications is now becoming an even greater source of concern. The same is true for facial affect analysis technologies [9]. Several studies have highlighted the pervasiveness of such discrimination [4, 19, 24] and a number of works have sought to address the problem by proposing solutions for mitigation [5, 8, 49]. In order to assess whether bias has been mitigated, a reliable measure of bias and fairness is sorely needed. A significant number of fairness definitions have been proposed [1, 18, 37, 46]. The more prevalent and long-standing definitions are often based upon statistical measures. However, statistical measures of fairness are mired with gaps. The definitions can often end up being mutually exclusive [7] and fail to distinguish spurious correlations between a sensitive attribute and the predicted outcome [15, 29, 39].

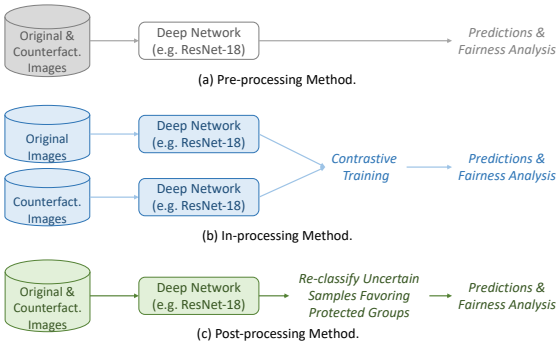

Fig. 1. Three methods for bias mitigation with counterfactual images.

Causal reasoning has been proposed as a potential instrument to address such gaps [29,34,39] and counterfactuals, one of the key ideas within causal reasoning, are increasingly used as a tool to achieve such goals. One such use case is to rely on counterfactuals as a *data augmentation* strategy. Such an approach has proved promising for several use cases within the field of natural language processing [11, 12, 14, 35, 36], recommendation systems [47] as well as visual question answering systems [41]. However, this approach has yet to be explored within the field of facial expression analysis. Our first contribution thus involves exploring the use of counterfactuals as a data augmentation strategy for the task of facial expression recognition. Second, as bias mitigation can be performed at the pre-processing, in-processing or post-processing stage [5], we will make use of counterfactuals at all three stages to mitigate bias as illustrated in Figure 1. To the best of our knowledge, no existing works describe a comprehensive system for mitigating bias at all three stages. Our key contributions are summarised as follows:

1. We make use of counterfactuals at the pre-processing, in-processing and post-processing stage for the first time in the literature in order to mitigate for bias in facial expression recognition.
2. We do an in-depth analysis of bias at the pre-processing, in-processing and post-processing stages using both statistical and causal measures of fairness. We show that different forms of bias can exist at different levels and are captured by the different measures used.

2 Literature Review

2.1 Fairness in Machine Learning

Fairness is now recognised as a significantly important component of Machine Learning (ML) given how the problem of bias can result in significant impact on human lives. The general Machine Learning (ML) approaches tackle bias typically at the pre-processing, in-processing and post-processing stages [5, 17,

44]. The pre-processing methods address the problem of bias at the data-level [44]. This typically involves some form of data augmentation or modification to the input data. The in-processing methods involve making modification to the learning algorithm itself [8]. The post-processing methods occur at the end of the learning and prediction process [28]. This usually involves altering the output to achieve fairer predictions. The pre- and post-processing methods are usually model agnostic and can be applied without having to re-train the model. In contrast, the in-processing method will involve making changes in the model or the training method.

Different fairness definitions result in very different quantitative measures which can result in very different algorithmic outcomes [1, 18]. To exacerbate matters, the different definitions can even sometimes be at odds with each other and improving the score on one fairness metric may very well involve a trade-off on another [2]. Hence, selecting the right definition and metric is a highly challenging and important task. A more thorough examination of these issues can be found in the following papers [1, 2, 18, 37].

There are two main groups of fairness measures. Statistical notion of fairness is based on statistical measures of the variables of interest, e.g. accuracy, true positive rate. As statistical measures are only able to capture correlation and not causation, this form of fairness is sometimes also referred to as *associational* fairness. Some well-known examples of such forms of fairness include demographic parity [50] and equality of opportunity [22]. As highlighted in several recent research, there are many gaps stemming from statistical fairness. As a result, causal notion of fairness has been proposed to address these gaps [29, 31, 44]. Causal fairness assumes the existence of an underlying causal model. Some examples of causal fairness include counterfactual fairness [31] and proxy fairness [29]. In this research, we will build upon the definition of *equality of opportunity* and *counterfactual fairness* to conduct a comparison between both types of fairness.

2.2 Facial Affect Fairness

Facial affect recognition involves automatically analysing and predicting facial affect [45]. The most common method is the discrete category method which assumes six basic emotion categories of facial expressions recognized universally (i.e., happiness, surprise, fear, disgust, anger and sadness) [16]. Another method is to rely on the Facial Action Coding System (FACS), a taxonomy of facial expressions in the form of Action Units (AUs), where the emotions can then be defined according to the combination of AUs activated [16]. The six basic categories have been criticized for being unrealistic and limited at representing a full spectrum of emotions using only a handful of categories [20]. Another alternative is to use a dimensional description of facial affect which views any affective state as being represented as a bipolar entity existing on a continuum. [43]. Depending on the method of distinguishing facial affect, this can be achieved by either training an algorithm to classify the facial expressions of emotion [49], predict the valence and the arousal value of the displayed facial expression or detect the activated facial action units [8]. To date, investigating bias and fairness

in facial affect recognition is still very much a understudied problem [5, 49]. Only a handful of studies have been done to highlight the bias and propose fairer solutions for facial affect analysis [8, 25, 40, 49]. In this paper, we focus on the task of classifying facial expressions and attempt to investigate a solution which addresses bias at the pre-, in- and post-processing stages with the use of counterfactuals.

