Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES 2024)

Breaking the Global North Stereotype: A Global South-centric Benchmark Dataset for Auditing and Mitigating Biases in Facial Recognition Systems

Siddharth Jaiswal[1], Animesh Ganai[1], Abhisek Dash[2], Saptarshi Ghosh[1], Animesh Mukherjee[1]
[1]Indian Institute of Technology Kharagpur, India
[2]Max Planck Institute for Software Systems, Germany
siddsjaiswal@kgpian.iitkgp.ac.in, animeshganai0000@gmail.com, adash@mpi-sws.org,
saptarshi@cse.iitkgp.ac.in, animeshm@cse.iitkgp.ac.in

Abstract

Facial Recognition Systems (FRSs) are being developed and deployed all around the world at unprecedented rates. Most platforms are designed in a limited set of countries, but deployed in other regions too, without adequate checkpoints for region-specific requirements. This is especially problematic for Global South countries which lack strong legislation to safeguard persons facing disparate performance of these systems. A combination of unavailability of datasets, lack of understanding of how FRSs function and low-resource bias mitigation measures accentuate the problems at hand. In this work, we propose a self-curated face dataset composed of 6,579 unique male and female sports-persons (cricket players) from eight countries around the world. More than 50% of the dataset is composed of individuals from the Global South countries and is demographically diverse. To aid adversarial audits and robust model training, we curate four adversarial variants of each image in the dataset, leading to more than 40,000 distinct images. We also use this dataset to benchmark five popular facial recognition systems (FRSs), including both commercial and open-source FRSs, for the task of gender prediction (and country prediction for one of the open-source models as an example of red-teaming). Experiments on industrial FRSs reveal accuracies ranging from 98.2% (in case of Azure) to 38.1% (in case of Face++), with a large disparity between males and females in the Global South (max difference of 38.5% in case of Face++). Biases are also observed in all FRSs between females of the Global North and South (max difference of $\approx$ 50%). A Grad-CAM analysis shows that the nose, forehead and mouth are the regions of interest for one of the open-source FRSs. Based on this crucial observation, we design simple, low-resource bias mitigation solutions using few-shot and novel contrastive learning techniques that demonstrate a significant improvement in accuracy with disparity between males and females reducing from 50% to 1.5% in one of the settings. For the red-teaming experiment using the open-source Deepface model we observe that simple fine-tuning is not very useful while contrastive learning brings steady benefits.

Introduction

Artificial Intelligence (AI) systems are being developed and deployed at an unprecedented rate around the world for various applications ranging from face recognition (Face++ 2023; Amazon 2023) to web search (Google 2023) and chatbots (OpenAI 2023). Often, the development is done in a single place but the deployment is done worldwide. For example, facial recognition system AWS Rekognition (Amazon 2023) has been developed by Amazon in the USA but is deployed in other countries for tasks like ID verification (Sur 2023) at airports, with no clear distinction between different geographical deployments. Thus, if the AI model is designed without considering the deployment context, the system may end up propagating stereotypes (Olier and Spadavecchia 2022) and biases (Buolamwini and Gebru 2018).

Facial recognition systems: Face recognition is a group of classification tasks that involve detecting a face in an image followed by downstream tasks like gender/age/emotion detection and/or matching the input face against images stored in a database (face identification). Facial recognition systems (FRSs) are increasingly deployed for highly sensitive applications like surveillance (Kamgar-Parsi, Lawson, and Kamgar-Parsi 2011), person re-identification (Rao, Lu, and Zhou 2019), policing (Krueckeberg and Ferris 2018), etc. This has put FRSs under the scanner of researchers and policymakers alike (Fitzpatrick 2023; Dizikes 2023).

Impact on deployment in the Global South: FRSs have historically been developed in countries which are defined by the UN as economically developed nations, i.e., the Global North (UN 2022) that includes Western countries like USA, UK and Germany, and others like Australia, Japan and Korea. A majority of the large-scale training datasets also consist of faces from this socio-economic region (Raji et al. 2020; Ma, Correll, and Wittenbrink 2015; Ma, Kantner, and Wittenbrink 2020; Zhang, Song, and Qi 2017; Yang et al. 2016; Karkkainen and Joo 2021; Parkhi, Vedaldi, and Zisserman 2015; Rothe, Timofte, and Van Gool 2018; Eidinger, Enbar, and Hassner 2014). Due to economical pricing, and relaxed licensing rules, commercial (Amazon 2023; Face++ 2023; Microsoft 2023) and open-source systems (Parkhi, Vedaldi, and Zisserman 2015; Umagat 2023; Serengil and Ozpinar 2021) trained on such datasets are available all around the world for use by individuals, governments and corporations alike. Multiple studies have reported large-scale biases in these platforms against minority races and genders in the Global North (Buolamwini and Gebru 2018; Raji et al. 2020; Jaiswal et al. 2022; Dooley et al. 2022). The US and EU have recently strengthened their rules and introduced various measures to ensure fair and proper use of such AI platforms (NIST 2020; Parliament 2023), but the Global South countries are still lagging behind in terms of such interventions. Thus, unrestricted deployment of FRSs in the Global South may have more serious ramifications like denial of services or facilities (Bansal 2022), without legal recourse and impacting individuals' quality of life.

A face dataset from the Global South: Developing and benchmarking any deep learning model like FRSs, requires a large amount of well-annotated data. Unfortunately, there is a lack of such face image data from the Global South, be it South Asia or South America. To evaluate and improve model efficiency such datasets are very important. This brings us to the first research gap in literature– *Unavailability of well-annotated data for faces from the Global South.* Moreover, the existing face datasets do not have adversarial realistic variants for real-world training. Prior research (Jaiswal et al. 2022; Dooley et al. 2022) has shown that FRS biases increase manifold for such adversarial input. Thus, the second gap that we identify is– *Unavailability of realistic adversarial variants for face images.* While FRSs may perform well on standard color images, they may be confounded on realistic adversarial inputs that can impact real-world deployment. To bridge the gaps stated above, we present a new inclusive, adversarial and robust dataset – **FARFACE**, of cricketers (famous sports individuals, see Figure 1 for some examples), which is composed of more than 50% individuals from the Global South. This dataset allows for fair, adversarial and robust face recognition model training (as our experimental results indicate later).

Benchmarking FRSs for the Global South: The first step toward evaluating the efficacy of any AI model is to benchmark it for a specific deployment scenario. In this work, we benchmark FRS models for the task of *gender prediction*. This is one of the most commonly deployed applications of FRSs, used for personalized recommendations (Chen et al. 2007; Dai, Ye, and Gong 2009), surveillance, assistive technologies, safety, and marketing (Ali et al. 2019). Some of these use cases are sensitive, and any discriminatory performance can have a damaging impact on the targets (Buolamwini and Gebru 2018; Raji et al. 2020; Jaiswal et al. 2022). Since the FRSs are often developed in one region (geographical or socio-economic) and deployed in many, without any target-specific modifications, it is important to benchmark these models' performance, especially in the Global South, where aggrieved individuals may have fewer options for recourse, legal or otherwise. This brings us to our first research question, **(RQ1.)** *How do existing Facial Recognition Systems perform on the* FARFACE *dataset, which has more than half of the faces from the Global South?*

Explaining predictions of FRSs: As a next step, we attempted to understand how a model converges on its prediction. Note that this question is relevant only for the open-source models, where the internal operations are known. We use a state-of-the-art explainability tool for computer vision tasks – Grad-CAM (Selvaraju et al. 2017), to identify the regions of interest in the images for the Deepface FRS by analysing the activation maps. This gives us **(RQ2.)** *Why do the misclassifications occur for different faces in the case of open-source models?* A suitable answer to this question could play a key role in designing solutions that can reduce the number of misclassifications and, hence, the disparity.

