Signal, Image and Video Processing
https://doi.org/10.1007/s11760-022-02246-8

ORIGINAL PAPER

Ethical AI in facial expression analysis: racial bias

Abdallah Hussein Sham[1] · Kadir Aktas[2,3] · Davit Rizhinashvili[2] · Danila Kuklianov[2] · Fatih Alisinanoglu[4] · Ikechukwu Ofodile[2] · Cagri Ozcinar[2] · Gholamreza Anbarjafari[2,3,5,6]

Received: 27 February 2022 / Accepted: 12 April 2022

Abstract

Facial expression recognition using deep neural networks has become very popular due to their successful performances. However, the datasets used during the development and testing of these methods lack a balanced distribution of races among the sample images. This leaves a possibility of the methods being biased toward certain races. Therefore, a concern about fairness arises, and the lack of research aimed at investigating racial bias only increases the concern. On the other hand, such bias in the method would decrease the real-world performance due to the wrong generalization. For these reasons, in this study, we investigated the racial bias within popular state-of-the-art facial expression recognition methods such as Deep Emotion, Self-Cure Network, ResNet50, InceptionV3, and DenseNet121. We compiled an elaborated dataset with images of different races, cross-checked the bias for methods trained, and tested on images of people of other races. We observed that the methods are inclined towards the races included in the training data. Moreover, an increase in the performance increases the bias as well if the training dataset is imbalanced. Some methods can make up for the bias if enough variance is provided in the training set. However, this does not mitigate the bias completely. Our findings suggest that an unbiased performance can be obtained by adding the missing races into the training data equally.

Keywords Facial expression recognition (FER) · Deep neural networks · Reaction emotion · LSTM

1 Introduction

Facial expressions are utilized in many fields such as security, medicine, social sciences, marketing, and human-machine interaction to obtain better capabilities from different aspects [1]. Therefore, automated FER became one of the widely studied computer vision topics in the last two decades [2].

With the advancement of technology, deep learning methods in the computer vision tasks [3,4] became a popular approach for tackling facial expression recognition tasks. These methods use datasets containing facial images annotated with facial expressions and deep learning techniques to characterize and learn the samples in more detail as the datasets get bigger. Due to this, they are prone to biases in the datasets [5,6]. If the collected data do not represent the variety of the real-life samples properly, then inconsistent results can be obtained, leading to a drop in performance

Abdallah Hussein Sham and Kadir Aktas are both equally led this work.

Our thanks to Pexels API for granting us the rights for the data collection of the database.

✉ Abdallah Hussein Sham
ahsham@tlu.ee

Kadir Aktas
kadir.aktas@ut.ee

Davit Rizhinashvili
davit.rizhinashvili@ut.ee

Danila Kuklianov
danila.kuklianov@ut.ee

Fatih Alisinanoglu
fatih.alisinanoglu@hku.edu.tr

Ikechukwu Ofodile
i.ofodile@ut.ee

Gholamreza Anbarjafari
shb@ut.ee

[1] Enactive Virtuality Lab, Baltic Film, Media, and Arts School, Tallinn University, Narva mnt 25, 10120 Tallinn, Estonia

[2] iCV Lab, University of Tartu, Tartu, Estonia

[3] iVCV, Tartu 51011, Estonia

[4] Faculty of Engineering, Hasan Kalyoncu University, Gaziantep, Turkey

[5] PwC Advisory, Helsinki, Finland

[6] Yildiz Technical University, Istanbul, Turkey

Published online: 09 May 2022

[7]. Additionally, a bias towards a certain race in FER raises concerns about fairness [8,9] and often leads to high-profile cases. This is especially problematic in fields such as policing and surveillance [10], where criminal charges and prosecution can be at stake. For example, in 2020 a black male was wrongfully arrested due to an inaccurate facial recognition algorithm [11].

Various researches have been completed to investigate the bias in sensitive topics such as face recognition, gender recognition, and age estimation [12–14]. In FER, different aspects are investigated to understand and mitigate the bias. In [15], the annotation bias was explored and an AU-Calibrated FER framework was proposed to remove it. They utilized facial action units (AUs) in their framework, hence the name. In [7], the authors compared a disentangled and an attribute-aware method with their baseline method to assess the bias. They also demonstrated the impact of data augmentation on the bias. They utilized RAF-DB [16] and CelebA [17] datasets in their study.

There is still a research gap in the investigation of racial bias in FER. Most of the publicly available datasets do not include the racial information of the people in the dataset. Moreover, the distribution of the races in the datasets is not equal. Additionally, the methods do not provide information about the preventions or performances regarding racial bias. Such points assess bias as a challenge [7]. To address the gap in racial bias investigation in FER, we gathered a new dataset that contains samples from four different races, i.e., Caucasian, Black, Asian Indian, and East Asian [18].