2.3 Counterfactuals and Bias

A counterfactual is the result of an intervention after modifying the state of a set of variables X to some value $X = x$ and observing the effect on some output Y. Using Pearl's notation [42], this intervention is captured by the $do(\cdot)$ operator and the resulting computation then becomes $P(Y|do(X = x))$.Several existing frameworks offer methods for countering bias with the use of counterfactuals. Existing methods typically rely on using counterfactuals as a data augmentation strategy at the pre-processing stage to mitigate for bias. This method has proven to be successful within the field of natural language processing [11, 12, 14, 35, 36]. Its use case includes hate speech detection [11, 12], machine translation [35, 48] and dialogue generation [14]. Experiments done on recommendation systems [47] and more recently, in Visual Question Answering (VQA) systems [41] indicated that such an approach is promising.

Such an approach has yet to be explored for facial *expression* analysis. In the domain of facial analysis, counterfactuals have been used to identify [13, 27] and mitigate for bias [10]. Our research resembles that of [13] and [27] in that we used a generative adversarial network (GAN), STGAN [33], to generate adversarial counterfactual facial images to assess for counterfactual fairness. Though alike in spirit, our paper differs as follow. First, the above studies focused on investigating different methods for counterfactual generation [10,13,27]. In our case, we do not propose an alternative method to generate adversarial or counterfactual images. Instead, we use a pre-trained GAN [33] to generate images which would then be used to augment the original dataset.

Second, [13] and [27] focused on using the generated counterfactual images to measure the bias present in either a publicly available dataset [13] or existing black-box image analysis APIs [27] but did not propose any method to mitigate bias. Though we do conduct a bias analysis of the model's performance on counterfactuals, the focus is chiefly on deploying counterfactuals explicitly to mitigate for bias which was attempted by [10] as well. However, the goal in [10] focuses on "attractiveness" prediction rather than facial expression recognition. In addition, we focus on investigating methods to mitigate bias at all three stages which is distinctly different from all the previous research mentioned above.

3 Methodology

As highlighted in [5], we can intervene to mitigate bias at either the pre-processing, in-processing or post-processing stage. To investigate our bias mitigation proposal for the task of facial expression recognition, we conduct a comparative

study using counterfactuals at the three different stages. The first method is the pre-processing method which involves augmenting the training set using counterfactual images. Subsequently, we implemented an in-processing approach using a Siamese Network [3] to investigate further downstream methods for mitigating bias with the use of counterfactuals. Finally, we explore the use of a post-processing method, the Reject-Option Classification proposed by Kamiran et al. to mitigate bias [28].

3.1 Notation and Problem Definition

We adopt a machine learning approach where we have a dataset $D = \{(\mathbf{x}_i, y_i)\}_i$ such that $\mathbf{x}_i \in X$ is a tensor representing information (e.g., facial image, health record, legal history) about an individual I and $y_i \in Y$ is an outcome (e.g., identity, age, emotion, facial action unit labels) that we wish to predict. In other words, we assume that we have a classification problem and we are interested in finding a parametric predictor/mapping f with $f : X \rightarrow Y$. We use $\hat{y_i}$ to denote the predicted outcome for input $\mathbf{x}_i$ and $p(y_i|\mathbf{x}_i)$ is the predicted probability for $\mathbf{x}_i$ to be assigned to the correct class y_i. Each input $\mathbf{x}_i$ is associated (through an individual I) with a set of sensitive attributes $\{s_{j\in a}\}_a \subset S$ where a is e.g. *race* and $j \in \{$Caucasian, African-American, Asian$\}$. The minority group are those with sensitive attributes which are fewer in numbers (e.g. African-American) compared to the main group (e.g. Caucasian). Note that there are other attributes $\{z_{j\in a}\}_a \subset Z$ that are not sensitive. In bias mitigation, we are interested in diminishing the discrepancy between $p(\hat{y_i} = c|\mathbf{x}_i)$ and $p(\hat{y_j} = c|\mathbf{x}_j)$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ are facial images for different individuals and their sensitive attributes $s^{\mathbf{x}_i}$ and $s^{\mathbf{x}_j}$ are different.

