# Anatomizing Bias in Facial Analysis

## Richa Singh,[1] Puspita Majumdar, [1,2] Surbhi Mittal, [1] Mayank Vatsa [1]

[1] IIT Jodhpur
[2] IIIT-Delhi
richa@iitj.ac.in, pushpitam@iiitd.ac.in, mittal.5@iitj.ac.in, mvatsa@iitj.ac.in

## Abstract

Existing facial analysis systems have been shown to yield *biased* results against certain demographic subgroups. Due to its impact on society, it has become imperative to ensure that these systems do not discriminate based on gender, identity, or skin tone of individuals. This has led to research in the identification and mitigation of bias in AI systems. In this paper, we encapsulate bias detection/estimation and mitigation algorithms for facial analysis. Our main contributions include a systematic review of algorithms proposed for understanding bias, along with a taxonomy and extensive overview of existing bias mitigation algorithms. We also discuss open challenges in the field of biased facial analysis.

## Introduction

Artificial Intelligence (AI) systems are used in complex decision-making tasks, including risk assessment, criminal sentencing, and healthcare diagnostics. These decisions affect every aspect of our life, especially when AI systems are used for making predictions about individuals. However, as shown in Figure 1, several of these systems are found to be biased against certain demographic groups which raises towards trustability and fairness of these systems (Osoba and Welser IV 2017). Multiple instances of biased predictions are observed in AI applications, such as facial analysis tasks, including face detection, attribute prediction, and face recognition. (Buolamwini and Gebru 2018) have shown the biased behavior of commercial gender classifiers against darker-skinned females. Other instances, such as false identification of members of US Congress as criminals (Paolini-Subramanya 2018) and misclassification of an African American couple as *gorilla* (Team 2015), highlight the biased behavior of existing systems. Due to several of these instances, some corporate and government organizations have banned the use of facial analysis systems (Conger, Fausset, and Kovaleski 2019).

These instances have motivated the research community towards designing mechanisms to improve the fairness and trustability of these systems. The focus of researchers is towards understanding bias in model prediction and mitigating its effect to obtain unbiased outcomes (Ntoutsi et al.
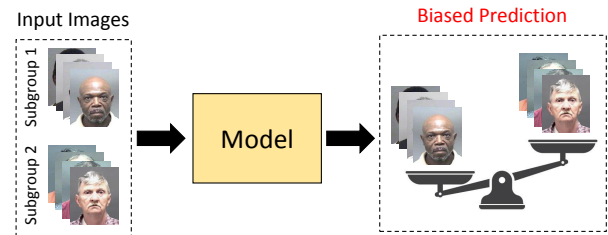
Figure 1: Illustrating biased predictions of deep models to favor/disfavor certain demographic subgroups.

2020). In this regard, multiple analyses have been performed to detect the sources of bias (Celis and Rao 2019; Krishnapriya et al. 2020), databases are proposed to understand the effect of bias (Wang et al. 2019a; Karkkainen and Joo 2021), and algorithms are developed to mitigate the effect of bias in model predictions (Majumdar et al. 2020; Joo and Kärkkäinen 2020).

This research presents a systematic survey of bias in facial analysis tasks as opposed to the broader topic of biometrics addressed in (Drozdowski et al. 2020). We provide a synopsis of the analysis towards understanding bias, discuss the bias mitigation algorithms proposed in the literature, and summarize the details of the databases developed for studying bias. As a part of this research, we also present a taxonomy to systematically categorize the techniques proposed for mitigating bias in facial analysis tasks. In the end, with a help of meta-analysis, we discuss open challenges that require attention of the research community.

## Understanding Bias in Facial Analysis

The prevalence of bias has adverse effects on modern technology, and various attempts have been made to understand and detect the presence of bias. Since bias can occur in a system from various sources (Figure 2), different research efforts involve understanding the effect of bias from different perspectives. A large body of work is dedicated towards analyzing biased predictions of models for protected attributes such as gender and ethnic subgroups. In the following subsections, we discuss the research towards understanding bias in facial analysis.
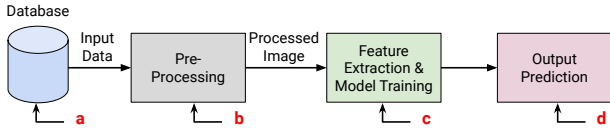
Figure 2: Sources of bias in a facial analytics system pipeline. (a) Dataset bias, (b) bias in the pre-processing step, (c) bias in feature extraction and model training, and (d) bias in prediction.