Moreover, to investigate the racial bias, we trained and tested the state-of-the-art deep learning methods and transfer learning methods in FER using different combinations of the races in our dataset. We trained and tested each network repeatedly using various combinations of races from our dataset. Our results show that the majority of the time whenever a race is missing from the training data, a lower accuracy is obtained for that race. We observed that a minority of the methods could make up for the missing races if enough variety is provided in the training data. However, this solution works only for some instances, even for these methods. In addition to these points, we observed that an increase in accuracy does not necessarily mean less bias. Sometimes, the higher accuracy results in more bias as the network optimize itself more towards a particular race. Finally, we consistently obtained the least number of biases when all the races in the dataset were included in the training. There are three main contributions in this paper:

- We compiled a new facial expressions dataset which consists of four races.
- We investigated the racial bias, which can occur in the state-of-the-art deep learning methods, to address the concern about fairness.
- We demonstrated the racial bias for each method. And, our findings show that training data should be as varied as possible to mitigate the racial bias.

The remaining of the paper is organized as follows: Sect. 2 gives an overview of the related works. Section 3 describes the dataset, the experimented methods and the training. Section 4 presents the experimental results and the discussion. Lastly, Sect. 5 concludes the paper.

2 Related work

With the advancement of AI, many are concerned there is a high chance that AI development can lead to significant and irreversible repercussions since there are substantial developments in affective computing, considerable amounts of work concerning facial expression and deep learning. In [19], the authors provided an overview of FER biases in the context of responsible AI. They focused on gender bias and stated that race is another bias that needs to be addressed. To understand the question of race and FER, we can look at the field of psychology. Since psychologists began by conducting studies covering the correlation between culture, ethnicity, gender, and cultural differences to emotions [20–25], we can use their contributions as prior work. One of the studies showed that a group of people within the same race are able to distinguish the emotions of other people who belong to either the same or different races [26]. However, there were a few situations of misclassification of people. Since computers are less prone to making mistakes as compared to humans, AI has a higher probability of outperforming humans as they can learn faster and more precise. On the other hand, outperforming humans does not make them superior as they are prone to overfitting. Consequently, we need to observe all the measures that have been set to avoid overfitting.

Secondly, we can look at the adjustments needed to make AI more fair or responsible. In [27], the authors published a position paper in which they highlighted the requirements and obstacles for responsible AI concerning two intertwined objectives: efforts toward socially beneficial applications and human and social dangers of AI systems. They also mentioned several reported cases of bias in different fields due to lack of transparency, intelligibility and biased training data. These data are biased in hidden ways challenging to discover and mitigate. Similarly, the authors of [28] explained how to create a responsible AI by design methodology. With the right technical tools, such as Facebook's Fairness Flow tool, IBM's Fairness 360 toolkit, or even Accenture's AI Fairness tool, one can detect bias in sensitive datasets and even see correlations in datasets. In [29], the researchers wanted to look at the flaws in traditional definitions of AI as a rational agent and argue that a broader set of driving ethical principles

is needed to create more socially responsible AI agents. They mentioned a flow in a credit card dataset with a high risk of racial bias. It does not contain the protected attribute race, but other personal information can be used in a discriminatory way while analyzing datasets. Livingston, M. (2020) proposed three activities that government organizations should take to forestall racial inclination in Federal AI by expanding racial variety in AI fashioners, carrying out AI sway appraisal and setting up methods for staff to challenge computerized choices. He contended that the absence of racial mixture in technologists is all because of the necessities of the affected populaces that are underrepresented in the field. Hence, AI predictions are more accurate when the model is trained over a dataset where samples from Caucasian and Black are of the same ratio [30]. We can also find similar arguments in [31,32].

Finally, we can look at the results that have been the same. From [33], the authors proposed a method that performs ethnicity classification and evaluated it in three different classification scenarios such as between Black and Caucasian people, Chinese and non-Chinese, and Han, Uyghurs and non-Chinese. Their method uses deep convolution neural networks (DCNN) to extract features and classify the extracted features simultaneously. They perform facial image processing by detecting the face, aligning, normalizing, and cropping the face. The latter is then trained using the DCNN model. They compared their method with other methods such as Biologically Inspired Features (BIF), Kernel Class-dependent Feature Analysis (KCFA), and Local Binary Pattern (LBP). They concluded that their method outperforms the other ways as it has been trained using large-scale datasets such as MORPH-II, CASIA-PEAL, CASIA-WebFace. Using the identical manner, the authors in [34] trained their model using different datasets such as the CK+, JAFFE, and BU-3DFE so that they can predict facial emotions on them. We can find similar work in [35] where the authors presented an analysis of Western Caucasian and East Asian facial expressions of emotions based on visual representations and cross-cultural FER, and in [36], the authors proposed a joint deep learning approach called racial identity aware deep convolutional neural network, one developed to recognize multicultural facial expressions.