3.2 Counterfactual Fairness

Causality-based fairness reasoning [29, 31, 44] assumes that there exists a cause-and-effect relationship between the attribute variables and the outcome. We follow the counterfactual fairness notation used by Kusner et al. [31]. Given a predictive problem with fairness considerations, where S, X and Y represent the sensitive attributes, the input, and the output of interest respectively, let us assume that we are given a causal model (U, V, F), where U is a set of latent background variables, which are factors not caused by any variable in the set V of observable variables and F is a set of functions over $V_i \in V$. In our problem setup, $V \equiv S \cup X$. Kusner et al. [31] postulate the following criterion for predictors of Y: Predictor $f(\,\cdot\,;\theta)$ is counterfactually fair if, under any context $X = \mathbf{x}$ and $S = s$, the following is true:

$$p(f(\mathbf{x}; \theta, U) = y \mid \mathbf{x}, s) = p(f(\mathbf{x}^{S\leftarrow s'}; \theta, U) = y \mid \mathbf{x}, s), \tag{1}$$

where $\mathbf{x}^{S\leftarrow s'}$ denotes the counterfactual input (image) obtained by changing the sensitive attribute to s'. To simplify notation, we will use $\mathbf{x}'$ to denote the counterfactual image, $\mathbf{x}^{S\leftarrow s'}$, in the rest of the paper. With reference to Equation

1, we would like to highlight that this definition is an individual-level fairness definition. For the set of experiments that we will be doing in this paper, we will be aggregating the counterfactual *counts* dissected according to sensitive attributes and class in order to facilitate measurement and comparison.

3.3 Counterfactual Image Generation

We used a state-of-the-art GAN, a pre-trained STGAN model [33] trained on the CelebA dataset to generate a set of counterfactual images for RAF-DB. In our experiments, the attribute that we have chosen to manipulate is that of skin tone – see Figure 2 for samples. This is because manipulating skin tone produces more consistent results than manipulating other sensitive attributes such as age or gender. The adversarial counterfactual images modified across the other sensitive attributes are less stable and more likely to be corrupted. Our counterfactuals involve lightening but not darkening skin tone as GANs are currently still incapable of effectively doing so [26]. We are not conflating skin tone with race. We recognise that they are separate entities with some overlaps. Our analysis is focused on mitigating bias stemming from difference in skin tone which aligns with the approach taken in other bias investigation research [4, 13]. Moreover, though the images generated may not be completely satisfactory, this is due to the limitations of GANs. This is a contextual challenge and the solutions proposed here can still be deployed when GANs have been further improved.

Fig. 2. Sample counterfactual images obtained by changing the skin tone of the original images (without changing facial expression) using a pre-trained STGAN model [33].

3.4 Baseline Approach

We take prior of Tian et al. [49] as baseline. For this, we use a 18-layer Residual Network (ResNet) [23] architecture and train it from scratch with a Cross Entropy loss to predict a single expression label y_i for each $\mathbf{x}_i$:

$$\mathcal{L}_{CE}(\mathbf{x}_i, y_i) = -\sum_{y \in Y} \mathbb{1}[y_i = \hat{y}_i] \log p(y|\mathbf{x}_i), \tag{2}$$

where $p(y|\mathbf{x}_i)$ denotes the predicted probability for $\mathbf{x}_i$ being assigned to class $y_i \in Y$ and $\mathbb{1}[\cdot]$ denotes the indicator function. The baseline model is trained on the original images.

3.5 Pre-processing: Data Augmentation with Counterfactuals

Similar to the works done in the field of natural language processing (discussed in Section 2.3), we make use of counterfactuals as a data augmentation strategy. In this approach, we generate a counterfactual for each image and feed them as input to train a new network (Figure 1). Hence, instead of having N image samples in D, we now have $2N$ number of training samples. The network is trained with these $2N$ images using a Cross Entropy loss defined in Equation 2.

3.6 In-processing: Contrastive Counterfactual Fairness

Counterfactual fairness is defined with respect to the discrepancy between the predictions on an image $\mathbf{x}_i$ and its counterfactual version $\mathbf{x}'_i$. To be specific, as discussed in Section 3.2, counterfactual fairness requires the gap between $p(y_i|\mathbf{x}_i)$ and $p(y_i|\mathbf{x}'_i)$ to be minimal. An in-processing solution that fits very well to this requirement is contrastive learning.

In general, the goal of contrastive learning [6] is to learn an embedding space which minimises the embedding distance between a pair of images which are of the same class (positive pair) and maximizes the distance between a pair of "unrelated" images (negative pair). However, in our setting, we seek to minimise the difference between the prediction probabilities on an image and its counterfactual version. We realize contrastive counterfactual fairness by feeding $\mathbf{x}_i$ and its counterfactual version $\mathbf{x}'_i$ through a Siamese network (Figure 3). The following contrastive loss then seeks to minimise the discrepancy (bias) between the predictions of the two branches:

$$\mathcal{L}_{con}(\mathbf{x}_i, \mathbf{x}'_i) = -\sum_{y \in Y} \mathbb{1}[f(\mathbf{x}_i; \theta) = \hat{y}_i] \log p(y_i|\mathbf{x}'_i), \tag{3}$$

where we penalize the Siamese network if the counterfactual prediction is not consistent with respect to the predicted label $f(\mathbf{x}_i; \theta)$ of the original image $\mathbf{x}_i$.