## Bias in Face Detection and Recognition

The performance of face recognition systems has been observed to be inconsistent across different demographic groups. An early observation in this regard is made by (Klare et al. 2012) where they demonstrated the difference in recognition performance for different commercial and non-trainable algorithms. They observed a consistently low performance for darker-skinned individuals and prompted the usage of either balanced datasets for training algorithms or separate algorithms for different subgroups. These observations are made in the pre-deep learning era for face recognition algorithms using LBP and Gabor filters. With the onset of the deep learning era, it is observed that demographic bias is an ongoing problem. Research has shown that females tend to have a higher false match rate and a higher false non-match rate as compared to males in face verification applications. To study this phenomenon, the score distributions obtained for genuine as well as impostor pairs across different demographic subgroups have been analyzed. In (Albiero et al. 2020), the authors showcase how females with impostor distribution have higher similarity scores while females with genuine distribution have lower similarity scores. In (Robinson et al. 2020), the authors analyzed the decision threshold for the face verification task and observed different thresholds to be optimal for different demographic subgroups. They highlighted how learning a global threshold for matching leads to the incorporation of bias in the system. Similarly, for race, (Vangara et al. 2019) used the MORPH dataset and observed that the genuine and impostor distributions are significantly different across Caucasian and African-American subgroups. In another work focusing on bias in face recognition, (Krishnapriya et al. 2020) made similar observations for face verification decision thresholds. They further explore optimal decision thresholds for one-to-many identification search.

The presence of gender and ethnic subgroup information in face recognition technology clearly indicates that deep models embed the aforementioned information and utilize it for predictions. In this spirit, (Acien et al. 2018) attempted to infer gender and ethnic group information from pre-trained deep models. They observed that these models classify gender and ethnicity with nearly 95% accuracy on the LFW database. To further understand how deep models incorporate demographic information, (Serna et al. 2019) used feature space visualizations along with class activation maps (CAMs) which forms a popular technique for inspecting relevant pixels in the input image. They further comment upon how over-representation of certain demographic groups in

popular face databases (dataset bias Figure 2(a)) has led to popular pre-trained deep face models being biased. Many face recognition systems have a mandatory face quality assessment step while enrolment of an individual. The assessment step ensures the face image meets a certain quality threshold thereby providing a high-quality image for comparison at query time. (Terhörst et al. 2020c) study the correlation between face quality estimation and demographic bias in face recognition (bias in pre-processing step Figure 2(b)). On the evaluation of four algorithms for face quality assessment towards biases to pose, ethnicity, and age, they observed bias towards frontal poses against Asian and African-American ethnicities and towards face images of individuals below 7 years.

Most face recognition databases collected in the wild lack annotation information for protected attributes such as race and gender. This leads to incomplete information about a model's ability to generalize across different subgroups (dataset bias Figure 2(a)). In (Kortylewski et al. 2019), the authors used synthetically generated images for their study and observed significant influence of pose variation on the generalization performance. The authors leveraged synthetic data for analysis and showcased how facial pose and facial identity cannot be completely disentangled by deep networks (bias in model training Figure 2(c)). To further study the impact of dataset bias, (Gwilliam et al. 2021) analyzed facial recognition performance by training on various imbalanced distributions across race. They observed less biased model predictions after training on a specific subgroup than training on a balanced distribution. Further, the addition of more samples for existing identities in the database improved performance across racial subgroups.

(Celis and Rao 2019) analyzed the latent representations of faces to understand potential sources of bias. They observed that the images became brighter with increasing latent values and darker as values got lower, highlighting the importance of skin color in latent representations. To understand where bias is encoded in face recognition, certain works used CAMs and showed how activated regions vary across different demographic subgroups. In such an attempt to understand the cause of bias in deep models, (Nagpal et al. 2019) observed the presence of own-race and own-age bias in popular deep networks. They observed that deep models have a tendency to focus on selected facial regions for a particular ethnic subgroup, with these regions varying across different subgroups. Similarly, in (Majumdar et al. 2021), the authors study the incorporation of bias in model predictions in the presence of real-world distortions (Figure 2(d)).

## Bias in Attribute Prediction

It has been observed in the literature that many deep learning-based systems are biased in their predictions when measured across different subgroups. (Buolamwini and Gebru 2018) evaluated the performance of three commercial gender-classification systems across faces with different skin tones. They observed a huge disparity in classification error rates for darker females versus lighter males. (Deuschel, Finzel, and Rieger 2020) studied the impact of gender and skin tone on facial expression detection and

used classification accuracy and heatmaps for quantitative and qualitative evaluations. (Krishnan, Almadan, and Rattani 2020) investigated the impact of different deep models and training set imbalance on gender classification across different gender-race groups. They highlighted how training set imbalance widens the gap in performance accuracy. The authors further studied facial morphology for different ethnic subgroups using facial landmark detection and obtained interesting insights about probable causes of disparity. To further investigate the impact of different skin tones on gender classification, (Muthukumar 2019) used luminance mode-shift and optimal transport techniques to vary the skin tones. They reported that skin tone alone is not the driving factor for observed bias, and broader differences in ethnicity must be considered. (Joo and Kärkkäinen 2020) proposed another approach for understanding bias involves using counterfactuals where they synthesized counterfactual face images with varying gender and ethnic groups keeping the other signals constant. Using these samples, they analyzed performance on different downstream tasks and commented on different hidden biases in the system. Similarly, (Denton et al. 2019) performed sensitivity analysis based on the performance of deep models using generated counterfactuals. (Quadrianto, Sharmanska, and Thomas 2019) translated the data from the input domain to a fair target domain. They observed interesting outcomes where the model adjusts eyes and lips regions in males to enforce fairness in predictions. Further studying the interdependence between factors leading to biased model predictions, (Barlas et al 2021) analyzed the correlation between context and gender in five proprietary image tagging algorithms.