In [18], the authors did an excellent survey about what we can learn from faces. They started from the fundamental and analytical understanding of race-based on interdisciplinary expertise. They also specify that it is good to exclude the hair and keep only the face to keep the models unbiased. In most studies, human races databases are classified as follows: African/African American, Caucasian, East Asian, Native American/American Indian, Pacific Islander, Asian Indian, and Hispanic/Latino. The authors in [18] also mentioned the different databases that are commonly related to racial face such as: CAS-PEAL, IFDB, Texas 3DFRD, KFDB, JAFEE, CAFE, FGRC 2.0, CMU DB, BU-3DFE, FERET, Cohn-Kanade, Asian PFOI, HAJJ & UMRAH, JACFEE, CUN, Indian DB, FEI, and EGA.

3 The proposed method

3.1 Database

In the scope of this study, we focus on three emotions, Happy, Neutral, and Sad, to represent Positive, Neutral, and Negative valence, respectively. We did so since imagery showing these emotions is the most common online and minimizes the risk of other biases that may occur at any point from data source till evaluation. It is also good to mention that our focus is to demonstrate the racial bias in emotional recognition and its dependence on the training dataset; we need to separate sets for different races. Hence, we compiled our database **iCV's RDB** also known as Races Database.

For this reason, we opted to compile standard size datasets of Black, East Asian, and Asian Indian (See Figs. 1, 2, and 3) races expressing one of the aforementioned emotions, and they are tabulated in Table 1. We used publicly available movies and Pexels.com API as the source of images. For those images, race and emotions were separated and labeled manually. Since the age and gender were not annotated from the source, we chose young adults and tried to keep the ratio

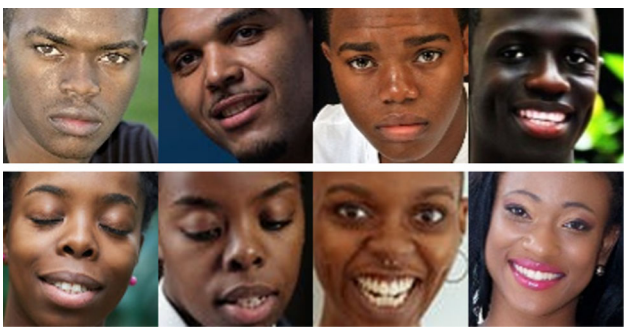

Fig. 1 Some example images from Black dataset

Fig. 2 Some example images from East Asian dataset

Table 1 Datasets, their content and description

Database	Abbr.	Total	Happy	Neutral	Sad	Male%	Female%
Caucasian	C	5020	2240	320	2240	52	48
Black	B	3900	1600	1900	800	55	45
Asian Indian	I	3900	1840	1196	908	53	47
East Asian	A	3900	1434	1320	1146	53	47

of male and female as balanced as possible. Additionally, we used ICV's Multi-Emotion Facial Expression Dataset (MEFED) Fig. 4 for Caucasian samples.

We used Python's face recognition library to crop every image as explained in [18], to keep consistent frame position of human's faces. We manually discarded the images with partially-covered or decorated faces. We made sure that to check the images and the annotations so that our data is not biased (content production bias- normally arises from lexical, semantic and structural differences in the contents generated by users[37]). In all used datasets, gender distribution was almost uniform.

As you can see in Table 1, our datasets are imbalanced regarding the samples per emotions. Therefore, we created smaller datasets with balanced class samples. To keep the emotion distribution uniform, we reduced the size of each set to only 300 images for each emotion, 900 in total. We did this so that we can avoid the population bias and investigate the effects of having imbalanced classes compared to balanced emotion samples count. The reason why we chose 300 is because from Table 1, the smallest number of samples among all the emotions is 320.

Fig. 3 Some example images from Asian Indian dataset

Fig. 4 Some example images from Caucasian dataset

3.2 Networks

Novel deep architectures and transfer learning methods are commonly used in facial expression recognition due to their good performances. Therefore, in this paper, we evaluate two state-of-the-art architectures, i.e., Deep Emotion and Self-Cure Network, and three commonly used state-of-the-art transfer learning methods, i.e., ResNet50, InceptionV3, and DenseNet121, to provide a proof of concept in our bias investigation study.

3.2.1 Deep Emotion

Deep Emotion [38] propose using an attentional convolutional network for emotion recognition. The network is structured to try and learn which region of face is more important when predicting a certain emotion. The feature extraction part consists of four convolutional layers, each two followed by max-pooling layer and ReLU activation function. They are then followed by a dropout layer and two fully connected layers. The localization network consists of two convolution layers (each followed by max-pooling and ReLU) and two fully connected layers. After regressing the transformation parameters, the input is transformed to the sampling grid. The spatial transformer module essentially tries to focus on the most relevant part of the image, by estimating a sample over the attended region.

3.2.2 Self-Cure Network

Another method utilized was Self-Cure Network [39]. In the paper, the authors detail an approach using self-attention importance weighting to relabel images in batches internally; that is, images with low importance weights can be relabeled if the maximum predicted probability exceeds the probability of a given class by a threshold. The authors have experimentally found correlation between importance weights and class uncertainty in images. The network is expected to "cure" itself by relabeling and reweighting, hence the name.