Figure 1: Images from our FARFACE dataset. The first row has images from Global North – Australia, New Zealand, England and South Africa. The second row has images from the Global South – India, Bangladesh, Pakistan and West Indies. The third row shows the average face for each region – Global North male, Global North female, Global South male and Global South female, generated by superimposing the images of individuals from each region.

Mitigating biases in FRSs for the Global South: Despite multiple reports on biases in FRSs for sensitive tasks like gender prediction, commercial vendors continue to develop and release these models for public use. As these black-box models cannot be debiased, researchers and activists have resorted to either calling for a complete ban of FRS usage (for the Future 2023) or various bias mitigation measures on open-source models. It is already known that training a deep model from scratch is extremely data and resource intensive and incurs a significant energy cost impacting the environment, hence re-training models may not be a good technique. It is beneficial to fine-tune existing pre-trained models to adapt them for a particular task and domain. As already pointed out, there is a significant lack of proper large-scale datasets from the Global South, a gap we aim to address with our work. We use low-resource, low-data and smart techniques like few-shot and contrastive learning to improve the efficiency of the Deepface FRS, with a specific focus on reducing the reported bias against individuals from the Global South. To perform this, we utilize our FARFace dataset, a more inclusive, complementary dataset to existing ones, with all its adversarial variants. Thus, our final research question is **(RQ3.)** *Can existing biases and lack of robustness in open-source FRS models be mitigated, leading to an overall improved performance, using simple interven-*

tions?

Our contributions: Here, we present a new large-scale, geographically diverse and adversarial face dataset FARFACE and benchmark it on *five* facial recognition systems – three commercial ones viz. Amazon AWS Rekognition (Amazon 2023), Microsoft Azure Face (Microsoft 2023) and Face++ (Face++ 2023), and two open-source ones viz. DeepFace (Serengil and Ozpinar 2021) and Libfaceid (Umagat 2023) – for the gender prediction task, for both the standard condition (normal face images), as well as for various adversarial conditions – noisy filters of RGB, Greyscale, Spread, and an occlusion filter of face mask. We observe that the gender prediction accuracy for females in the Global South is worse consistently across the different FRSs. To further understand the reasons behind the misclassifications (for the open-source models), we use the Grad-CAM explainability tool and analyse its activation maps. Finally, we adopt two different paradigms, (a) few shot learning and (b) contrastive learning, to mitigate the biases observed in the open source FRS. We show that our fine-tuned model also transfers well to other datasets from the Global North, thus being equally useful in that region. We also perform an alternative red-teaming experiment with the Deepface model to expose biases in tasks other than gender prediction– we use the Deepface model for the task of country prediction (and by extensions, ethnicity) from face images (Batsukh and Tsend 2016; AlBdairi, Xiao, and Alghaili 2020), especially for individuals from the Global South. The results from this experiment reinforce the cautionary message that such tasks, while easy to design, are unfair and biased and should not be propagated. The primary insights from our work are as follows:

• Our extensive experiments show that the disparity in accuracy between males and females for gender prediction is lower in the Global North (maximum disparity of 10% for commercial FRS– Face++ and 66.77% for open-source FRS– Libfaceid) than in the Global South (maximum disparity 38.51% for commercial FRS– Face++ and 83.35% for open-source FRS– Libfaceid).

• There is a consistent trend of high disparity for adversarial inputs of females from Global South (similar to Jaiswal et al. (2022)), indicating a lack of model robustness. Grad-CAM analysis shows systematic regions of interest for male classification, whereas they are random for females.

• Finally, our proposed approaches to overcome the gender disparity in FRSs (for gender prediction task) by adopting few-shot and contrastive learning demonstrate improvements in female accuracy by 59% and 60%, respectively.

• As an additional red-teaming task, we also study the performance of FRSs for predicting the country from the face. Our experiment shows that such a task is highly biased and can lead to unfavourable outcomes for individuals from both the Global North and South. Simple fine-tuning is not very useful; however, we get benefits in both accuracy and disparity reduction when contrastive learning is effectively used.

Dataset availability: The FARFACE dataset is available for research purposes, on request[1]. We would like to state here that the original dataset collected by us is gender imbalanced (86% males). This is reflective of the gender imbalance in society, including sports (Singh 2023). To prevent further misuse, we are releasing only a balanced subset (with all corresponding adversarial variants).

[1]Dataset Request: https://forms.gle/2Nd9ntNcc71vvwEJ7

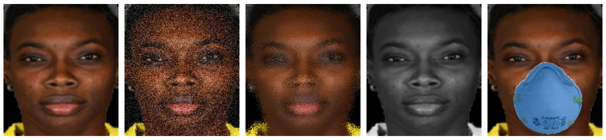
Background & Related Work

We now present a brief overview of the literature on face datasets and bias mitigation in FRSs.

Face datasets: There exist a multitude of face datasets in literature meant for different face recognition tasks. Some are designed specifically for audit studies (Buolamwini and Gebru 2018) and could be balanced (Raji et al. 2020), be collected from volunteers (Ma, Correll, and Wittenbrink 2015; Ma, Kantner, and Wittenbrink 2020; Lakshmi et al. 2021), have multiple age (Zhang, Song, and Qi 2017; Eidinger, Enbar, and Hassner 2014; Parkhi, Vedaldi, and Zisserman 2015) and racial (Rothe, Timofte, and Van Gool 2018; Zhang, Song, and Qi 2017; Karkkainen and Joo 2021) groups. All these datasets have images mainly of individuals from the Global North and have no adversarial images for robust model training or testing. We *address both these gaps* through our FARFACE dataset.

Audit & bias mitigation on FRSs: Audits on commercial FRSs (Buolamwini and Gebru 2018; Raji et al. 2020; Jaiswal et al. 2022) and open-source FRSs (Dooley et al. 2022), have shown temporal persistence to biases. Adversarial audits (Jaiswal et al. 2022; Dooley et al. 2022; Majumdar et al. 2021) on existing datasets show that disparities get magnified for noisy inputs. Our audit lies at the intersection of (Jaiswal et al. 2022) and (Dooley et al. 2022; Majumdar et al. 2021) – we audit both commercial and open-source FRSs for the task of gender prediction. Similar to Majumdar et al. (2021), we also study the activation maps for the prediction outputs but for the task of gender prediction instead of landmark detection. Another prior work by Wang et al. (2022) proposes bias mitigation strategies by developing adversarial filters on input images in real time. However, our approach differs from the mentioned approach (FAAP (Wang et al. 2022)) in the following ways: (i) Their adversarial filters do not simulate any real-world situation, and the authors do not disclose what type of filter is developed, hence it is difficult to compare against or audit for; (ii) Their adversarial filters conceal the gender and race information and improve performance for tasks like *smile* or *hair color* identification, whereas we explicitly audit the platforms for gender prediction, which has a large number of downstream applications ranging from marketing to surveillance. Furthermore, most existing approaches are either too complex or need significant changes to the pipeline (Gong, Liu, and Jain 2020; Wang and Deng 2020; Wang, Zhang, and Deng 2021; Conti et al. 2022; Wang et al. 2022). We, on the other hand, perform simple fine-tuning and contrastive learning to mitigate the observed biases.