The fairness of models is generally attributed to the difference in performance across subgroups. In a different direction, (Serna et al. 2021b) proposed *InsideBias* in which they studied how the model represents the information instead of how it performs. They analyzed the features learned by the models while training using an unbalanced dataset in the context of bias. (Serna et al. 2021a) showed how analyzing model weights provides insights into its biased behavior. (Li and Xu 2021) have proposed a variation loss that optimizes the hyperplane in the latent space to obtain biased attribute information. Researchers continue to study the influence of bias in predictions (Figure 2(d)), and predominantly how it might have been incorporated at the data and algorithm level (Figures 2(a) and 2(c)). A large body of work focuses on analyzing popular deep model architectures and COTS algorithms. Popular tools include feature map visualizations, fairness and performance evaluation metrics across subgroups, generated counterfactuals, and skewed training. However, a limited number of studies have been performed to detect bias-inducing factors in an automated manner and require the attention of the research community.

# Bias Mitigation in Facial Analysis

In this section, we provide a taxonomy of the bias mitigation techniques proposed for various face analysis tasks.

## Face Detection and Recognition

Majority of the algorithms developed for mitigating bias in face detection and recognition are based on **deep learning** based approaches. These approaches include designing novel loss functions, custom networks, and discrimination-aware learning methods. (Amini et al. 2019) proposed a novel algorithm for mitigating hidden bias in face detection algorithms. The proposed algorithm uses a variational autoencoder to learn the latent structure within the database. The learned latent distributions are used to re-weight the importance of certain data points during training. In the literature, it is shown that imbalanced class distribution leads to biased predictions of deep models. To handle the problem of imbalanced class distribution, (Huang et al. 2019) proposed to learn Cluster-based Large Margin Local Embedding, and combined it with k-nearest cluster algorithm for improved recognition performance. A deep information maximization adaptation network is proposed by (Wang et al. 2019a) for bias mitigation using deep unsupervised domain adaptation techniques. The authors considered Caucasian as the source domain and other races as target domains to decrease race gap at domain-level. (Terhörst et al. 2020b) proposed a fairness driven neural network classifier that works on the comparison-level of a biometric system. A novel penalization term in the loss function is used to train the proposed classifier to reduce the intra-ethnic performance differences and introduce both group and individual fairness to the decision process. (Bruveris et al. 2020) proposed to mitigate bias in face recognition due to imbalanced data distribution by employing sampling strategies that balance the training procedure. Attention mechanisms are also used for enhancing the fairness of recognition models. In this direction (Gong, Liu, and Jain 2021) proposed group adaptive classifier that uses adaptive convolution kernels and attention mechanism for bias mitigation. The adaptive module and the attention maps help to activate different facial regions to learn discriminative features for each demographic group. (Yang et al. 2021) proposed the race adaptive margin based face recognition (RamFace) model to enhance the discriminability of the features. The authors also proposed a race adaptive margin loss function to automatically derive different optimal margins to mitigate the effect of racial bias.

Some researchers have used **adversarial learning** based approaches for bias mitigation. (Alasadi, Al Hilli, and Singh 2019) presented a framework for matching low resolution and high-resolution facial images. The aim is to mitigate bias in cross-domain face recognition. The proposed framework consists of two parts, the first part maximizes the face matching quality and the second part minimizes the prediction of demographic properties to reduce bias in model prediction. A novel de-biasing adversarial network is proposed by (Gong, Liu, and Jain 2020) that adversarially learns to generate disentangled representations for unbiased face and demographics recognition. The proposed network consists of four classifiers to distinguish the identity, gender, race, and age of the facial images. (Dhar et al. 2020) presented a novel Adversarial Gender De-biasing algorithm to reduce the gender prediction ability of face descriptors. The proposed algorithm unlearns the gender information in descrip-

tors while training them for classification.

Researchers have also proposed techniques that could be combined with **pre-trained** models to improve their performance and reduce bias in model prediction. (Serna et al. 2020) proposed a discrimination-aware learning method, termed as Sensitive loss for bias mitigation. The proposed loss function is based on the triplet loss function and a sensitive triplet generator to improve the performance of pre-trained models. A novel unsupervised fair score normalization approach based on individual fairness is proposed by (Terhörst et al. 2020d). The proposed solution is easily integrable into existing systems that reduce bias and improves the overall recognition performance.

**Generative** approaches are adopted by researchers to synthesize images of the under-represented class for bias mitigation. (McDuff et al. 2019) proposed a simulation-based approach using generative adversarial models for synthesizing facial images to mitigate gender and racial biases in commercial systems. A novel data augmentation methodology is proposed by (Yucer et al. 2020) to balance the training database at a per-subject level. The authors used image-to-image transformation for transferring facial features with sensitive racial characteristics while preserving the identity-related features for mitigating racial bias. Recently, **reinforcement learning** based approach is used by (Wang and Deng 2020) to learn balanced features and remove racial bias in face recognition using the idea of adaptive margin.

It is observed that the majority of the algorithms for mitigating bias in face recognition are designed to mitigate bias for a specific demographic group. Therefore, these algorithms may not generalize for other demographic groups (Xu et al. 2021). It is our belief that researchers should focus more on designing solutions that are generalizable across different demographic groups.