3.2.3 Transfer learning

Transfer learning is a technique which allows to use pre-trained neural networks on another task to transfer the knowledge gained during the initial training. This helps with the performance and generalization of the architecture even though the dataset at hand is small. Therefore, transfer learning is a common technique used in tasks such as emo-

tion recognition. In this paper, we evaluated three methods which utilizes three common deep CNN architectures. We built three different methods to evaluate the performances of Resnet50 [40], InceptionV3 [41], and Dense121 [42] individually. We fine-tuned each model by training on our data.

- *ResNet50*: Deep convolutional neural networks are getting used more and more in a wide range of tasks due to their good performance. However, as the problems get more complex and the networks go deeper, training becomes harder. Problems like degradation and vanishing gradient prevents a successful training [43]. ResNet50 addresses these problems by proposing to use residual blocks in the network. A residue, i.e., x, is carried to the output. This way, residual learning is done. ResNet50 includes 50 layers and other variants of the network exist as well [40].
- *InceptionV3*: Inception is a deep neural architecture which is introduced by GoogleNet to increase the computational efficiency in terms of memory and speed compared to other deep neural network architectures. With time, the model is improved step by step and several versions of it are published. InceptionV3 includes 42 layers. Feature channels in the network are highly reduced by using 1x1 convolutional kernels. Furthermore, large convolutions in the previous Inception models are split into smaller convolutions, hence decreasing the parameters. Thanks to such implementations, the training is accelerated compared to previous Inception architectures. [41]
- *DenseNet121*: DenseNet is another deep network which aims to solve vanishing gradient problem. DenseNet utilizes structures called dense blocks. Layers in a dense block have a feed-forward connection. And, in every layer, the feature maps from all of the previous layers are used as a part of the input. This way, information flow in the model is tried to be ensured by exploiting feature reuse [42].

3.3 Training

We held the same training conditions for all the methods. We used input image size as 224 $\times$ 224. All the methods accept this image size but they may resize it in their process. We used batch size of 32. We used 100 epochs. But, we utilized early stopping mechanism. The training is stopped if the loss has not improved for 10 epochs to prevent overfitting. The learning rate is selected as 0.01. But, if the loss has not improved for 5 epochs, the learning rate is divided by 10 to continue the learning.

In order to see the effects of the dataset on the racial bias, we used various combinations of training and test sets. As explained in Sect. 3.1, we have 2 groups of datasets as balanced (900 images per race) and imbalanced (5020 images for Caucasian and 3900 images for others) datasets. Each group has 4 datasets, i.e., I, A, C, B. We repeatedly trained all of our 5 models using 11 different combinations, i.e., C, B, A, I, IA, IB, IC, AB, AC, CB, IACB, of the datasets. Therefore, we had 11 $\times$ 5 = 55 training for each group of datasets. In the end, 55 $\times$ 2 = 110 models are trained using different combinations of balanced and imbalanced datasets. In all the trainings, we used the dataset split ratio of 0.6, 0.2, and 0.2 for train, validation, and test, respectively.

4 Experimental results and discussion

As our first experiment, we evaluate the racial bias of the experimented methods. For this, we trained and tested each model 11 times, using different combinations of the collected datasets. To provide a balance between the sample numbers of each class, we used the balanced datasets (with 300 samples for each emotion as described in Sect. 3.1) in this experiment. Therefore, we ensure the dataset imbalance is not a factor.

Table 2 shows the results for Deep Emotion. First row in the table shows that when the model is trained on Caucasian dataset, the model performs best on Caucasian test set. However, it performs worse in all the other datasets during the test. This is a clear indication of bias. The same results repeat for all the models that are trained on different datasets. Certainly, the method could not make up for a lack of data from the missing races in the training set, hence had a lower accuracy on them. Racial bias is eliminated only when the model is trained using a combination of the datasets. For example, training on Asian Indian and East Asian images together increases the accuracy for all the tests including Asian Indian and East Asian. However there is still a bias against the missing races. Eventually, bias is eliminated when the training is done on four datasets together as seen in the last row. Therefore we can see that the method is unbiased as much as the training data cover the different races. And, it does not make

Table 2 Results of deep emotion trained and tested on different balanced datasets with 300 images per emotion

		TESTED ON										
	ACC %	C	B	A	I	IA	IB	IC	AB	AC	CB	IACB
T	C	73	39	45	41	42	41	60	41	62	58	51
R	B	42	72	50	55	51	67	47	62	43	57	53
A	A	47	55	74	57	65	55	54	65	64	47	55
I	I	40	54	50	74	66	67	62	52	47	47	55
N	IA	40	55	72	74	73	65	62	64	61	49	64
E	IB	39	73	40	72	65	76	66	64	47	62	64
D	IC	70	55	40	70	63	62	72	46	66	67	62
	AB	38	72	74	50	64	64	54	73	62	61	60
O	AC	72	48	74	46	66	48	62	61	74	62	62
N	CB	72	70	49	50	46	62	61	62	62	72	64
	IACB	71	70	71	71	74	74	73	70	71	72	73