Thus, in this work, we perform all three parts of a Responsible AI pipeline for FRSs viz. (i) benchmark audits of commercial and open-source FRSs on a large scale inclusive, adversarial dataset for gender bias (if any), (ii) explaining model predictions to better understand the reason behind



(a) ORIG (b) $RGB_{0.3}$ (c) Spread (d) Greyscale (e) N95 mask

Figure 2: Adversarial variants in the FARFACE dataset, shown for an example image (original image in (a)).

misclassifications of images of several demographics and, (iii) mitigation of the said biases.

FARFACE Dataset

In this section, we describe the collection & preparation process of the FARFACE dataset followed by its basic demographic distribution based on *gender* and *geographic region*.

Description of the Dataset

FARFACE is curated from the face images of male and female cricket players belonging to eight prominent cricket-playing countries around the world – Australia, New Zealand, England, South Africa (Global North) and, Bangladesh, Pakistan, India and West Indies (Global South). We collect these images by scraping the ESPNCRICINFO website's player pages[2] for *all* domestic and international cricketers of the above countries. While there are multiple other cricketing nations all around the world, our choice was guided by the following reasons– (i) most players in the Global North countries have lighter skin tones, whereas a majority of the players in the Global South countries have darker skin tones, giving us more diversity in terms of skin tone (see Fig. 1), face structure and geography, and (ii) these countries have both male and female cricketers, giving us gender diversity. Table 1 presents the number of images, and gender distribution for all the countries.

We next describe the methodology of dataset collection and generating their adversarial variants.

Dataset Curation

To collect the dataset, we use the Selenium tool to collect the images and the player metadata like country, name and gender. We perform the following steps to clean and preprocess the images:

• We remove all images which have drawings instead of photographs. We also set aside all images that are greyscale by default (these are added back later).

• We crop and resize all resulting images to display only the face area using YOLOv5 (Jocher et al. 2022). The resized images have a resolution of 200×256px. We choose this resolution heuristically as it creates the least distortion of the images. YOLOv5 also provides the bounding box coordinates of the final faces. We will release only the cropped images themselves.

We are finally left with 5,648 male images and 931 female images, all in JPEG format. We henceforth refer to this set of original images as the ORIG set.

Dataset accuracy: We manually checked the country, name and gender for a randomly chosen 15% subset of the dataset, and matched the data with the image. The information was accurate in all cases, showing the sanctity of the dataset.

Adversarial variants: After the above ORIG set of images was prepared, we used the GIMP[3] image editor to create the following adversarial variants – RGB (with noise values 0.3 and 0.5), Spread and Greyscale. RGB and Spread have been used previously by Jaiswal et al. (2022) to create realistic adversarial inputs for auditing FRSs. RGB simulates the effect of edited images from social media, and Spread simulates the effect of blurry photos captured by cameras, which may be exposed to natural elements like rainwater. We introduce Greyscale for the first time to simulate the effect of a black and white photograph, thus removing the skin-tone of the individual in the photograph. We also use a popular tool MaskTheFace (Anwar and Raychowdhury 2020) to create *masked* variants of the ORIG set by applying blue N95 face masks on the images, to create an occlusion-based adversary. Example images for each of the adversarial variants are shown in Figure 2.

Basic Statistics

Our downloaded set has 7,324 images, which reduces to 6,579 after cleaning. In the cleaned dataset, 85.85% images are of males, and 14.15% are of females. The size of the two RGB sets and the Spread set are the same as the ORIG set. The Greyscale set is slightly larger as it also includes some images which were greyscale by default. The Masked set is slightly smaller because the masking tool did not identify faces in all images. Thus, we effectively get 40,055 images in the FARFACE dataset after generating the five adversarial variants from the original set of images. The gender distribution in this final dataset is 85.98% males and 14.02% females. Some sample images are present in Figure 1.

Task description: While FRSs can be used for multiple tasks like face detection, gender/age detection and identification, in this paper, we only perform gender detection from the input face since we have the ground truth data for only this task. Even though the player profiles are labelled with age, we do not experiment for it as there is no guarantee on the age being the same as when the photographs were clicked. Further, a manual inspection reveals that a majority of the dataset is of young people (expected of sportspersons) and may cause a model to perform poorly when predicting the age for old people. We acknowledge that gender is a spectrum, but in this work, we consider it to be binary because all FRS models only predict either male or female as a label for the task of gender prediction (Keyes 2018). The ground truth labels for all individuals in our dataset are also only male or female. We also do not study face verification/identification because we only have 1 image per identity.

As an additional red-teaming task, we also perform country prediction from the input face for one of the open-source

[2] https://www.espncricinfo.com/cricketers

[3] https://www.gimp.org/

Region	Country	Type of Image						Total (A×4 + B + C)
		ORIG / $RGB_{0.3}$ / $RGB_{0.5}$ / SPRD (A)		GREY (B)		MASK (C)		
		Male	Female	Male	Female	Male	Female	
GN	Australia	666	159	759	181	651	158	5,049
	New Zealand	322	232	365	232	312	221	3,346
	England	947	112	1,276	133	917	111	6,673
	South Africa	444	59	482	59	422	58	3,033
GS	Bangladesh	328	39	328	39	319	39	2,193
	India	1,966	187	2,066	188	1,966	187	13,019
	Pakistan	442	79	478	79	435	78	3,154
	West Indies	533	64	584	65	488	63	3,588
All		**5,648**	**931**	**6,338**	**976**	**5,510**	**915**	**40,055**

Table 1: Distribution of images across countries, gender and image type in the FARFACE dataset. A majority of the dataset is composed of images from the Global South countries (54.8%), and has male individuals (85.98%). Among all the types, greyscale type images have the largest share (18.25%).

FRSs. The results show that such tasks are not only inaccurate but also highly biased.

FRS Platforms & Audit Methodology

We now give a brief overview of the Face Recognition tools/platforms we audit using our benchmark dataset. Next we describe the methodology of our audits, Grad-CAM analysis and bias mitigation on one of the classical and highly accurate face detection models– VGG-Face (Parkhi, Vedaldi, and Zisserman 2015). We use the implementation in Deepface (Serengil and Ozpinar 2021)[4], which is a modified version of VGG-Face.

Platforms Audited

In this study, we audit five FRSs– three popular, economically-priced plug-and-play API-based commercial FRSs and two popular pre-trained open-source FRSs.

Commercial FRSs: We perform our audits on Amazon AWS Rekognition (Amazon 2023), Microsoft Azure Face (Microsoft 2023) and Face++ (Face++ 2023). These models are made available through easy-to-use APIs and charge nominal fees for their services; thus are easy to deploy for non-domain experts at scale. However, no information is shared publicly on the training dataset or the model architecture, making it impossible for third-party researchers to address any performance or bias issues. Multiple previous audits (Buolamwini and Gebru 2018; Raji et al. 2020; Jaiswal et al. 2022) have exposed large-scale biases in these platforms for standard benchmark datasets.