### Attribute Prediction

Several algorithms have been proposed for bias mitigation in attribute prediction. Among **deep learning** approaches, one set of algorithms aim to unlearn the model's dependency on sensitive attributes. (Kim et al. 2019) proposed a regularization loss to minimize the mutual information between feature embedding and bias to unlearn the bias information. Attribute aware filter drop is proposed by (Nagpal et al. 2020a) that performs the primary attribute classification task while unlearning the dependency of the model on sensitive attributes. (Tartaglione, Barbano, and Grangetto 2021) proposed a regularization strategy to disentangle the biased features while entangling the features belonging to the same target class. Apart from this some techniques are proposed that use feature distillation (Jung et al. 2021) and mutual information between the learned representation (Ragonesi et al. 2021). For understanding different bias mitigation algorithms (Wang et al. 2020) provided a thorough analysis of the existing bias mitigation techniques. They further designed a domain-independent training technique for bias mitigation. An interesting data augmentation strategy, *fair mixup* is proposed by (Chuang and Mroueh 2021) to optimize group fairness constraints. The authors proposed to regularize the model on interpolated distributions between

different subgroups of a demographic group. Recently, (Park et al. 2021) argued that removing the information of sensitive attributes in the decision process has the limitation of eliminating beneficial information for target tasks. To overcome this, the authors proposed Fairness-aware Disentangling Variational Auto-Encoder for disentangling data representation into three latent subspaces. A decorrelation loss is proposed to align the overall information into each subspace, instead of removing the information of sensitive attributes.

(Dwork et al. 2018) proposed a decoupling technique for bias mitigation of **black-box** ML algorithms. The proposed technique learns separate classifiers for different groups to increase fairness and accuracy in classification systems. (Nagpal et al. 2020b) proposed diversity blocks to de-bias existing models. The diversity block is trained using small training data and can be added to any black-box model. (Roh et al. 2020) addressed the problem of bias mitigation using bi-level optimization. They proposed to adaptively select mini-batches for improving model fairness. A bias mitigation algorithm based on adversarial perturbation is proposed by (Majumdar et al. 2020). The proposed algorithm learns a subgroup invariant perturbation to be added to the input database to generate a transformed database. The transformed database, when given as input to a model, produces unbiased outcomes. The proposed algorithm is able to mitigate bias in **pre-trained** model prediction without retraining. Recently, (Majumdar, Singh, and Vatsa 2021) proposed an algorithm based on attention mechanism for mitigating bias in pre-trained models.

Oversampling the minority class is one of the popular techniques for handling the class imbalance problem. Imbalance in class distribution leads to biased predictions and multiple **generative** approaches have been proposed for mitigation. (Ramaswamy, Kim, and Russakovsky 2021) used generative adversarial networks (GANs) to generate images for data augmentation. The generated images are perturbed in the latent space to generate balanced training data w.r.t protected attribute for bias mitigation. A weakly-supervised algorithm is proposed by (Choi et al. 2020) to overcome database bias for deep generative models. An additional unlabeled database is required by the proposed approach for bias detection in existing databases. (Tan, Shen, and Zhou 2020) proposed an effective method for improving the fairness of image generation for a pre-trained GAN model without retraining. Images generated using the proposed method are applied for bias quantification in commercial face classifiers. A multi-attribute framework is proposed by (Georgopoulos et al. 2021) to transfer facial patterns even for the underrepresented subgroups. The proposed method helps to mitigate dataset bias by data augmentation. Apart from the generative approaches, (Wang et al. 2019b) proposed an **adversarial** approach to remove bias from intermediate representations of a deep neural network. The proposed approach reduces gender bias amplification and maintains the overall model performance. (Adeli et al. 2021) used adversarial training to maximize the discriminative power of the learned features with respect to the main task and minimize the statistical mean dependence with the bias variable. By minimizing the dependency on the bias variable, the authors have

| Database | Identity | Gender | Race | Age | No. of Subjects | No. of Images |
|---|---|---|---|---|---|---|
| MORPH (Rawls and Ricanek 2009) | ✓ | ✓ | ✓ | ✓ | 13,000+ | 55,000+ |
| IMFDB (Setty et al. 2013) | ✓ | ✓ | ✗ | ✓ | 100 | 34,512 |
| Adience (Eidinger, Enbar, and Hassner 2014) | ✓ | ✓ | ✗ | ✓ | 2,284 | 26,580 |
| CACD (Chen, Chen, and Hsu 2015) | ✓ | ✗ | ✗ | ✓ | 2,000 | 163,446 |
| LFWA (Huang et al. 2008) | ✓ | ✓ | ✗ | ✗ | 5,749 | 13,233 |
| CelebA (Liu et al. 2015) | ✗ | ✓ | ✗ | ✗ | 10,000+ | 2,02,599 |
| AgeDb (Moschoglou et al. 2017) | ✓ | ✗ | ✗ | ✓ | 568 | 16,488 |
| AAF (Cheng et al. 2019) | ✗ | ✓ | ✗ | ✓ | - | 13,298 |
| IJB-C (Maze et al. 2018) | ✓ | ✓ | ✗ | ✗ | 3500 | 33,000 |
| UTKFace (Zhang, Song, and Qi 2017) | ✗ | ✓ | ✓ | ✓ | - | 20,000+ |
| RFW (Wang et al. 2019a) | ✓ | ✗ | ✓ | ✗ | 40,607 | 11,430 |
| BUPT-Balancedface, BUPT-Globalface, BUPT-Transferface (Wang and Deng 2020) | ✓ | ✗ | ✓ | ✗ | 28,000, 38,000, 10,000 | 1.3M, 2M, 0.6M+ |
| DiveFace (Morales et al. 2020) | ✓ | ✓ | ✓ | ✗ | 24,000 | 72,000 |
| FairFace (Karkkainen and Joo 2021) | ✗ | ✓ | ✓ | ✓ | - | 108,501 |