Table 3 Results of Self-Cure Network trained and tested on different balanced datasets with 300 images per emotion

TRAINED ON	ACC%	C	B	A	I	AI	IB	IC	AB	AC	CB	IACB
C	76	44	39	36	37	40	56	42	57	60	49
B	56	72	66	68	67	70	62	69	61	64	66
A	51	67	72	72	72	69	61	69	61	59	65
I	43	64	56	79	67	72	61	60	49	54	60
IA	47	72	71	77	74	75	62	72	59	60	67
IB	51	74	71	77	74	76	64	73	61	63	68
IC	72	66	68	78	73	72	75	67	70	69	71
AB	55	72	76	73	75	73	64	74	66	64	69
AC	74	64	73	65	69	64	70	68	74	69	69
CB	72	71	64	70	67	70	71	68	68	71	69
IACB	76	75	76	79	78	77	78	76	76	76	77

Table 4 Results of ResNet50 on 300 images per emotion

TRAINED ON	ACC %	C	B	A	I	IA	IB	IC	AB	AC	CB	IACB
C	69	41	40	41	41	41	55	40	55	55	48
B	48	67	62	65	64	66	57	64	55	58	61
A	50	64	67	68	68	66	59	66	58	57	62
I	45	63	54	75	64	69	60	59	50	54	59
IA	47	61	66	70	68	66	58	64	56	54	61
IB	49	67	62	70	66	68	60	65	56	58	62
IC	66	59	59	67	63	63	67	59	62	62	63
AB	49	63	56	64	60	63	56	60	53	56	58
AC	67	57	61	61	61	59	64	59	64	62	62
CB	62	54	44	50	47	52	56	59	53	58	52
IACB	61	58	64	67	65	62	64	61	62	60	62

up for the lack of samples from the different race during the training.

In Table 3, Self-Cure Network has slightly different results. In this case, the overall accuracy is higher and the bias is much lower compared to Deep Emotion. For the models that are trained and tested on combinations of East Asian, Asian Indian, and Black datasets, we observe a small bias among each other. Moreover, models that are trained on two datasets, i.e., Caucasian and Black, reduces the bias further. Therefore, Self-Cure Network has a better capability to generalize compared to Deep Emotion. However, bias still exists in the results. We can clearly see that the model trained on Caucasian dataset is biased toward Caucasian test samples. Also, the results show that whenever a model is trained without Caucasian samples, the accuracy for Caucasian test samples reduces. Thus, a bias for Caucasian dataset still exists, although the bias between the other races are reduced. Finally, the model which is trained on all the datasets performs without bias, similarly with Deep Emotion.

For transfer learning methods ResNet50, DenseNet121, and InceptionV3 we observe a similar trend to Self-Cure Network (See Table 4, 5, and 6). Again, there is a huge bias in the models trained on only Caucasian dataset. Models trained on combinations of East Asian, Asian Indian, and Black datasets have lower bias toward each other. However, bias is reduced properly only when all the datasets are included.

Additionally, we held another experiment to observe the effect of imbalanced dataset. In this experiment, we repeated the same training and testing process, but this time we used

Table 5 Results of DenseNet121 on 300 images per emotion

TRAINED ON	ACC %	C	B	A	I	IA	IB	IC	AB	AC	CB	IACB
C	79	49	45	40	42	44	59	47	62	64	53
B	46	55	53	55	54	55	51	54	50	50	52
A	55	60	69	73	71	66	64	64	62	57	64
I	45	67	58	79	68	73	62	62	51	56	62
IA	50	73	75	77	76	75	64	74	62	62	69
IB	48	65	72	61	66	73	54	68	60	56	61
IC	76	66	62	77	69	71	76	64	69	71	70
AB	51	73	69	77	73	75	64	71	60	62	67
AC	76	64	64	70	67	67	73	64	70	70	69
CB	79	69	60	61	61	65	70	64	69	74	67
IACB	71	72	73	71	72	71	71	72	72	71	72

Table 6 Results of InceptionV3 on 300 images per emotion

TRAINED ON	ACC %	C	B	A	I	IA	IB	IC	AB	AC	CB	IACB
C	71	44	50	46	48	45	59	47	61	58	53
B	53	71	57	65	61	68	59	64	55	62	61
A	48	66	66	72	69	69	60	66	57	57	63
I	50	68	64	79	71	74	64	66	57	60	65
IA	51	70	68	80	74	75	65	69	59	60	67
IB	52	71	62	75	68	73	63	66	57	61	65
IC	75	66	74	76	75	71	76	70	74	71	73
AB	48	72	71	75	73	74	62	72	60	60	67
AC	79	65	65	71	68	68	75	65	72	72	70
CB	72	70	67	64	66	67	68	69	70	71	68
IACB	77	72	75	79	77	76	78	74	76	75	76