Each branch of the Siamese network has its own Cross Entropy Loss so that each branch can predict the correct label independently. The overall loss is then defined as:

$$\mathcal{L}_i = \alpha(\mathcal{L}_{CE}(\mathbf{x}_i, y_i) + \mathcal{L}_{CE}(\mathbf{x}'_i, y_i)) + \mathcal{L}_{con}(\mathbf{x}_i, \mathbf{x}'_i), \tag{4}$$

where $\mathcal{L}_{CE}$ is as defined in Equation 2, and α is a hyper-parameter which we tuned as 1.5. By jointly minimizing the two Cross Entropy objectives and $\mathcal{L}_{con}$, the network learns a representation that minimises the difference between the predictions on the original and counterfactual images as well as the individual prediction errors for both the original and counterfactual images. The authors of [10] leveraged on a similar idea though both research were done independently. It is similar in that, to enforce fairness, they added a regularizer between the logits of the image and its counterfactual which is similar to the functionality of $\mathcal{L}_{con}$. The slight difference is that the overall loss function only accounts for the CE loss of the original image whereas ours attempt to account for both the original and counterfactual loss via $\mathcal{L}_{CE}(\mathbf{x}_i, y_i)$ and $\mathcal{L}_{CE}(\mathbf{x}'_i, y_i)$ respectively.

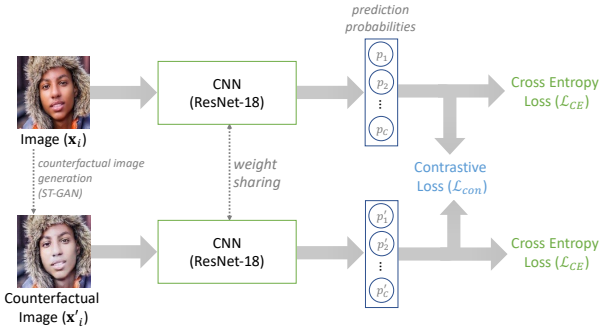

Fig. 3. An overview of the in-processing method: Original image $\mathbf{x}_i$ and its counterfactual $\mathbf{x}'_i$ are fed through a Siamese network, which is trained to minimize the discrepancy between the branches is penalized with a contrastive loss ($\mathcal{L}_{con}$) and to maximize the prediction probabilities of both branches using Cross Entropy loss ($\mathcal{L}_{CE}$).

3.7 Post-processing: Reject Option Classification

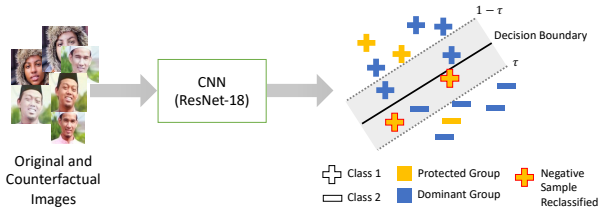

Fig. 4. Post-processing method by Kamiran et al. [28] reclassifies samples around the decision boundary in favor of protected groups.

Post-processing approaches take the output of the model and modify the output in a manner to achieve greater fairness. Here, we employ the Reject Option Classification suggested by Kamiran et al. [28]. This approach re-classifies the outputs where the model is less certain about, i.e., the predictions that fall in a region around the decision boundary parameterised by τ (see Figure 4). For instance, given a typical classification probability threshold of 0.5, if the model prediction is 0.85 or 0.15, this means that the model is highly certain of its prediction. If the values range around 0.53 or 0.44, this means that the input falls very close to the decision boundary and the model is less certain about its prediction. In order to improve fairness of the prediction, we reclassify the predicted output if it belongs to that of the minority group. More formally, if a sample $\mathbf{x}_i$ that falls in the "critical" region $1 - \tau \leq p(y|\mathbf{x}_i) \leq \tau$ where $0.5 \leq \tau \leq 1$, we reclassify $\mathbf{x}_i$ as y if $\mathbf{x}_i$ belongs to a minority group. Otherwise, i.e. when $p(y|\mathbf{x}_i) > \tau$, we accept the predicted output class y. In our experiments,

we set $\tau = 0.6$ as suggested by Kamiran et al. [28]. This method captures the innate human intuition to give the benefit of the doubt to samples from the minority group which they are unsure of.

4 Experimental Setup

4.1 Dataset

We chiefly conducted our experiments on the RAF-DB [32] dataset. RAF-DB contains labels in terms of facial expressions of emotions (Surprise, Fear, Disgust, Happy, Sad, Anger and Neutral) and sensitive attribute labels along gender, race and age. We excluded images labelled as "Unsure" for gender. In addition, the age binning system is likely to cause greater variation noise. For instance, an individual who is age 4 is likely to look very different from someone who is age 19 but yet they are categorised in the same category. As such, we have chosen to restrict our analysis to the sensitive attributes gender and race. We utilised a subset of the dataset consisting of 14,388 images, with 11,512 samples used for training and 2,876 samples used for testing within our experiments. This training and testing split has been pre-defined according to the instructions in the original dataset [32].

4.2 Implementation and Training Details

We first generated a set of counterfactual images using the method delineated in Section 3.3. This is done for both the training and testing images within RAF-DB. Our task subsequently focuses on categorising the seven categories of facial expressions of emotion. We then reclassified the counterfactual RAF-DB images (Figure 2) to evaluate for counterfactual biases as shown in Table 2.