Table 1: Details of the databases used for bias study.

shown reduced effect of bias in model prediction.

Majority of the algorithms are focused on alleviating the influence of demographics on model predictions for enhancing fairness. However, (Terhörst et al. 2020a) demonstrated that face templates store non-demographic characteristics as well. Our assertion is that the biased prediction of models could be due to the non-demographic characteristics stored in the face images. Therefore, we believe that considering the non-demographic attributes during bias mitigation is important for designing robust systems.

## Facial Analysis Databases for Bias Study

Several facial analysis databases have been proposed in the literature. Initially, existing databases with demographic information were used for studying the effect of bias. (Rawls and Ricanek 2009) proposed the MORPH database with subjects belonging to different gender subgroups and ethnicity within the age range of 16 to 77 years. It is one of the largest databases for studying the effect of age on face recognition algorithms. The Indian Movie Face Database (IMFDB) proposed by (Setty et al. 2013) consists of face images of Indian actors and actresses. The database is annotated with age, gender, pose, expression, and the amount of occlusion present in an image. (Eidinger, Enbar, and Hassner 2014) proposed the Adience database containing face images labeled with age and gender information. The database is mainly used for age estimation. The Cross-Age Celebrity Dataset (CACD) proposed by (Chen, Chen, and Hsu 2015) contains images of celebrities with age ranging from 16 to 62 years. The LFWA (Huang et al. 2008) and CelebA (Liu et al. 2015) databases contain 40 annotated facial attributes and are commonly used for attribute prediction tasks. These databases are used for recent research that focuses on analyzing the performance of attribute prediction across protected attributes (male and young), followed by developing algorithms for bias mitigation (Tan, Shen, and Zhou 2020). The AgeDb (Moschoglou et al. 2017) and All-Age Faces

(AAF) (Cheng et al. 2019) databases are used for analyzing the performance of algorithms across age subgroups. Apart from this, the IJB-C database (Maze et al. 2018) is one of the largest databases with skin-tone information on the Fitzpatrick scale used for studying bias.

With the increased attention on understanding different aspects of bias in model prediction and developing fair algorithms, multiple databases are proposed for studying bias. (Zhang, Song, and Qi 2017) proposed the UTKFace database with more than 20K images having variations in pose, illumination, expression, occlusion, and resolution. (Buolamwini and Gebru 2018) proposed the Pilot Parliaments Benchmark (PPB) database for studying the effect of bias in gender classifiers w.r.t different skin tones. The PPB database consists of 1270 subjects from three African and three European countries. Authors labeled the skin tones of the subjects using the Fitzpatrick six-point labeling system. Racial Faces in-the-Wild (RFW) database proposed by (Wang et al. 2019a) is an unconstrained testing database for studying racial bias in face recognition. The database consists of Caucasian, Indian, Asian, and African racial subgroups. (Wang and Deng 2020) proposed four different training databases, namely BUPT-Balancedface (race-balanced), BUPT-Globalface (racial distribution approximately equal to the real distribution of world's population), BUPT-Transferface containing labeled data for Caucasian and unlabeled data for other subgroups (created for unsupervised domain adaptation), and MS1M-wo-RFW containing non-overlapping subjects of MS-Celeb-1M database for studying the effect of race on face recognition algorithms. (Morales et al. 2020) proposed the DiveFace database with annotations for gender and ethnicity (Caucasian, European, Asian) subgroups. Recently, the FairFace database is proposed (Karkkainen and Joo 2021), which is balanced across seven race subgroups: White, Black, East Asian, Middle Eastern, Southeast Asian, Indian, and Latino. The databases proposed for studying the effect of bias, shown in Table 1,
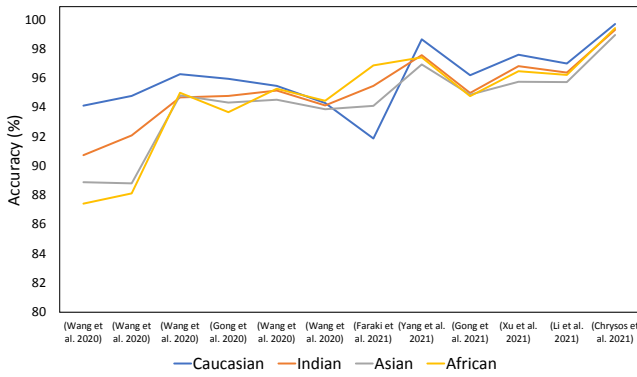
Figure 3: Meta-analysis of the verification accuracy reported on the RFW dataset across the four racial subgroups.

have escalated the research towards understanding bias and boosted the development of algorithms for bias mitigation.

