Hindawi
IET Biometrics
Volume 2024, Article ID 1808587, 18 pages
https://doi.org/10.1049/2024/1808587

Research Article

On the Potential of Algorithm Fusion for Demographic Bias Mitigation in Face Recognition

Jascha Kolberg, Yannik Schäfer, Christian Rathgeb, and Christoph Busch

da/sec–Biometrics and Security Research Group, Hochschule Darmstadt, Darmstadt, Germany

Correspondence should be addressed to Christian Rathgeb; christian.rathgeb@h-da.de

Received 1 September 2023; Revised 4 December 2023; Accepted 29 January 2024; Published 23 February 2024

Academic Editor: Marjory Da Costa Abreu

With the rise of deep neural networks, the performance of biometric systems has increased tremendously. Biometric systems for face recognition are now used in everyday life, e.g., border control, crime prevention, or personal device access control. Although the accuracy of face recognition systems is generally high, they are not without flaws. Many biometric systems have been found to exhibit demographic bias, resulting in different demographic groups being not recognized with the same accuracy. This is especially true for facial recognition due to demographic factors, e.g., gender and skin color. While many previous works already reported demographic bias, this work aims to reduce demographic bias for biometric face recognition applications. In this regard, 12 face recognition systems are benchmarked regarding biometric recognition performance as well as demographic differentials, i.e., fairness. Subsequently, multiple fusion techniques are applied with the goal to improve the fairness in contrast to single systems. The experimental results show that it is possible to improve the fairness regarding single demographics, e.g., skin color or gender, while improving fairness for demographic subgroups turns out to be more challenging.

1. Introduction

Biometrics are already employed in many areas of life as automated algorithms. According to recent market value analyses, the biometrics market is expected to grow even more in the next years [1]. Automated algorithms, such as face recognition, have already outperformed human capabilities [2]. Therefore, these algorithms are also used in areas that can immediately and strongly impact an individual's life. For example, automated algorithms are used in the judiciary [3], healthcare [4], credit scoring [5], and other fields [6]. However, face recognition technologies are also error prone. For example, in the U.S., there are known cases where misidentifying a person as a wanted criminal has led to a wrongful arrest, accompanied by at least temporary imprisonment and inappropriate treatment from the police [7–9]. In this context, Garvie and Bedoya [10] documented a disproportional higher arrest and search rate of African-Americans based on face recognition software decisions. In addition to these individual cases, researchers have reported a difference in the performance of face recognition algorithms based on the demographic characteristics (skin color/ethnicity, gender, age) of the individual being identified or verified. Demographic bias in face recognition is already known in the field of human expert analysis: The so-called other-race effect describes the fact that people can recognize faces within their own demographic group better than faces of another demographic group [11]. Many researchers even refer to algorithmic bias as one of the most critical challenges in the field of biometrics [12–14].

In response to said issues, organizations, such as the Association of Computing Machinery, call for an immediate suspension of face recognition software [15]. Both in the U.S. [16] and in the EU [17], standards have been created to regulate automated algorithms with respect to demographic bias. There are now several proposed measurements to evaluate the fairness and demographic differentials of biometric algorithms [18–21]. Also, a vast number of techniques and algorithms have been put forward to mitigate demographic bias, mainly focused on face recognition [22]. Many different approaches are published trying to mitigate demographic bias [23], which include methods during the training process [24], the removal of sensitive attributes [25], and domain adaptation [26]. Most approaches focus on the verification

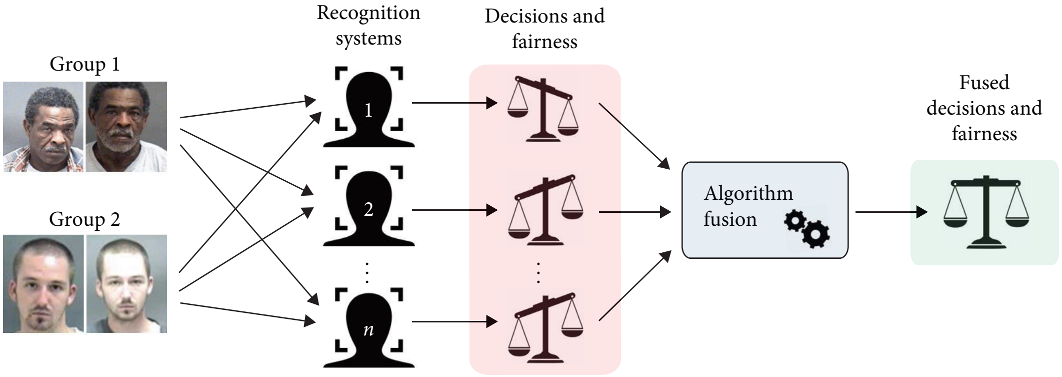


Figure 1: Overview of the proposed concept on different demographic (sub-) groups face recognition systems may exhibit performance differentials, whereas a fusion of multiple algorithms is expected to improve the overall fairness.

scenario, while only a few approaches consider the identification scenario [22].