Open-source FRSs: We audit two popular pre-trained open-source models– Libfaceid (Umagat 2023)[5] (a modified version of CaffeNet (Krizhevsky, Sutskever, and Hinton 2012)) and Deepface (Serengil and Ozpinar 2021)[4] (implementation of VGG-Face (Parkhi, Vedaldi, and Zisserman 2015)). The training data information for Libfaceid is not available; Deepface has been pre-trained on two datasets– VGG-Face for the initial face recognition task and WikiData (from IMDB-Wiki (Rothe, Timofte, and Van Gool 2018)) for the gender prediction task. Both these datasets have $> 50\%$ faces of White individuals (primarily found in Global North countries). As there are no licensing restrictions on their deployment, these models can propagate their pre-training biases, if deployed in geographies which are not represented well during training. These models are available for download and use, but require domain knowledge for deployment. They can be fine-tuned and modified for specific usecases allowing improvement of performance and reducing observed biases. We can extract output explanations to better understand the model's decisions. Both employ standard CNN-based architectures with Libfaceid having 11 layers (3 convolution layers) and Deepface having 22 layers (13 convolution layers); Deepface is a deeper model and captures more detailed facial features compared to Libfaceid.

Methodology of Experiments

In this work, we perform three sets of experiments described as follows.

- **Audit of all FRSs**: We audit all five FRSs on the entire FARFACE dataset for the task of gender prediction. Each FRS is supplied with raw image files in JPEG format and the response for gender prediction is collected.
- **Grad-CAM analysis of Deepface**: We analyse the predictions for the Deepface model using the Grad-CAM explainability tool to better understand the correct and incorrect predictions for the two socio-political regions and genders.
- **Bias-mitigation of Deepface**: As has been observed in previous literature, and as we shall also show through our audits in the subsequent sections, all FRSs demonstrate biases against the females and more so in the Global South (faces of individuals with darker skin tones and different facial structures). We shall use two simple techniques to show that it is quite straightforward to mitigate the observed biases in such open-source models. We choose Deepface for this as it is a deeper model and is better designed to detect more interesting facial features as opposed to Libfaceid; thus an analysis on this model is expected be more generalizable.

(a) Few-shot fine-tuning: We perform one-shot and two-shot fine-tuning on the pre-trained Deepface model. To choose our one-shot data points, we randomly sample one

[4]https://github.com/serengil/deepface

[5]https://github.com/richmondu/libfaceid

image from each country and gender, thereby giving us 16 unique data points; similarly 32 unique data points are chosen for the two-shot scenario. For each learning setup, we have three settings – involving 0%, 50% and 100% adversarial examples from the $\text{RGB}_{0.3}$ set. In our experiments we find the RGB filter to be most adversarial resulting in the highest performance drop (see Figure 3), and hence the choice. This allows us to target two objectives– reducing observed gender and demographic bias and improving model robustness to adversarial inputs. The hyperparameters are – $\alpha : 1e{-}5$, epochs: 10, optimizer: ADAM.

(b) Contrastive learning: Contrastive learning has proved to be very effective in handling adversarial inputs (Aggarwal et al. 2023). In this paper we take dual advantage of the contrastive learning setup. While on one hand, we teach the model the contrast between males and females, on the other hand we also teach it that the original and the adversarial versions of a face are the same. Based on this intuition, we propose a triplet loss (Schroff, Kalenichenko, and Philbin 2015) function to perform a *novel* contrastive learning based fine-tuning. The anchor points (any data point in a given class) (x_a) are chosen from each gender and country, the positive example for each anchor (x_a^+) is an adversarial image of the same individual, whereas the negative example (x_a^-) is an image of another person from the ORIG set of the opposite gender. We choose the (x_a^+) from $\text{RGB}_{0.3}$ in this case to improve model robustness while reducing bias simultaneously and we choose the (x_a^-) from the opposite gender to create a stronger negative example for every anchor. We also experiment with a setup where the negative example is chosen from the same or opposite gender with a probability (more results are shared in an extended version of the paper[6]). The hyperparameters are – $\alpha : 1e{-}5$, epochs: 40, optimizer: ADAM, $L_{triplet} : L_{BCE}$ = 0.8:0.2.

Held-out test set for all experiments: A held out test set of 480 images is created, composed of 60 images from each country, in a 2:1 ratio of males to females. Each setting is evaluated three times, and the avg. percentages are reported. We run our experiments on an Ubuntu 18.04 LTS Intel(R) Xeon(R) Gold 6126 CPU server with NVIDIA Tesla P100 GPU (CUDA v11.4), 128 GB RAM and 48 cores.

Results & Observations

We now present the results of our benchmarking audit, followed by the Grad-CAM analysis, bias mitigation and redteaming task of country prediction.

Benchmark Audit of FRSs (RQ1)

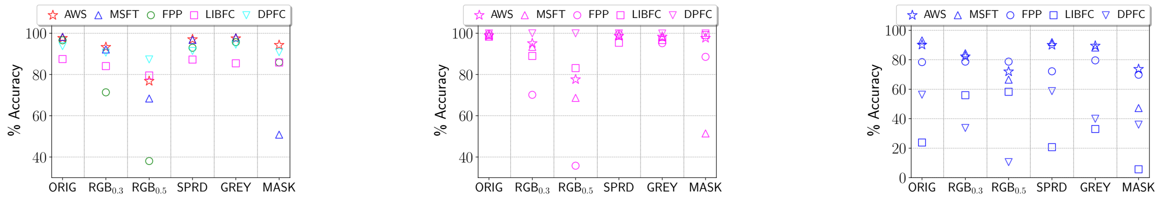
In Figure 3, we show the overall accuracy and the genderwise results of the different FRSs for the benchmark audit of FARFACE dataset separated by image type. From Fig. 3a, we observe that all FRSs perform well on the original set (ORIG), with Microsoft Azure Face being the best and identifying the correct gender for 98.22% of the images. On the other hand, Libfaceid correctly identifies the gender in 87.48% images. All FRSs report low accuracies for the RGB variants (highest – 93.3% on AWS for $\text{RGB}_{0.3}$; lowest – 38.05% on Face++ for $\text{RGB}_{0.5}$), but are robust to the Spread and Greyscale variants (performance on both these variants is comparable to the ORIG set). For the masked variant, Azure reports the lowest accuracy of 51%, with others reporting $> 85\%$. Among the FRSs studied in this paper, AWS Rekognition and Deepface are the best performing commercial and open-source FRSs respectively. Commercial FRSs perform better than open-source ones on five of the image sets. Deepface open-source FRS is the best performing FRS for the $\text{RGB}_{0.5}$ variant. We calculate the standard deviation of the accuracy distribution across different image sets for each FRS. In general, commercial FRSs are less stable, with the standard deviation varying from 8% (AWS) to 22.6% (Face++), but for open-source FRSs, this is less than 3%.

We also perform a deeper analysis of accuracies for males (Fig. 3b) and females (Fig. 3c). From Fig. 3b, for gender prediction on male images, we see that all FRSs perform exceptionally well on the ORIG set, with Deepface being the best, correctly predicting the male gender for 99.91% of the images. On the other hand, the lowest accuracy is reported by Libfaceid at 98.31%. All FRSs except Deepface report the lowest accuracy for $\text{RGB}_{0.5}$, but are generally robust to Spread and Greyscale variants. For the masked variant, the lowest accuracy is reported by Microsoft Azure.