## Open Challenges

Research towards bias and fairness has gained significant advancements, and several solutions have been proposed to improve the trustability and dependability of facial analysis systems. The trend in deep learning models has been towards improving fairness for different demographic subgroups. As illustrated in Figure 3, newer algorithms have shown improved fairness across different racial subgroups. Despite the progress achieved in designing fair solutions, there are various open challenges that require the attention of the research community. Here, we discuss some of the challenges that require more attention and focused research efforts.

**Fairness in Presence of Occlusion:** Wearing face masks has become a mandate in public places worldwide due to the COVID-19 pandemic. Thus, face recognition algorithms are required to recognize faces in the presence of masks. Masks occlude a major portion of the facial region that poses challenges to face recognition algorithms. To facilitate research in this direction, researchers have proposed multiple masked face databases. However, these databases contain limited demographic information. In a real-world scenario, it is important that the face recognition algorithms perform equally well across different demographic groups in the presence of occlusion. Limited research is done towards understanding the effect of bias in the presence of occlusion and more attention is required to develop fair algorithms.

**Fairness Across Intersectional Subgroups:** Majority of the research is performed to mitigate bias due to a single demographic group. Less attention is paid towards bias mitigation across intersectional subgroups. For instance, a model that is fair across gender subgroups may be biased towards old darker-skinned females. To ensure trustability, it is important that the output predictions of a model are fair across individual as well as intersectional demographic subgroups. Thus, more research is required towards identifying and mitigating bias across intersectional subgroups.

**Trade-off Between Fairness and Model Performance:** Bias mitigation may affect the overall model performance.

While mitigation algorithms increase fairness, and the trained models achieve equal performance across different demographic subgroups, the model performance on over-represented subgroups may reduce. It is challenging to simultaneously reduce the effect of bias without hampering the overall model performance.

**Benchmark Databases:** Multiple databases have been proposed in recent years for studying bias as shown in Table 1. A very limited set of databases provide identity information along with demographic information. Even for databases that provide both information, there is a clear lack of consistency among the demographic information provided. For example, different databases segregate data for different number of ethnic subgroups. We need to account for intra-class variations within demographic groups such as Indian and Asian which have huge diversity with respect to skin-tone and facial appearance. None of the existing databases provide the distribution of subgroups (based on skin tone or facial appearance) within an ethnicity. This poses challenge regarding interpretability of model predictions. Further, an imbalance in database w.r.t the unlabeled demographic groups may introduce bias in model prediction. In such scenarios, it becomes difficult to interpret the source of bias. Therefore, we believe that construction of large-scale benchmark databases with details of demographic information will help to decipher the cause of bias in model prediction and develop algorithms for bias mitigation.

**Unavailability of Complete Information:** Existing algorithms are based on the assumption that demographic information is available during training. However, due to privacy concerns and regulations in the real-world, the collection of demographic information or their use during training is precluded. Further, the protected attribute information can be noisy. This severely limits the applicability of existing bias mitigation algorithms and demands the need for the development of algorithms which do not require demographic information for bias mitigation.

## Conclusion

Fairness in model prediction is important for the trustability of AI systems. In recent years, growing attention is observed towards handling the problem of bias in model predictions, and significant progress is made towards understanding bias and developing mitigation algorithms. We presented a systematic review of bias in facial analysis tasks and provided a taxonomy of the algorithms proposed for bias mitigation. We also discuss some of the open challenges that require proper attention and continuous research efforts. We believe that solving or providing solutions for these open research problems is necessary for further reducing the accuracy gap across different subgroups and building trusted AI systems.

## Acknowledgements

# References

Acien et al., A. 2018. Measuring the gender and ethnicity bias in deep models for face recognition. In *CIARP*, 584–593.

Adeli et al., E. 2021. Representation learning with statistical independence to mitigate bias. In *WACV*, 2513–2523.

Alasadi, J.; Al Hilli, A.; and Singh, V. K. 2019. Toward fairness in face matching algorithms. In *FAT/MM Workshops*, 19–25.

Albiero et al., V. 2020. Analysis of gender inequality in face recognition accuracy. In *WACVW*, 81–89.

Amini et al., A. 2019. Uncovering and mitigating algorithmic bias through learned latent structure. In *AIES*, 289–295.

Barlas et al, P. 2021. To" See" is to Stereotype: Image Tagging Algorithms, Gender Recognition, and the Accuracy-Fairness Trade-off. *ACM HCI*, 4(CSCW3): 1–31.

Bruveris, M.; Gietema, J.; Mortazavian, P.; and Mahadevan, M. 2020. Reducing geographic performance differentials for face recognition. In *WACVW*, 98–106.

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

Celis, D.; and Rao, M. 2019. Learning facial recognition biases through VAE latent representations. In *FAT/MM*, 26–32.