Training Details ResNet-18 [23] is used as the baseline model as well as for the mitigation models as illustrated in Figure 1. For all models, we take the PyTorch implementation of ResNet and train it from scratch with the Adam optimizer [30], with a mini-batch size of 64, and an initial learning rate of 0.001 (except for the in-processing method for which 0.0005 worked slightly better). The learning rate decays linearly by a factor of 0.1 every 40 epochs. The maximum training epochs is 100, but early stopping is applied if the accuracy does not increase after 30 epochs. For the pre-processing method, we train a network with both the original and counterfactual images (Figure 1). For the in-processing method, we have two Siamese branches with shared weights (Figure 3). For the post processing approach, we train a network with both the original and counterfactual images (Figure 4) but take the output predictions and reclassify them according to the methodology delineated in Section 3.7.

Image Pre-processing and Augmentation All images are cropped to ensure faces appear in approximately similar positions. The images are subsequently normalized to a size of 128×128 pixels which are then fed into the networks as input. During the training step, we apply the following commonly used augmentation methods: Randomly cropping the images to a slightly smaller size (i.e. 96×96); rotating them with a small angle (i.e. range from -15^o to 15^o); and horizontally mirroring them in a randomized manner.

4.3 Evaluation Measures

In this paper, we use two measures: accuracy and F1-score, to evaluate prediction quality and two measures: equal opportunity and causal fairness, to evaluate fairness. *Equal opportunity* ($\mathcal{M}_{EO}$), a group-based metric, is used to compare the group fairness between models [22]. This can be understood as the largest accuracy gap among all demographic groups:

$$\mathcal{M}_{EO} = \frac{\min_{s \in S} \mathcal{M}_{ACC}(s)}{\max_{s \in S} \mathcal{M}_{ACC}(s)}, \tag{5}$$

where $\mathcal{M}_{ACC}(s)$ is the accuracy for a certain demographic group s. We also include a causality-based fairness metric *Counterfactual fairness* ($\mathcal{M}_{CF}$) because it has often been noted that commonly-used fairness metrics based on classification evaluation metrics such as accuracy, precision, recall and TP rate are insufficient to capture the bias present [29, 44]. $\mathcal{M}_{CF}$ is defined as:

$$\mathcal{M}_{CF} = \frac{1}{N} \sum_{i \in N} \mathbb{1}[f(\mathbf{x}_i; \theta) = f(\mathbf{x}'_i; \theta)], \tag{6}$$

where we compare the labels predicted by $f(\cdot; \theta)$. This is not a newly defined metric but a prevalent one based on an *aggregated* form of Counterfactual Fairness defined accordingly in Section 3.2 and [31].

Table 1. RAF-DB Test Set Distribution (Cauc: Caucasian, AA: African-American).

	Gender		Race			
Emotion	Male	Female	Cauc	AA	Asian	Percent.
Surprise	138	159	260	16	21	10.3%
Fear	43	36	61	5	13	2.7%
Disgust	69	89	125	6	27	5.5%
Happy	429	712	855	98	188	39.7%
Sad	147	239	291	30	65	13.4%
Angry	119	45	144	10	10	5.7%
Neutral	312	339	489	39	123	22.6%
Percent.	43.7%	56.3%	77.4%	7.1%	16.4%	

5 Results

5.1 An Analysis of Dataset Bias and Counterfactual Bias

Dataset Bias Analysis First, we conduct a preliminary bias analysis by attempting to highlight the different biases present. As highlighted in Table 1, there is a slight dataset bias across gender. 56.3% of the subjects are female, while 43.7% are male. There is a greater bias across race. 77.4% of the subjects are Caucasian, 15.5% are Asian, and only 7.1% are African-American.

Table 2. Proportion of samples that stayed consistent with the original classification after counterfactual manipulation (skin tone change). The values are the $\mathcal{M}_{CF}$ values. Classification is performed using the baseline model.

	Gender		Race		
Emotion	Male	Female	Cauc	AA	Asian
Surprise	0.34	0.31	0.33	0.38	0.24
Fear	0.16	0.25	0.16	0.20	0.38
Disgust	0.20	0.13	0.15	0.17	0.22
Happy	0.44	0.56	0.52	0.53	0.48
Sad	0.22	0.27	0.26	0.20	0.25
Angry	0.29	0.27	0.28	0.20	0.30
Neutral	0.33	0.36	0.36	0.28	0.33
$\mathcal{M}_{CF}$	0.34	0.41	0.39	0.39	0.37

Counterfactual Bias Analysis This involves calculating the proportion of the baseline model's predictions that remained the same between the original and counterfactual images. The specific formulation is captured by Equation 6. Though simple in nature, it forms a crucial cornerstone in evaluating Counterfactual Fairness. With reference to Table 2, we see that, for a majority of samples, the predicted labels did not remain the same for the counterfactual images. Across the sensitive attribute Gender, performance accuracy is slightly more consistent for Females. Across the sensitive attribute Race, performance accuracy is comparatively more consistent for Caucasians and Asians. This phenomena may be correlated with class size numbers as evidenced in Table 1. A similar trend is true across emotions. We see that the emotion "Happy" has the highest consistency. This may be due to the larger sample size and the fact that "Happy" is considered an emotion that is relatively easier to recognise and label. On the other hand, we see that the "Fear" class has the lowest consistency. This might be due to the fact that the emotion fear has lesser samples and is more ambiguous, variable and perhaps harder to identify and label. This hints that bias may not only vary across sensitive attribute but across emotion categories as well. We would like to highlight that this counterfactual analysis only analyses

Table 3. Accuracy and fairness scores from the fine-tuned standalone models trained on the combined training and tested on the combined test set which includes both the original and counterfactual images.