Next, looking at Fig. 3c, we note that the trends are less systematic for female faces. Even on the ORIG set, the accuracies range from 90.12% (AWS Rekognition) to 23.83% (Libfaceid). In fact, the open-source FRSs consistently report poor accuracies (independent of the image type) with the highest being 59% by Deepface on GREY and lowest being 5.7% for Libfaceid on MASK. The standard deviation trends reverse here, with commercial FRSs reporting a more stable performance – 4.14% (Face++) to 18.18% (Azure) as opposed to open-source FRSs which report a standard deviation of 17.52% (Deepface) and 20.72% (Libfaceid).

Disparity in accuracy is higher for images from GS: Table 2 shows the average accuracy (micro avg. for all images in a given region) of gender prediction for male and female cricketers for each region in the original (ORIG) and $\text{RGB}_{0.3}$ sets for all the five FRSs (more results for a balanced set of males and females are in an extended version of the paper[6]). In the rest of this section, we choose the RGB filter to demonstrate our results since it has the highest adversarial effect on the model performances (see Figure 3). For the ORIG set, all five FRSs report high accuracy (min. 97.02%) of gender prediction for males in both Global North (GN) and Global South (GS) countries, with three FRSs performing marginally better for males in GN than in GS. In contrast, the accuracy of gender prediction for images of females from Global South is significantly lower than those from Global North for all FRSs consistently. This observation corroborates the observations made in prior literature (Buolamwini and Gebru 2018; Raji et al. 2020; Jaiswal et al. 2022) regarding less accurate performance of FRSs on darker female subgroups. Irrespective of the FRS, the gender prediction accuracy for females is always lower than males within the same region. The disparity observed for commercial FRSs ranges from nearly 3% in MSFT to nearly 10% in Face++ for the images of crick-

[6]https://arxiv.org/abs/2407.15810



(a) Overall accuracy for all images (b) Overall accuracy for males (c) Overall accuracy for females

Figure 3: Overall accuracy for all FRSs, segregated by image type, for all images (a) and for each gender group (b,c). On average, AWS and Deepface are the best performing commercial and open-source FRSs respectively, independent of gender. All the FRSs are least robust to $RGB_{0.5}$ for both genders and MASK for females. The FRSs are AWS Rekognition (AWS), Microsoft Azure Face (MSFT), Face++ (FPP), Libfaceid (LIBFC), Deepface (DPFC).

Type	Region	AWS		MSFT		FPP		LIBFC		DPFC	
		M	F	M	F	M	F	M	F	M	F
ORIG	GN	98.95	95.37	99.24	96.26	99.33	89.68	97.02	30.25	99.92	75.98
	GS	98.81	**82.11**	98.93	**88.35**	99.76	**61.25**	98.53	**15.18**	99.91	**26.29**
$RGB_{0.3}$	GN	93.57	88.61	94.12	91.64	59.77	85.05	84.95	65.48	99.92	47.51
	GS	96.21	**72.36**	92.84	**73.44**	77.7	**69.11**	91.28	**42.82**	100	**12.74**

Table 2: Average male and female gender prediction accuracies for the images belonging to the ORIG and the $RGB_{0.3}$ set from Global North and Global South. M & F refer to the male and female groups, the FRSs are AWS Rekognition (AWS), Microsoft Azure Face (MSFT), Face++ (FPP), Libfaceid (LIBFC), Deepface (DPFC). Accuracy for male images are comparable between the two regions, but that for female images are lower than males consistently across regions and FRSs. FRSs report significantly lower accuracy for female images from Global South than for male images within the same region. This accuracy is also lower than accuracy observed for female images from Global North.

eters from Global North. However, this disparity increases to 10% in MSFT and 38.51% in Face++, respectively, for images from Global South. The disparity in accuracy across gender is even more pronounced in the open-source FRSs. For images from Global North, the disparity ranges from 23.94% (for Deepface) to 66.67% (for Libfaceid). However, for images from the Global South countries, this disparity goes upwards of 70% for both the open-source FRSs.

Looking at the $RGB_{0.3}$ set, we see that there is a drop in accuracy for all FRSs on images from the Global North, but a slight improvement in accuracy for some FRSs on Global South images. Interestingly, for the Global North images, the minimum and maximum disparity between the males and females reduce to 2.5% and 52.41%, indicating that the accuracy for males reduces more than that for females. For the Global South images, the max disparity now becomes 87.26%. The disparity in accuracy between females of the two regions reduces for two of the FRSs.

Takeaways: *(a) Gender prediction accuracy for females is worse than that for males for all the FRSs. This phenomenon is even more pronounced in case of open source FRSs. (b) Gender prediction accuracy for females is worse for images from Global South than for those from Global North across all FRSs.*

Grad-CAM Analysis of Predictions (RQ2)

Figure 4 shows the Grad-CAM activation maps for the Deepface predictions using 12 representative images from the FARFACE dataset. We observe that the FRS model's attention heatmap when classifying images as male is very different compared to when it classifies them as female. For the male classifications (the first two rows in Figure 4), the model's focus is on a narrow vertical region stretching from the forehead to the mouth, with the most important region being the nose, as seen from the color grading on the activation map. On the other hand, when the model predicts an image as female, the region of interest varies and random pixels are highlighted in the image. The patterns are independent of the region to which the individual belongs (GN or GS). We present some representative images here, but the observations can be generalized over all images in the dataset (last column in Fig. 4).

Takeaways: *There is a significant difference in the activation maps of the Deepface model's classification for the two genders. For the male classifications, it has a more systematic region of interest, whereas the region of interest is random for images classified as female.*

Mitigating Biases in FRSs (RQ3)

Previously, we presented the audit study for FRSs on the FARFACE dataset. We observed that all FRSs report disparate performances for females, specifically from the Global South. We now study the results for the two bias mitigation strategies (discussed earlier) for the Deepface open-source FRS – (1) **few-shot learning**, and (2) **contrastive learning**.

Results for few-shot learning: In Table 3, we report the results for the few-shot learning setup (more results are in an extended version of the paper[6]). These are compared against the baseline values from the pre-trained zero-shot setup on

Type	Region	Pre-trained		One-shot						Two-shot					
				0% Adv		50% Adv		100% Adv		0% Adv		50% Adv		100% Adv	
		M	F	M	F	M	F	M	F	M	F	M	F	M	F
ORIG	GN	**100**	70	86.04	**100**	88.96	**100**	88.96	**100**	93.33	**99.58**	91.67	**99.58**	88.75	**100**
	GS	**100**	30	91.25	88.75	94.38	86.25	94.58	86.25	96.67	80.42	95.21	82.08	92.71	86.67
$RGB_{0.3}$	GN	**100**	51.25	91.88	91.67	85.21	**97.5**	83.96	**98.33**	94.58	91.67	89.58	**97.5**	87.5	**95.42**
	GS	**100**	16.25	**95.42**	65.83	91.46	77.92	87.29	79.17	**98.54**	59.17	96.25	72.5	91.04	75.42

Table 3: Average male and female gender prediction accuracies for the images belonging to the ORIG and the $RGB_{0.3}$ set from Global North (GN) and Global South (GS) on a test set of 480 images for few-shot fine-tuning on Deepface. One-shot and two-shot refer to the number of fine-tuning examples chosen from the dataset for each gender from each country. 0%, 50% and 100% refer to the number of $RGB_{0.3}$ adversarial examples used during fine-tuning, the rest being from the ORIG set. Accuracies for females improve significantly, irrespective of the region. One-shot performs better on the ORIG set and two-shot on the $RGB_{0.3}$ set. Max values for each set and mitigation setup are in bold.