Table 7 Results of Deep Emotion on full datasets

TRAINED ON	ACC %	C	B	A	I	IA	IB	IC	AB	AC	CB	IACB
C	87	47	51	53	52	51	72	48	70	58	58
B	34	85	57	69	62	75	53	66	45	56	63
A	35	53	92	68	78	64	50	72	56	45	61
I	34	60	62	93	84	72	53	56	40	47	64
IA	34	68	90	91	93	71	70	72	70	54	72
IB	35	82	65	90	68	88	69	71	52	70	71
IC	85	59	61	90	70	72	90	51	69	70	70
AB	36	80	89	70	70	70	53	90	69	71	73
AC	86	47	88	60	71	53	71	72	89	68	71
CB	82	80	55	61	50	72	71	72	69	91	72
IACB	82	83	85	86	85	84	84	84	83	83	85

the full datasets (See Table 1) instead of the balanced datasets. As explained in Sect. 3.1, full datasets include more samples but the sample distribution among the classes are not balanced. Therefore, there is a bias in class accuracies caused by the dataset. Moreover, the higher sample count in the full datasets helps with the models' learning.

For this experiment, the observed accuracies are higher compared to the previous experiment in general (See Table 7, 8, 9, 10, 11). However, bias is also more clear in this case. This is expected as the data count increases, the models can optimize better. The weights fit better to the data, as the model gets better opportunity to characterize the dataset in detail thanks to the high number of samples. From the results of Self-Cure Network and transfer learning methods, we can see that models trained on Asian Indian, East Asian, and Black now have a more clear bias compared to the previous experiment.

Table 8 Results of Self-Cure Network on full datasets

		TESTED ON										
	ACC%	C	B	A	I	IA	IB	IC	AB	AC	CB	IACB
T	C	87	44	49	53	51	48	71	46	69	67	59
R	B	37	84	70	73	72	79	54	77	53	59	65
A	A	36	75	98	79	89	77	56	86	65	54	71
I	I	31	78	68	100	84	89	63	73	48	53	68
N	IA	32	78	98	99	98	89	63	88	62	54	75
E	IB	24	82	73	99	86	90	59	77	47	51	68
D	IC	82	71	64	99	82	85	90	67	74	77	79
	AB	38	81	96	82	89	82	59	88	65	58	73
O	AC	86	60	98	76	87	68	82	79	92	74	80
N	CB	84	81	67	75	71	78	80	74	76	83	77
	IACB	86	81	97	99	98	90	92	89	91	84	91

Table 9 Results of ResNet50 on full datasets

		TESTED ON										
	ACC %	C	B	A	I	IA	IB	IC	AB	AC	CB	IACB
T	C	86	29	46	49	48	39	69	37	68	59	54
R	B	36	80	69	75	72	78	55	75	51	57	64
A	A	40	55	95	74	84	65	56	75	65	47	65
I	I	26	77	68	99	83	88	60	72	45	50	66
N	IA	38	76	97	99	98	88	66	86	65	56	76
E	IB	30	84	75	99	87	91	62	79	51	55	70
D	IC	87	76	71	100	85	88	93	73	79	82	83
	AB	32	81	97	82	89	82	56	94	62	55	72
O	AC	96	85	96	74	85	65	81	75	91	72	78
N	CB	85	81	70	74	72	77	80	75	78	83	78
	IACB	84	81	95	99	97	90	91	88	89	83	90

Table 10 Results of DenseNet121 on full datasets

		TESTED ON										
	ACC %	C	B	A	I	IA	IB	IC	AB	AC	CB	IACB
T	C	76	45	44	50	47	48	64	44	61	62	55
R	B	23	80	69	74	71	77	47	75	44	50	60
A	A	37	69	98	79	88	74	57	83	65	52	69
I	I	28	79	73	99	86	89	61	76	49	52	68
N	IA	35	79	97	99	98	89	65	88	64	56	76
E	IB	25	82	76	98	87	90	59	79	48	52	69
D	IC	85	77	75	99	87	88	92	76	81	82	84
	AB	34	79	96	80	88	80	56	88	63	55	71
O	AC	86	63	97	77	87	70	82	80	91	75	81
N	CB	70	75	63	67	65	71	69	69	67	72	69
	IACB	82	80	94	98	96	89	90	87	88	81	88

Table 11 Results of InceptionV3 on full datasets

		TESTED ON										
	ACC %	C	B	A	I	IA	IB	IC	AB	AC	CB	IACB
T	C	80	34	46	49	48	42	66	40	64	59	53
R	B	26	81	66	72	69	76	48	73	44	52	60
A	A	52	65	92	72	82	68	61	78	70	58	69
I	I	24	79	72	100	87	89	60	77	48	50	68
N	IA	37	79	96	100	98	89	66	87	64	57	76
E	IB	33	83	70	97	84	90	63	77	50	56	70
D	IC	86	74	71	99	85	87	92	73	79	80	83
	AB	35	82	94	80	87	81	56	94	62	57	71
O	AC	86	60	96	75	86	68	81	78	91	74	80
N	CB	84	82	67	75	71	78	80	74	76	83	77
	IACB	81	78	94	98	96	88	89	86	87	80	88