	1. Pre-processing					2. In-processing					3. Post-processing				
	Gender		Race			Gender		Race			Gender		Race		
Emotion	M	F	Cau	AA	A	M	F	Cau	AA	A	M	F	Cau	AA	A
Surprise	0.58	0.59	0.58	0.59	0.60	0.59	0.63	0.63	0.50	0.52	0.55	0.67	0.61	0.72	0.60
Fear	0.35	0.26	0.34	0.30	0.19	0.31	0.28	0.33	0.30	0.15	0.37	0.22	0.34	0.40	0.12
Disgust	0.29	0.26	0.26	0.33	0.31	0.25	0.31	0.30	0.33	0.24	0.24	0.35	0.28	0.25	0.39
Happy	0.75	0.76	0.72	0.87	0.84	0.89	0.92	0.90	0.90	0.91	0.90	0.93	0.92	0.88	0.94
Sad	0.44	0.48	0.47	0.35	0.51	0.45	0.49	0.47	0.42	0.50	0.46	0.47	0.46	0.37	0.52
Angry	0.52	0.47	0.52	0.45	0.30	0.51	0.43	0.50	0.50	0.25	0.48	0.37	0.45	0.45	0.40
Neutral	0.66	0.63	0.65	0.51	0.66	0.63	0.68	0.67	0.58	0.64	0.61	0.62	0.62	0.58	0.62
$\mathcal{M}_{ACC}$	0.61	0.63	0.61	0.65	0.67	0.65	0.72	0.69	0.69	0.68	0.65	0.71	0.68	0.68	0.70
$\mathcal{M}_{EO}$	**0.97**		0.91			0.91		**0.99**			0.91		0.96		
$\mathcal{M}_{CF}$	0.53	0.57	0.54	0.56	0.61	0.58	0.63	0.59	0.63	0.65	0.39	0.44	0.41	0.44	0.39
$\mathcal{M}_{CF}$ (Avg.)	0.55		0.57			**0.60**		**0.62**			0.42		0.41		

whether predictions remained consistent between the original and counterfactual images and has no bearing on whether the initial prediction was correct.

5.2 Bias Mitigation Results with Counterfactual Images

Next, we evaluate to what extent we are able to mitigate bias via the methods proposed in the Methodology section: At the pre-processing, in-processing and post-processing stage. With reference to Table 3, in terms of accuracy, there does not seem to be a difference between the pre-processing, in-processing and post-processing methods. All were able to improve accuracy prediction to largely the same effect. However, we witness a difference in outcome across $\mathcal{M}_{EO}$. According to $\mathcal{M}_{EO}$, the pre-processing method is best for achieving fairness across gender whilst the in-processing method is best for achieving fairness across race. Though this measure highlight different effectiveness, all three methods seem comparable in terms of their ability to improve $\mathcal{M}_{EO}$ across board.

It is only in terms of $\mathcal{M}_{CF}$ where we manage to observe a wider difference in performance disparity. The pre-processing and in-processing methods were able to improve $\mathcal{M}_{CF}$ to a greater extent compared to the post-processing method. Out of the first two, it is evident that the in-processing method manages to outperform the other two across both sensitive attributes. It gives the highest $\mathcal{M}_{CF}$ score of 0.60 and 0.62 across gender and race respectively compared to 0.55 and 0.57 for the pre-processing method. On the other hand, we see that the post-processing method only gives 0.42 and 0.41 across gender and race respectively. This result is noteworthy in several ways. First, this highlights the importance of using different metrics for bias evaluation as this methodological gap would not have been picked up by the other two metrics ($\mathcal{M}_{ACC}$ and $\mathcal{M}_{EO}$). Second,

Table 4. Stacked model combines the pre-, in- and post-processing stages.

	Gender		Race		
Emotion	Male	Female	Caucasian	AA	Asian
Surprise	0.68	0.74	0.73	0.72	0.55
Fear	0.47	0.44	0.47	0.50	0.38
Disgust	0.37	0.42	0.42	0.33	0.30
Happy	0.95	0.98	0.98	0.91	0.93
Sad	0.59	0.62	0.60	0.63	0.61
Angry	0.67	0.50	0.64	0.60	0.50
Neutral	0.75	0.84	0.80	0.72	0.78
$\mathcal{M}_{ACC}$	0.75	0.81	0.79	0.78	0.76
$\mathcal{M}_{EO}$	0.93		0.96		
$\mathcal{M}_{CF}$	0.59	0.65	0.62	0.61	0.63
$\hookrightarrow \mathcal{M}_{CF}(Avg.)$	0.62		0.62		

this underlines the need to use a variety of methods to tackle the problem of bias as a standalone post-processing method might be inadequate. Finally, with reference to Table 4, we see that the stacked approach improved scores across all evaluation metrics. This model comprises of a combination of the fine-tuned pre, in and post-processing methods. In Table 5, we see that the combined approach is comparatively better than all the standalone methods across most metrics. Out of the standalone methods, the in-processing method seems to be best in terms of achieving both $\mathcal{M}_{EO}$ and $\mathcal{M}_{CF}$ fairness.

Table 5. Results summary showing that a combined stacked approach supersedes all other standalone models on most metrics.