Figure 4: Example activation maps from the Grad-CAM analysis of the ORIG set for Deepface. The images are ordered as – Row 1: *Correctly predicted* males (New Zealand, Pakistan, Bangladesh), Row 2: *Misclassified* females (New Zealand, West Indies, South Africa), Row 3: *Correctly predicted* females (England, West Indies, Australia) and, Row 4: *Misclassified* males (England, India, India). For images classified as male, the region of interest is systematic, while for images classified as female, it appears random. The last column in each row shows the average Grad-CAM activation maps, indicating that the analysis generalizes.

the held out test set of 480 images. In the pre-trained model, the accuracy of females is inferior to males, irrespective of the region on the ORIG as well as the $RGB_{0.3}$ set. When the FRS is fine-tuned with one-shot examples, the accuracy for females improves significantly – by 30% for Global North (giving *100% accuracy*) and by more than 55% for Global South in the ORIG set. This improvement is similar for the $RGB_{0.3}$ test set ($> 40\%$). Finally, we see that the accuracy for females on the RGB set improves with an increase in adversarial examples, whereas the opposite is true for the ORIG set. We see similar trends for the two-shot scenario, wherein the absolute values are lower than the one-shot scenario. Here, the accuracy on both sets improves with an increasing number of $RGB_{0.3}$ fine-tuning examples.

Next, observing the accuracy for predicting the male gender for the one-shot learning scenario, we see that the accuracy decreases from 100% in the pre-trained model by as much as 14% (ORIG) – 16% ($RGB_{0.3}$). Increasing the ratio of adversarial examples in the fine-tuning set reduces the accuracy for male prediction, the opposite of what we observe for females. Next, in the two-shot scenario, the accuracy of predicting males increases again – maximum of 98.54% for males from the Global South in the $RGB_{0.3}$ set.

Finally, we note two consistent trends – (i) the accuracy for gender prediction of females from the Global South is always lower than that of the females from the Global North and, (ii) the accuracy on females (independent of the fine-tuning setting) is always higher than the males except in the $RGB_{0.3}$ set, when fine-tuned using ORIG examples only. This shows that only few examples are needed to significantly reduce the bias and improve the accuracy of the marginalized class.

Results for generalizability to other datasets: We benchmark all the five pre-trained FRSs and the Deepface model fine-tuned on 2-shot images from the FARFace dataset on five highly popular diverse face datasets viz. CelebSET (Raji et al. 2020), CFD-USA (Ma, Correll, and Wittenbrink 2015), CFD-India (Lakshmi et al. 2021), CFD-MR (Ma, Kantner, and Wittenbrink 2020) and Fairface (Karkkainen and Joo 2021). From Table 4 we see that while the three commercial systems report significantly high accuracies on all datasets, Libfaceid and Deepface lag far behind. Thus, these open-source models, despite being trained on faces from the Global North do not generalize well to even these datasets. Our fine-tuned model (DPFC-FT), on the other hand, reports a significantly high accuracy for other datasets as well exhibiting good generalizability prowess. We also observe a reduction in disparity in gender accuracy for all datasets for our fine-tuned model as compared to the pre-trained Deepface model (results are in an extended version of the paper[6]). For all datasets except CFD-MR, the disparity reduc-

Model	Dataset				
	CelebSET	**CFD-USA**	**CFD-India**	**CFD-MR**	**Fairface**
AWS	99.5%	97.32%	98.59%	97.73%	92.11%
FPP	99.06%	93.47%	92.96%	90.91%	90.83%
MSFT	99.31%	99.83%	100%	100%	80.01%
LIBFC	85.94%	72.70%	79.58%	76.14%	76.67%
DPFC	82.06%	79.56%	73.94%	67.05%	71.67%
DPFC-FT	97.75%	85.26%	93.66%	89.77%	84.62%

Table 4: Benchmark on other datasets for all FRSs and for 2-shot fine-tuned Deepface using the FARFace dataset (DPFC-FT). The Deepface model reports high accuracies on all datasets after fine-tuning proving generalizability of our technique.

Figure 5: Example activation maps from the Grad-CAM analysis of the ORIG set (held-out test set) for Deepface after two-shot fine-tuning on the ORIG set. The images are ordered as – Row 1: *Correctly predicted* males (Australia, New Zealand, Pakistan, Bangladesh), Row 2: *Misclassified* females (West Indies, Bangladesh, Pakistan, India), Row 3: *Correctly predicted* females (England, West Indies, South Africa, India) and, Row 4: *Misclassified* males (Australia, West Indies, India, England). Row 5 has the average Grad-CAM activation maps for the images of correctly predicted males, correctly predicted females, misclassified females, and misclassified males; it is apparent that there is a more systematic focus on the nose for correctly predicted females.

tion ranges from 29-60%. Thus, we can state that *models fine-tuned on images from the Global South can be re-used in other regions without loss in performance.*

Grad-CAM analysis: We evaluate the Grad-CAM activation maps of the held-out test set for the following setup: two-shot fine-tuning with all training examples chosen from the ORIG set. Here we observe a change in the highlighted area of the image when the model predicts the female gender – the model's region of interest is around the *nose* of the person, as opposed to random areas within the image. This indicates a clear change in the model's decision making process, specially for the true positive data points. Example activation maps are shown in Figure 5. Consider activation map in row 2, col 1 of Fig. 4 and in row 3, col 2 of Fig. 5– both belong to the same female and there is a clear change in the area of interest. This analysis generalizes over the entire held-out test set as is evident from the last row of Fig. 5.

Type	Region	Pre-trained		Contrastive	
		M	F	M	F
ORIG	GN	**100**	70	82.5	88.75
	GS	**100**	30	**90**	80
$RGB_{0.3}$	GN	**100**	51.25	81.88	**88.75**
	GS	**100**	16.25	87.5	76.25

Table 5: Average male and female gender prediction accuracies for the contrastive learning setup on test sets of 480 images. Female accuracies from both regions improve significantly. Accuracy for females is higher than males in Global North and lower in Global South, for both test sets.

Results for contrastive learning: Finally, we also experiment with a *novel* contrastive learning setup. Recall, for the computation of the triplet loss for an anchor point (any data point in a given class) (x_a), the positive example (x_a^+) is an adversarial image of the same individual from $RGB_{0.3}$, whereas the negative example (x_a^-) is an image of another person from the ORIG set of the opposite gender (more results for other variants are in an extended version of the paper[6]). Table 5 shows the accuracy in the ORIG set for gender prediction increases significantly for females from both Global North (18.75%) and Global South (50%). The increase is more significant for images in the $RGB_{0.3}$ set with an increment of 37.5% in the Global North and 60% in the Global South, thereby justifying our choice of adversarial samples in the learning setup. In compensation, the accuracy for males drops, with a max drop of $\approx$ 20%. We do not observe a large reduction in bias between the different groups here.