To sum up, there is a clear bias for Deep Emotion which inclines towards the data it is trained on. Self-Cure Network and the transfer learning methods have similar bias for Caucasian dataset, but among other datasets they perform with less bias. Between transfer learning methods, ResNet50 handles the bias slightly better. Although the difference with other methods is very small. Using only certain races in the training to increase the variety decreases the bias up to a degree. But it still does not mitigate it for all the races which are not included in the training. For all the methods, racial bias is eliminated only when all the races are included equally in the training. Furthermore, we observed that imbalanced classes does not effect the racial bias trend of a method. However, increasing the dataset size can effect the racial bias.

5 Conclusion

In view of the above, we investigated the racial bias in facial expression recognition by compiling datasets from different races, which were then used in various combinations to explore how the racial balance in the training dataset affects the racial bias of the model. We detailed the study by analyzing the test results for samples from four different races, both combined and individually. As a proof of concept, we used state-of-the-art facial expression recognition methods and transfer learning methods, i.e., Deep Emotion, Self-Cure Network, ResNet50, InceptionV3, and DenseNet121. Our bias analysis showed three significant conclusions. Firstly, the investigated methods are biased toward the races included in the training data. We included the missing races in the training phase to eliminate racial bias. Moreover, it is possible to reduce racial bias by having only some races for some methods. In other words, even if some races are not included in the training, the method can make up for it and learn in an unbiased way, but this is not always the case to mitigate the bias completely. Finally, an improvement in performance does not mean an improvement in racial bias. On the contrary, bias becomes more apparent as the networks fit the data more. It is important to note that our data and chosen methods bind our study. As far as future work is concerned, this study could be expanded by including the other datasets and methods.

Acknowledgements Our thanks to Pexels API for granting us the rights for the data collection of the database. This work has been partially supported by the EU MobilitasPluss grant (MOBTT90), Enactive Virtuality Lab, Tallinn University (2017-2022).

References

1. Roychowdhury, S., Emmons, M.: A survey of the trends in facial and expression recognition databases and methods. arXiv:1511.02407 (2015)
2. Li, S., Deng, W.: Deep facial expression recognition: a survey. In: IEEE Transactions on Affective Computing. IEEE, pp 1–20 (2020)
3. Kamińska, D., Aktas, K., Rizhinashvili, D., Kuklyanov, D., Sham, A.H., Escalera, S., Nasrollahi, K., Moeslund, T.B., Anbarjafari, G.: Two-stage recognition and beyond for compound facial emotion recognition. Electronics **10**(22), 2847 (2021)
4. Sang, D.V., Van Dat, N., et al.: Facial expression recognition using deep convolutional neural networks. In: 2017 9th International Conference on Knowledge and Systems Engineering (KSE), pp. 130–135. IEEE (2017)