Model	$\mathcal{M}_{ACC}$	$\mathcal{M}_{F1}$	$\mathcal{M}_{EO}$	$\mathcal{M}_{CF}$
Original	0.65	0.54	**0.97**	0.30
Pre-processing	0.63	0.67	0.94	0.56
In-processing	0.69	0.68	0.95	0.61
Post-processing	0.68	0.65	0.94	0.42
Combined	**0.78**	**0.71**	0.95	**0.62**

6 Conclusion and Discussion

Overall, the stacked approach supersedes the rest across most measures: accuracy ($\mathcal{M}_{ACC}$), F1-score ($\mathcal{M}_{F1}$), and Counterfactual Fairness ($\mathcal{M}_{CF}$). A significant point is that our work agrees with the findings in [17] in many ways. One of the key findings was how pre-processing methods can have a huge effect on disparity in prediction outcomes. Second, their evaluation showed that many

of the measures of bias and fairness strongly correlate with each other. This is evident in our findings too as we see that the equal opportunity fairness correlates with class-wise performance accuracy across board. Hence, we argue the importance of using an orthogonal measure to capture the bias which would otherwise go unnoticed. In our experiments, the Counterfactual Fairness measure $\mathcal{M}_{CF}$ fulfills this criteria. Indeed, we see that it captures the efficacy difference in achieving Counterfactual Fairness across the different bias mitigation strategies. This provides empirical evidence for the gaps highlighted in Section 2.1.

Further, though the overall evaluation metrics have improved, we still observe bias when conducting a disaggregated analysis partitioned across the different sensitive attributes or emotion categories. For instance, the mitigation algorithms consistently performed poorly for certain emotion categories, e.g. "Disgust". This aligns with the findings in [25, 49] as such expressions are hypothesised to be less well-recognised than other prevalent emotions e.g. "Happy".

A key limitation of our work is that the methods that we propose are limited by dataset availability. We have used the annotations as provided by the original dataset owners [32] which were crowd-sourced and labelled by humans. However, this approach of treating race as an attribute fails to take into account the multi-dimensionality of race [21] which represents a research area that future researchers can look into. In addition, we have solely relied upon the original training-test split provided by the data repository. As highlighted in [17], algorithms are highly sensitive to variation in dataset composition and changes in training-test splits resulted in great variability in the accuracy and fairness measure performance of all algorithms. Another limitation is that of robustness and further research on adversarial attacks on fairness [38] should be investigated.

We recognise that face recognition and by its extension, facial affect recognition has received criticism for its misuse and parallelism to facial phrenology. First, we would like to underscore that the ideas in this paper are meant to address the existing problem of bias and our intention is for it to be used for good. Second, we concur that many of the concerns raised are valid and we view this as an opportunity for future work extensions. Third, we hope this piece of work will encourage researchers and companies to shape solutions that ensure that the technology and applications developed are fair and ethical for all.

Data access statement: This study involved secondary analyses of pre-existing datasets. All datasets are described in the text and cited accordingly. Licensing restrictions prevent sharing of the datasets. The authors thank Shan Li, Prof Weihong Deng and JunPing Du from the Beijing University of Posts and Telecommunications (China) for providing access to RAF-DB.
Acknowledgement: J. Cheong is supported by the Alan Turing Institute doctoral studentship and the Cambridge Commonwealth Trust. H. Gunes' work is supported by the EPSRC under grant ref. EP/R030782/1.

References

1. Barocas, S., Hardt, M., Narayanan, A.: Fairness in machine learning. NIPS tutorial **1**, 2 (2017)
2. Binns, R.: Fairness in machine learning: Lessons from political philosophy. In: Conf. on Fairness, Accountability and Transparency (2018)
3. Bromley, J., Bentz, J.W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., Shah, R.: Signature verification using a "siamese" time delay neural network. International Journal of Pattern Recognition and Artificial Intelligence **7**(04), 669–688 (1993)
4. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. pp. 77–91. PMLR (2018)
5. Cheong, J., Kalkan, S., Gunes, H.: The hitchhiker's guide to bias and fairness in facial affective signal processing: Overview and techniques. IEEE Signal Processing Magazine **38**(6), 39–49 (2021)
6. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: CVPR (2005)
7. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data **5**(2), 153–163 (2017)
8. Churamani, N., Kara, O., Gunes, H.: Domain-incremental continual learning for mitigating bias in facial expression and action unit recognition. arXiv preprint arXiv:2103.08637 (2021)
9. Crawford, K.: Time to regulate ai that interprets human emotions. Nature **592**(7853), 167–167 (2021)
10. Dash, S., Balasubramanian, V.N., Sharma, A.: Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals. In: WACV (2022)
11. Davani, A.M., Omrani, A., Kennedy, B., Atari, M., Ren, X., Dehghani, M.: Fair hate speech detection through evaluation of social group counterfactuals. arXiv preprint arXiv:2010.12779 (2020)
12. Davani, A.M., Omrani, A., Kennedy, B., Atari, M., Ren, X., Dehghani, M.: Improving counterfactual generation for fair hate speech detection. In: Workshop on Online Abuse and Harms (WOAH) (2021)
13. Denton, E., Hutchinson, B., Mitchell, M., Gebru, T.: Detecting bias with generative counterfactual face attribute augmentation. arXiv e-prints pp. arXiv–1906 (2019)
14. Dinan, E., Fan, A., Williams, A., Urbanek, J., Kiela, D., Weston, J.: Queens are powerful too: Mitigating gender bias in dialogue generation. In: EMNLP (2020)
15. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. pp. 214–226 (2012)
16. Ekman, R.: What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA (1997)
17. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. In: Conf. on fairness, accountability, and transparency (2019)
18. Gajane, P., Pechenizkiy, M.: On formalizing fairness in prediction with machine learning. arXiv preprint arXiv:1710.03184 (2017)
19. Garcia, R., Wandzik, L., Grabner, L., Krueger, J.: The harms of demographic bias in deep face recognition research. In: Proc. Int. Conf. Biometr. (ICB). pp. 1–6 (2019)