Chen, B. C.; Chen, C. S.; and Hsu, W. H. 2015. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *IEEE Transactions on Multimedia*, 17(6): 804–815.

Cheng et al., J. 2019. Exploiting effective facial patches for robust gender recognition. *Tsinghua Science and Technology*, 24(3): 333–345.

Choi et al., K. 2020. Fair generative modeling via weak supervision. In *ICML*, 1887–1898.

Chuang, C.-Y.; and Mroueh, Y. 2021. Fair mixup: Fairness via interpolation. In *ICLR*.

Conger, K.; Fausset, R.; and Kovaleski, S. F. 2019. San Francisco Bans Facial Recognition Technology. https://tinyurl.com/y4x6wbos. Online; accessed 19 February 2021.

Denton, E.; Hutchinson, B.; Mitchell, M.; and Gebru, T. 2019. Detecting bias with generative counterfactual face attribute augmentation. *arXiv preprint arXiv:1906.06439*.

Deuschel, J.; Finzel, B.; and Rieger, I. 2020. Uncovering the Bias in Facial Expressions. *arXiv preprint arXiv:2011.11311*.

Dhar et al., P. 2020. An adversarial learning algorithm for mitigating gender bias in face recognition. *arXiv preprint arXiv:2006.07845*.

Drozdowski et al., P. 2020. Demographic bias in biometrics: A survey on an emerging challenge. *TTS*, 1(2): 89–103.

Dwork, C.; Immorlica, N.; Kalai, A. T.; and Leiserson, M. 2018. Decoupled classifiers for group-fair and efficient machine learning. In *FAT*, 119–133.

Eidinger, E.; Enbar, R.; and Hassner, T. 2014. Age and gender estimation of unfiltered faces. *TIFS*, 9(12): 2170–2179.

Georgopoulos et al., M. 2021. Mitigating Demographic Bias in Facial Datasets with Style-Based Multi-attribute Transfer. *IJCV*, 129(7): 2288–2307.

Gong, S.; Liu, X.; and Jain, A. K. 2020. Jointly de-biasing face recognition and demographic attribute estimation. In *ECCV*, 330–347.

Gong, S.; Liu, X.; and Jain, A. K. 2021. Mitigating face recognition bias via group adaptive classifier. In *CVPR*, 3414–3424.

Gwilliam, M.; Hegde, S.; Tinubu, L.; and Hanson, A. 2021. Rethinking Common Assumptions to Mitigate Racial Bias in Face Recognition Datasets. In *ICCVW*, 4123–4132.

Huang, C.; Li, Y.; Loy, C. C.; and Tang, X. 2019. Deep imbalanced learning for face recognition and attribute prediction. *T-PAMI*, 42(11): 2781–2794.

Huang, G. B.; Mattar, M.; Berg, T.; and Learned-Miller, E. 2008. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*.

Joo, J.; and Kärkkäinen, K. 2020. Gender slopes: Counterfactual fairness for computer vision models by attribute manipulation. In *FATE/MM*, 1–5.

Jung, S.; Lee, D.; Park, T.; and Moon, T. 2021. Fair Feature Distillation for Visual Recognition. In *CVPR*, 12115–12124.

Karkkainen, K.; and Joo, J. 2021. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *WACV*, 1548–1558.

Kim et al., B. 2019. Learning not to learn: Training deep neural networks with biased data. In *CVPR*, 9012–9020.

Klare et al., B. F. 2012. Face recognition performance: Role of demographic information. *TIFS*, 7(6): 1789–1801.

Kortylewski et al., A. 2019. Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In *CVPRW*.

Krishnan, A.; Almadan, A.; and Rattani, A. 2020. Understanding fairness of gender classification algorithms across gender-race groups. *arXiv preprint arXiv:2009.11491*.

Krishnapriya et al., K. 2020. Issues related to face recognition accuracy varying based on race and skin tone. *TTS*, 1(1): 8–20.

Li, Z.; and Xu, C. 2021. Discover the Unknown Biased Attribute of an Image Classifier. In *ICCV*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *ICCV*, 3730–3738.

Majumdar, P.; Chhabra, S.; Singh, R.; and Vatsa, M. 2020. Subgroup Invariant Perturbation for Unbiased Pre-trained Model Prediction. *Frontiers in Big Data*, 3: 52.

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 *ICCVW*, 3786–3795.

Majumdar, P.; Singh, R.; and Vatsa, M. 2021. Attention Aware Debiasing for Unbiased Model Prediction. In *ICCVW*, 4133–4141.

Maze et al., B. 2018. IARPA Janus Benchmark-C: Face dataset and protocol. In *ICB*, 158–165.

McDuff, D.; Ma, S.; Song, Y.; and Kapoor, A. 2019. Characterizing bias in classifiers using generative models. *arXiv preprint arXiv:1906.11891*.

Morales, A.; Fierrez, J.; Vera-Rodriguez, R.; and Tolosana, R. 2020. SensitiveNets: Learning agnostic representations with application to face images. *T-PAMI*.

Moschoglou et al., S. 2017. Agedb: the first manually collected, in-the-wild age database. In *CVPRW*, 51–59.

Muthukumar, V. 2019. Color-theoretic experiments to understand unequal gender classification accuracy from face images. In *CVPRW*.