5. Drozdowski, P., Rathgeb, C., Dantcheva, A., Damer, N., Busch, C.: Demographic bias in biometrics: a survey on an emerging challenge. IEEE Trans. Technol. Soc. **1**(2), 89–103 (2020)
6. Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J.R., Jurafsky, D., Goel, S.: Racial disparities in automated speech recognition. Proc. Natl. Acad. Sci. **117**(14), 7684–7689 (2020)
7. Xu, T., White, J., Kalkan, S., Gunes, H.: Investigating bias and fairness in facial expression recognition. In: European Conference on Computer Vision, pp. 506–523. Springer (2020)
8. Yang, K., Qinami, K., Fei-Fei, L., Deng, J., Russakovsky, O.: Towards fairer datasets: filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 547–558 (2020)
9. De Vries, T., Misra, I., Wang, C., Van der Maaten, L.: Does object recognition work for everyone? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 52–59 (2019)
10. Kasapoglu, T., Masso, A.: Attaining security through algorithms: perspectives of refugees and data experts. In: Theorizing Criminality and Policing in the Digital Media Age. Emerald Publishing Limited (2021)
11. Perkowitz, S.: The bias in the machine: Facial recognition technology and racial disparities. MIT Case Studies in Social and Ethical Responsibilities of Computing, no. Winter. https://mit-serc.pubpub.org/pub/bias-in-machine (2021)
12. Robinson, J.P., Livitz, G., Henon, Y., Qin, C., Fu, Y., Timoner, S.: Face recognition: too bias, or not too bias?" In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–1 (2020)
13. Das, A., Dantcheva, A., Bremond, F.: Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops (2018)
14. Guo, G., Mu, G.: Human age estimation: What is the influence across race and gender? In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, pp. 71–78. IEEE (2010)
15. Chen, Y., Joo, J.: Understanding and mitigating annotation bias in facial expression recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 14980–14991 (2021)
16. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2852–2861 (2017)
17. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3730–3738 (2015)
18. Fu, S., He, H., Hou, Z.-G.: Learning race from face: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **36**(12), 2483–2509 (2014)
19. Domnich, A., Anbarjafari, G.: Responsible ai: gender bias assessment in emotion recognition. arXiv:2103.11436 (2021)
20. Conley, M.I., Dellarco, D.V., Rubien-Thomas, E., Cohen, A.O., Cervera, A., Tottenham, N., Casey, B.: The racially diverse affective expression (radiate) face stimulus set. Psychiatry Res. **270**, 1059–1067 (2018)
21. Dailey, M.N., Joyce, C., Lyons, M.J., Kamachi, M., Ishi, H., Gyoba, J., Cottrell, G.W.: Evidence and a computational explanation of cultural differences in facial expression recognition. Emotion **10**(6), 874 (2010)
22. Fischer, A.H., Rodriguez Mosquera, P.M., Van Vianen, A.E., Manstead, A.S.: Gender and culture differences in emotion. Emotion **4**(1), 87 (2004)
23. Laurence, S., Zhou, X., Mondloch, C.J.: The flip side of the other-race coin: they all look different to me. Br. J. Psychol. **107**(2), 374–388 (2016)
24. Prado, C., Mellor, D., Byrne, L.K., Wilson, C., Xu, X., Liu, H.: Facial emotion recognition: a cross-cultural comparison of Chinese, Chinese living in Australia, and Anglo-Australians. Motiv. Emot. **38**(3), 420–428 (2014)
25. Strohminger, N., Gray, K., Chituc, V., Heffner, J., Schein, C., Heagins, T.B.: The mr2: a multi-racial, mega-resolution database of facial stimuli. Behav. Res. Methods **48**(3), 1197–1204 (2016)
26. Shimoda, K., Argyle, M., Bitti, P.R.: The intercultural recognition of emotional expressions by three national racial groups: English, Italian and Japanese. Eur. J. Soc. Psychol. **8**(2), 169–179 (1978)
27. Ghallab, M.: Responsible ai: requirements and challenges. AI Perspect. **1**(1), 1–7 (2019)
28. Benjamins, R., Barbado, A., Sierra, D.: Responsible ai by design in practice. arXiv:1909.12838 (2019)
29. Vetrò, A., Santangelo, A., Beretta, E., De Martin, J.C.: Ai: from rational agents to socially responsible agents. Digital Policy, Regulation and Governance (2019)
30. Livingston, M.: Preventing racial bias in federal ai, JSPG, vol. 16 (2020)
31. Benjamins, R.: A choices framework for the responsible use of ai. AI Ethics **1**(1), 49–53 (2021)
32. Shneiderman, B.: Responsible ai: bridging from ethics to practice. Commun. ACM **64**(8), 32–35 (2021)
33. Wang, W., He, F., Zhao, Q.: Facial ethnicity classification with deep convolutional neural networks. In: Chinese Conference on Biometric Recognition, pp. 176–185. Springer (2016)
34. Lopes, A.T., De Aguiar, E., De Souza, A.F., Oliveira-Santos, T.: Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern Recogn. **61**, 610–628 (2017)
35. Benitez-Garcia, G., Nakamura, T., Kaneko, M.: Multicultural facial expression recognition based on differences of Western-Caucasian and East-Asian facial expressions of emotions. IEICE Trans. Inf. Syst. **101**(5), 1317–1324 (2018)
36. Sohail, M., Ali, G., Rashid, J., Ahmad, I., Almotiri, S.H., AlGhamdi, M.A., Nagra, A.A., Masood, K.: Racial identity-aware facial expression recognition using deep convolutional neural networks. Appl. Sci. **12**(1), 88 (2022)
37. Olteanu, A., Castillo, C., Diaz, F., Kıcıman, E.: Social data: biases, methodological pitfalls, and ethical boundaries. Front. Big Data **2**, 13 (2019)
38. Minaee, S., Minaei, M., Abdolrashidi, A.: Deep-emotion: facial expression recognition using attentional convolutional network. Sensors **21**(9), 3046 (2021)
39. Kai, W., Xiaojiang, P., Jianfei, Y., Shijian, L., Yu, Q: Suppressing uncertainties for large-scale facial expression recognition. arXiv:2002.10392 (2020)
40. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)
41. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna,Z.: Rethinking the inception architecture for computer vision (2015)
42. Huang, G., Liu, Z., Weinberger, K.Q.: Densely connected convolutional networks. CoRR, vol. abs/1608.06993. http://arxiv.org/abs/1608.06993 (2016)
43. Hochreiter, S.: The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int. J. Uncertain. Fuzzin. Knowl. Based Syst. **6**(02), 107–116 (1998)