Fairness vs. accuracy: As this is a binary classification task,

Scheme	Region	Type					
		ORIG	$RGB_{0.3}$	$RGB_{0.5}$	GREY	SPRD	MASK
Fine-tuning	GN	54.17	60.83	64.17	53.33	52.50	62.50
	GS	75	53.33	41.67	76.11	49.44	57.22
Contrastive	GN	87.92	83.33	86.25	86.25	81.25	72.38
	GS	90.56	89.44	86.67	80.56	92.22	60.42

Table 6: Average country prediction accuracies for the two regions, for the fine-tuning and contrastive learning setup. In the fine-tuning setup, first, the model is fine-tuned on Geofaces, followed by a 2-shot fine-tuning on FARFace. In the second setup, the model is trained using contrastive learning first on the Geofaces dataset and followed by a second round of contrastive learning on the FARFace dataset. The two rounds of contrastive learning produce the best results.

we hypothesize that the drop in accuracy for males and significant improvement in accuracy for females is a result of an angular shift in the separating hyperplane, causing a small number of previously classified males to now be classified as female, but a larger number of females to be classified correctly. Thus we see that the cost of fairness (reduction in disparity between the performance for the two genders) is a drop in accuracy for males, the majority class, which is typically a standard observation in the literature. A more in-depth mathematical study to bound this shift in accuracy, while maintaining fairness is deferred to a future study.

Takeaways: *(a) The model is highly receptive to few-shot fine-tuning setups showing a significant improvement in accuracy, especially for females in the Global South,* and, *(b) Adversarial examples improve the accuracy as well as the robustness of the fine-tuning setup.*

Results of Red-Teaming Task (Country Prediction)

The results for predicting the country from face image are shown in Table 6. Here, we see that depending on the adversarial variant, one region reports a higher accuracy over the other. Overall, fine-tuning results are poor with accuracy values around 50% only. The 2-stage contrastive learning setup produces the best and least disparate results. However, as a note of caution we would like to state that the premise of predicting one's country or ethnicity from their face image is a flawed one and has primarily been used for surveillance (Mozur 2019). As an example, one can consider a country like India with its large ethnicity and individuals being profiled as belonging to other countries because of their facial phenotype, leading to unfair and biased outcomes.

Concluding Discussion

Benchmark audit of FRSs: In our benchmark audit, we see that all models have a good accuracy on the ORIG set, but commercial models are less robust to adversarial inputs; all FRSs report disparate results against Global South females (generally dark skin tone). Such biases were observed previously on commercial FRSs (Buolamwini and Gebru 2018; Raji et al. 2020; Jaiswal et al. 2022). Thus we observe both *temporal* and *emergent* biases against dark-skinned women (Mehrabi et al. 2021) through our study. Our current study supports existing observations and necessitates the ***need for continual temporal audits*** of FRSs, especially for adversarial real-world inputs.

Explanation of FRS outputs: Grad-CAM analysis on the Deepface FRS reinforced the audit observations that FRS models are better equipped to predict gender for males (the IMDB-Wiki (Rothe, Timofte, and Van Gool 2018) dataset has majority male faces (Raji et al. 2020)); these observations generalize across countries. This leads us to the final part of our pipeline – bias mitigation.

Bias mitigation on Deepface: Our few-shot and contrastive learning approaches result in improved accuracy for females in both regions. Accuracy for the $RGB_{0.3}$ set also improves, resulting in increased model robustness. Few-shot fine-tuning performs better than contrastive learning overall, but we defer further exploration of the reasoning to a future study. Simple techniques like fine-tuning and contrastive learning improve accuracy while reducing bias, and adversarial examples improve robustness.

Limitations and scope for future work: While the current dataset has images from both Global South and Global North, the age related information and gender diversity (as the sport of cricket is male dominated) is missing. We would like to improve the dataset on these dimensions in the future. The current study focuses on classification tasks. We intend to extend it to retrieval tasks like face verification. We also plan to study the accuracy vs fairness trade-offs for our mitigation strategies and introduce better algorithms that reconciles between the two notions better.

Acknowledgements

This research is supported partly by a European Research Council (ERC) Advanced Grant for the project "Foundations for Fair Social Computing", funded under the European Union's Horizon 2020 Framework Programme (grant agreement no. 789373), and by a grant from the Max Planck Society through a Max Planck Partner Group at IIT Kharagpur. S. Jaiswal is supported by the Prime Minister's Research Fellowship, Govt. of India.

Ethics Statement

In this work, we have introduced a new large-scale face dataset, curated from images of cricket players. We would like to state here that we use the gender labels already publicly available on the `ESPNCricInfo` website and do not attempt to modify these or any other ground truth attributes. We do not collect any private information. Our study is limited to binary gender classification, but this is primarily due to only binary gender labels – male & female being available in all face datasets (ours and in literature). We do not intend to propagate the perspective that gender is a binary label; we acknowledge that gender is a fluid spectrum, and hence all gender labels in our study are to be interpreted as *perceived* gender.

Next, we note that our dataset is unbalanced, with a larger ratio of males than females. This is no surprise as it is reflective of gender imbalance observed in many professions, including sports (Singh 2023) and is a societal issue that needs deeper reflection and redressal through governmental intervention. Multiple prior open-source models have also been trained on such imbalanced datasets as is evident from their bias against females in audit studies. We do not condone this and, as such, to prevent further misuse, will only share a balanced subset of our dataset for any and all research purposes. We also audit on a balanced subset of images and identify that similar biases as shown here persist (Results are in an extended version of the paper[6]). Even in our bias mitigation experiments, we have only used balanced subsets (for both fine-tuning and contrastive learning).

Images from the Global South can include additional sensitive features, for example, caste or religion in South Asia, which are correlated to one's socio-economic position in these societies. These attributes intersect with sensitive features like skin tone and can lead to newer skews beyond Global North and South. We acknowledge this as a limitation of our dataset and recommend future users of the dataset to be mindful when performing sociotechnical analyses.

Finally, our red teaming experiment has been done to show the relative ease of misusing FRSs for controversial and morally ambiguous experiments like predicting the country of a person from their face. Ironically, the results from this experiment too are biased, thus showing how models trained on datasets from the Global North encode strong biases. We will not be releasing either the code or the specific subset of images used for this experiment.

Our primary motivation for this work stems from the under-representation of the Global South in the larger AI development landscape and over-representation in the AI testbed landscape. We hope our dataset and bias mitigation algorithms will break this hegemony and provide some balance.

References

Aggarwal, P.; Chawla, P.; Das, M.; Saha, P.; Mathew, B.; Zesch, T.; and Mukherjee, A. 2023. HateProof: Are Hateful Meme Detection Systems Really Robust? In *WWW '23*.

AlBdairi, A. J. A.; Xiao, Z.; and Alghaili, M. 2020. Identifying ethnics of people through face recognition: A deep CNN approach. *Scientific Programming*, 2020: 1–7.

Ali, M.; Sapiezynski, P.; Bogen, M.; Korolova, A.; Mislove, A.; and Rieke, A. 2019. Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes. *Proceedings of the ACM on human-computer interaction*, 3(CSCW): 1–30.

Amazon. 2023. Amazon AWS Rekognition. https://aws.amazon.com/rekognition/faqs/. Accessed: 2023-04-01.

Anwar, A.; and Raychowdhury, A. 2020. Masked Face Recognition for Secure Authentication. arXiv:2008.11104.

Bansal, V. 2022. Uber's facial recognition is locking Indian drivers out of their accounts. https://www.technologyreview.com/2022/12/06/1064287/ubers-facial-recognition-is-locking-indian-drivers-out-of-their-accounts/. Accessed: 2023-04-01.