Nagpal, S.; Singh, M.; Singh, R.; and Vatsa, M. 2019. Deep learning for face recognition: Pride or prejudiced? *arXiv preprint arXiv:1904.01219*.

Nagpal, S.; Singh, M.; Singh, R.; and Vatsa, M. 2020a. Attribute Aware Filter-Drop for Bias Invariant Classification. In *CVPRW*, 32–33.

Nagpal, S.; Singh, M.; Singh, R.; and Vatsa, M. 2020b. Diversity Blocks for De-biasing Classification Models. In *IJCB*, 1–9.

Ntoutsi et al., E. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs: Data Mining and Knowledge Discovery*, 10(3): e1356.

Osoba, O. A.; and Welser IV, W. 2017. *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Rand Corporation.

Paolini-Subramanya, M. 2018. Facial Recognition, and Bias. https://tinyurl.com/y7rat8vb. Online; accessed 19 February 2021.

Park, S.; Hwang, S.; Kim, D.; and Byun, H. 2021. Learning Disentangled Representation for Fair Facial Attribute Classification via Fairness-aware Information Alignment. In *AAAI*, volume 35, 2403–2411.

Quadrianto, N.; Sharmanska, V.; and Thomas, O. 2019. Discovering fair representations in the data domain. In *CVPR*, 8227–8236.

Ragonesi, R.; Volpi, R.; Cavazza, J.; and Murino, V. 2021. Learning unbiased representations via mutual information backpropagation. In *CVPR*, 2729–2738.

Ramaswamy, V. V.; Kim, S. S.; and Russakovsky, O. 2021. Fair attribute classification through latent space de-biasing. In *CVPR*, 9301–9310.

Rawls, A. W.; and Ricanek, K. 2009. MORPH: Development and optimization of a longitudinal age progression database. In *European Workshop, BioID*, 17–24.

Robinson et al., J. P. 2020. Face recognition: too bias, or not too bias? In *CVPRW*.

Roh, Y.; Lee, K.; Whang, S. E.; and Suh, C. 2020. FairBatch: Batch Selection for Model Fairness. *arXiv preprint arXiv:2012.01696*.

Serna, I.; Morales, A.; Fierrez, J.; and Ortega-Garcia, J. 2021a. IFBiD: Inference-Free Bias Detection. *arXiv preprint arXiv:2109.04374*.

Serna, I.; Peña, A.; Morales, A.; and Fierrez, J. 2021b. InsideBias: Measuring bias in deep networks and application to face gender biometrics. In *ICPR*, 3720–3727.

Serna et al., I. 2019. Algorithmic discrimination: Formulation and exploration in deep learning-based face biometrics. *arXiv preprint arXiv:1912.01842*.

Serna et al., I. 2020. SensitiveLoss: Improving Accuracy and Fairness of Face Representations with Discrimination-Aware Deep Learning. *arXiv preprint arXiv:2004.11246*.

Setty et al., S. 2013. Indian movie face database: a benchmark for face recognition under wide variations. In *NCVPRIPG*, 1–5.

Tan, S.; Shen, Y.; and Zhou, B. 2020. Improving the Fairness of Deep Generative Models without Retraining. *arXiv preprint arXiv:2012.04842*.

Tartaglione, E.; Barbano, C. A.; and Grangetto, M. 2021. EnD: Entangling and Disentangling deep representations for bias correction. In *CVPR*, 13508–13517.

Team, S. N. U. 2015. Google Photo App Labels Black Couple Gorillas. https://tinyurl.com/3npuwwbn. Online; accessed 19 February 2021.

Terhörst et al., P. 2020a. Beyond Identity: What Information Is Stored in Biometric Face Templates? In *IJCB*, 1–10.

Terhörst et al., P. 2020b. Comparison-level mitigation of ethnic bias in face recognition. In *IWBF*, 1–6.

Terhörst et al., P. 2020c. Face quality estimation and its correlation to demographic and non-demographic bias in face recognition. In *IJCB*.

Terhörst et al., P. 2020d. Post-comparison mitigation of demographic bias in face recognition using fair score normalization. *PRL*, 140: 332–338.

Vangara et al., K. 2019. Characterizing the variability in face recognition accuracy relative to race. In *CVPRW*, 0–0.

Wang, M.; and Deng, W. 2020. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *CVPR*, 9322–9331.

Wang et al., M. 2019a. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *ICCV*, 692–702.

Wang et al., T. 2019b. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *ICCV*, 5310–5319.

Wang et al., Z. 2020. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *CVPR*, 8919–8928.

Xu et al., X. 2021. Consistent Instance False Positive Improves Fairness in Face Recognition. In *CVPR*, 578–586.

Yang et al., Z. 2021. RamFace: Race Adaptive Margin Based Face Recognition for Racial Bias Mitigation. In *IJCB*, 1–8.

Yucer, S.; Akçay, S.; Al-Moubayed, N.; and Breckon, T. P. 2020. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *CVPRW*, 18–19.

Zhang, Z.; Song, Y.; and Qi, H. 2017. Age progression/regression by conditional adversarial autoencoder. In *CVPR*, 5810–5818.