20. Gunes, H., Schuller, B.: Categorical and dimensional affect analysis in continuous input: Current trends and future directions. Image and Vision Computing **31**(2), 120–136 (2013)
21. Hanna, A., Denton, E., Smart, A., Smith-Loud, J.: Towards a critical race methodology in algorithmic fairness. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. pp. 501–512 (2020)
22. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. NIPS (2016)
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
24. Hoffman, A.: Where Fairness Fails: Data, Algorithms and the Limits of Antidiscrimination Discourse. Journal of Information, Communication & Society **22**, 900–915 (2019)
25. Howard, A., Zhang, C., Horvitz, E.: Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems. In: Proc. Adv. Robot. Social Impacts (ARSO) (2017)
26. Jain, N., Olmo, A., Sengupta, S., Manikonda, L., Kambhampati, S.: Imperfect imaganation: Implications of gans exacerbating biases on facial data augmentation and snapchat face lenses. Artificial Intelligence **304**, 103652 (2022)
27. Joo, J., Kärkkäinen, K.: Gender slopes: Counterfactual fairness for computer vision models by attribute manipulation. In: Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia (2020)
28. Kamiran, F., Karim, A., Zhang, X.: Decision theory for discrimination-aware classification. In: Int. Conference on Data Mining (2012)
29. Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., Schölkopf, B.: Avoiding discrimination through causal reasoning. In: NIPS. pp. 656–666 (2017)
30. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
31. Kusner, M., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. In: NIPS (2017)
32. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: CVPR (2017)
33. Liu, M., Ding, Y., Xia, M., Liu, X., Ding, E., Zuo, W., Wen, S.: Stgan: A unified selective transfer network for arbitrary image attribute editing. In: CVPR (2019)
34. Loftus, J.R., Russell, C., Kusner, M.J., Silva, R.: Causal reasoning for algorithmic fairness. arXiv preprint arXiv:1805.05859 (2018)
35. Lu, K., Mardziel, P., Wu, F., Amancharla, P., Datta, A.: Gender bias in neural natural language processing. In: Logic, Language, and Security, pp. 189–202. Springer (2020)
36. Maudslay, R.H., Gonen, H., Cotterell, R., Teufel, S.: It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In: EMNLP-IJCNLP (2019)
37. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) **54**(6), 1–35 (2019)
38. Mehrabi, N., Naveed, M., Morstatter, F., Galstyan, A.: Exacerbating algorithmic bias through fairness attacks. In: AAAI (2021)
39. Nabi, R., Shpitser, I.: Fair inference on outcomes. In: AAAI (2018)

40. Ngxande, M., Tapamo, J., Burke, M.: Bias remediation in driver drowsiness detection systems using generative adversarial networks. IEEE Access **8**, 55592–55601 (2020). https://doi.org/10.1109/ACCESS.2020.2981912
41. Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X.S., Wen, J.R.: Counterfactual vqa: A cause-effect look at language bias. In: CVPR (2021)
42. Pearl, J.: Causality. Cambridge university press (2009)
43. Russell, J.A.: A circumplex model of affect. Journal of personality and social psychology **39**(6), 1161 (1980)
44. Salimi, B., Rodriguez, L., Howe, B., Suciu, D.: Interventional fairness: Causal database repair for algorithmic fairness. In: International Conference on Management of Data (2019)
45. Sariyanidi, E., Gunes, H., Cavallaro, A.: Automatic analysis of facial affect: A survey of registration, representation, and recognition. IEEE TPAMI **37**(6), 1113–1133 (2014)
46. Verma, S., Rubin, J.: Fairness definitions explained. In: International workshop on software fairness (fairware). pp. 1–7. IEEE (2018)
47. Wang, W., Feng, F., He, X., Zhang, H., Chua, T.S.: Clicks can be cheating: Counterfactual recommendation for mitigating clickbait issue. In: ACM SIGIR Conference on Research and Development in Information Retrieval (2021)
48. Wong, A.: Mitigating gender bias in neural machine translation using counterfactual data. M.A. Thesis, City University of New York (2020)
49. Xu, T., White, J., Kalkan, S., Gunes, H.: Investigating bias and fairness in facial expression recognition. In: ECCV2020 Workshop: ChaLearn Looking at People (2020)
50. Zafar, M.B., Valera, I., Rogriguez, M.G., Gummadi, K.P.: Fairness constraints: Mechanisms for fair classification. In: Artificial Intelligence and Statistics. pp. 962–970. PMLR (2017)