Batsukh, B.-E.; and Tsend, G. 2016. Effective computer model for recognizing nationality from frontal image. *arXiv preprint arXiv:1603.04550*.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91. PMLR.

Chen, T.; Han, W.-L.; Wang, H.-D.; Zhou, Y.-X.; Xu, B.; and Zang, B.-Y. 2007. Content Recommendation System Based on Private Dynamic User Profile. In *2007 International Conference on Machine Learning and Cybernetics*.

Conti, J.-R.; Noiry, N.; Clemencon, S.; Despiegel, V.; and Gentric, S. 2022. Mitigating gender bias in face recognition using the von mises-fisher mixture model. In *International Conference on Machine Learning*, 4344–4369. PMLR.

Dai, Y.; Ye, H.; and Gong, S. 2009. Personalized recommendation algorithm using user demography information. In *2009 Second International Workshop on Knowledge Discovery and Data Mining*, 100–103. IEEE.

Dizikes, P. 2023. How an "AI-tocracy" emerges. https://news.mit.edu/2023/how-ai-tocracy-emerges-0713. Accessed: 2023-08-08.

Dooley, S.; Wei, G. Z.; Goldstein, T.; and Dickerson, J. 2022. Robustness Disparities in Face Detection. *Advances in Neural Information Processing Systems*, 35: 38245–38259.

Eidinger, E.; Enbar, R.; and Hassner, T. 2014. Age and gender estimation of unfiltered faces. *IEEE Transactions on information forensics and security*, 9(12): 2170–2179.

Face++. 2023. Face++ Detect. https://www.faceplusplus.com/face-detection/. Accessed: 2023-04-01.

Fitzpatrick, A. 2023. Facial recognition's alarming pitfalls. https://www.axios.com/2023/01/07/facial-recognition-issues-problems. Accessed: 2023-05-05.

for the Future, F. 2023. Ban Facial Recognition. https://www.banfacialrecognition.com/. Accessed: 2024-06-01.

Gong, S.; Liu, X.; and Jain, A. K. 2020. Jointly de-biasing face recognition and demographic attribute estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, 330–347. Springer.

Google. 2023. Search. https://www.google.com. Accessed: 2023-06-01.

Jaiswal, S.; Duggirala, K.; Dash, A.; and Mukherjee, A. 2022. Two-face: Adversarial audit of commercial face recognition systems. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 381–392.

Jocher, G.; Chaurasia, A.; Stoken, A.; Borovec, J.; NanoCode012; Kwon, Y.; Michael, K.; TaoXie; Fang, J.; imyhxy; Lorna; Yifu, Z.; Wong, C.; V, A.; Montes, D.; Wang, Z.; Fati, C.; Nadar, J.; Laughing; UnglvKitDe; Sonck, V.; tkianai; yxNONG; Skalski, P.; Hogan, A.; Nair, D.; Strobel, M.; and Jain, M. 2022. ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation.

Kamgar-Parsi, B.; Lawson, W.; and Kamgar-Parsi, B. 2011. Toward development of a face recognition system for watchlist surveillance. *IEEE TPAMI*.

Karkkainen, K.; and Joo, J. 2021. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *IEEE/CVF WACV*.

Keyes, O. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction*, 2(CSCW): 1–22.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Krueckeberg, J.; and Ferris, G. 2018. Face off: The lawless growth of facial recognition in UK policing. https://bigbrotherwatch.org.uk/wp-content/uploads/2018/05/Face-Off-final-digital-1.pdf. Accessed: 2024-01-01.

Lakshmi, A.; Wittenbrink, B.; Correll, J.; and Ma, D. S. 2021. The India Face Set: International and Cultural Boundaries Impact Face Impressions and Perceptions of Category Membership. *Frontiers in Psychology*.

Ma, D. S.; Correll, J.; and Wittenbrink, B. 2015. The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*.

Ma, D. S.; Kantner, J.; and Wittenbrink, B. 2020. Chicago Face Database: Multiracial expansion. *Behavior Research Methods*.

Majumdar, P.; Mittal, S.; Singh, R.; and Vatsa, M. 2021. Unravelling the effect of image distortions for biased prediction of pre-trained face recognition models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3786–3795.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Microsoft. 2023. Microsoft Azure Face. https://azure.microsoft.com/en-in/services/cognitive-services/face/. Accessed: 2023-04-01.

Mozur, P. 2019. One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority. https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html. Accessed: 2024-01-01.

NIST. 2020. Facial Recognition Technology (FRT). https://www.nist.gov/speech-testimony/facial-recognition-technology-frt-0. Accessed: 2024-01-01.

Olier, J. S.; and Spadavecchia, C. 2022. Stereotypes, disproportions, and power asymmetries in the visual portrayal of migrants in ten countries: an interdisciplinary AI-based approach. *Humanities and Social Sciences Communications*.

OpenAI. 2023. ChatGPT. https://chat.openai.com. Accessed: 2023-06-01.

Parkhi, O. M.; Vedaldi, A.; and Zisserman, A. 2015. Deep Face Recognition. In *British Machine Vision Conference*.

Parliament, E. 2023. EU AI Act: first regulation on artificial intelligence. https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence. Accessed: 2024-01-01.

Raji, I. D.; Gebru, T.; Mitchell, M.; Buolamwini, J.; Lee, J.; and Denton, E. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 145–151.

Rao, Y.; Lu, J.; and Zhou, J. 2019. Learning discriminative aggregation network for video-based face recognition and person re-identification. *IJCV*.

Rothe, R.; Timofte, R.; and Van Gool, L. 2018. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4): 144–157.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Serengil, S. I.; and Ozpinar, A. 2021. HyperExtended LightFace: A Facial Attribute Analysis Framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, 1–4. IEEE.

Singh, P. 2023. Gender parity in sports will need focused planning & training. https://timesofindia.indiatimes.com/blogs/voices/gender-parity-in-sports-will-need-focused-planning-training/. Accessed: 2023-09-09.

Sur, A. 2023. The startup behind DigiYatra reveals how it perfected the tool's facial recognition algorithm. https://www.moneycontrol.com/news/business/startup/the-startup-behind-digiyatra-reveals-how-it-perfected-the-tools-facial-recognition-algorithm-10016211.html. Accessed: 2024-01-01.

Umagat, R. 2023. Libfaceid. https://github.com/richmondu/libfaceid. Accessed: 2023-04-01.

UN. 2022. UNCTAD Handbook of Statistics 2022. https://unctad.org/system/files/official-document/tdstat47_en.pdf. Accessed: 2023-04-01.

Wang, M.; and Deng, W. 2020. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9322–9331.

Wang, M.; Zhang, Y.; and Deng, W. 2021. Meta balanced network for fair face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(11): 8433–8448.

Wang, Z.; Dong, X.; Xue, H.; Zhang, Z.; Chiu, W.; Wei, T.; and Ren, K. 2022. Fairness-aware adversarial perturbation towards bias mitigation for deployed deep models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10379–10388.

Yang, S.; Luo, P.; Loy, C.-C.; and Tang, X. 2016. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5525–5533.

Zhang, Z.; Song, Y.; and Qi, H. 2017. Age Progression/Regression by Conditional Adversarial Autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.