In contrast to previously published works, this work investigates whether fusing multiple face recognition systems can mitigate demographic bias and make biometric systems fairer, as shown in Figure 1. Algorithm fusion has been successfully applied in the field of biometrics in order to achieve more robust recognition system. However, to the best of the authors' knowledge algorithm fusion has not yet been applied for the purpose of improving the fairness of face recognition. To do so, 12 different face recognition models are evaluated in verification mode with respect to accuracy and demographic fairness. The metrics used are general and demographic-specific false non-match rates (FNMRs) and false match rates (FMRs), as well as the resulting fairness metrics inequity rate (IR), fairness discrepancy rate (FDR), and Gini Aggregation Rate for Biometric Equitability (GARBE). In a case study, 33 different fusions are evaluated: These are composed of three selection criteria, three demographic attributes, and three to four types of fusions. The fusions applied are decision-level and score-level fusions. The decision-level fusions use the AND-, OR-, and Majority-Vote-operators. Score-level fusion is an equally weighted min–max normalized average fusion. The fusions are evaluated based on the selection criteria and the covariate under consideration. Fairness is evaluated within the three covariates of gender, skin color, and subgroups of gender and skin color.

In summary, this study presents a way to improve the fairness of biometric systems through carefully selected fusions. This gives providers of face recognition systems new opportunities to improve the fairness of their systems and helps to establish the equal treatment of individuals from different demographic groups. The key contribution of this paper can be summarized as follows:

(1) Twelve face recognition systems are benchmarked on the composite University of North Carolina at Wilmington (UNCW) dataset [27] to report their demographic bias toward the gender, skin color, and combined subgroups. The results are presented in terms of biometric performance in a verification scenario as well as a fairness score. Generally, it is observed that error rates are lower for males compared to females. Further, lower error rates are obtained for dark-skinned subjects compared to light-skinned subjects, while this different is less pronounced than the aforementioned gender accuracy gap.

(2) Multiple fusion schemes are implemented to combine the strengths of different face recognition systems. In this context, different fusion techniques are applied as well as different selection methods for possible fusion candidates.

(3) The fusion results are evaluated to understand whether the fairness score could be improved and how this fusion affects the biometric recognition performance. It is observed that biometric performance as well es fairness scores can be improved for distinct fusion approaches.

The rest of this paper is structured as follows: related work is reviewed in Section 2 and relevant metrics are defined in Section 3. Section 4 introduces terminology and concepts of our approach. The experimental evaluation is discussed in Section 5, and Section 6 concludes our findings.

2. Related Work

There are multiple works reporting the existence of demographic bias in face recognition. The following works estimate the demographic bias concerning different biometric applications, e.g., verification, identification, soft-biometric classification, and sample quality assessment. Most studies look at the verification scenario. Here, gender is the most commonly studied demographic attribute, followed by skin color, which is frequently referred to as ethnicity. A trend can be identified from the results of the different studies: The biometric performance is mostly better for male individuals [28–44], while only Lui et al. [45] found no difference in the performance of algorithms with respect to gender. Lower performance for females was also observed in classification tasks [46–48].

The analysis of bias in face recognition performance of different ethnicities is more challenging to assess due to a broader definition of ethnicity. The vast majority of studies only focus on the ethnicities East Asian, Caucasian, and Black. A common observation is that East Asians are the best-performing ethnicity, followed by Caucasians and dark-skinned people, who perform the worst in various studies [26, 28, 30–32, 38, 40, 42, 44, 46, 47, 49, 50]. However, other studies [11, 34, 41, 45, 51, 52] indicate that the performance differentials are not inherent to the different ethnicities and are a result of the own-race-effect and/or algorithm-specific training or implementation. The own-race effect causes algorithms to work best with ethnicities that originate from the same region as the algorithm training data.

For the identification scenario, the so-called watchlist imbalance effect has been examined [53, 54]. The effect describes the influence of the gallery composition on the performance of face recognition. Looking at the distribution of the gallery in terms of gender and skin color, the FMR is increasing for demographic groups with the proportion of the same demographic group in the watchlist.

Fairness measurement metrics have been introduced by different researchers. de Freitas Pereira and Marcel [55] proposed the FDR. The FDR is a fairness measurement that determines the fairness by the maximum absolute distance of the FMR and/or FNMR between two demographic groups at a certain decision threshold. Grother et al. [41] proposed the IR as a fairness measurement. In contrast to the FDR, the IR calculates the ratio between the worst and best FMR and/or FMR observed across demographic groups. Howard et al. [21] introduced a set of interpretable criteria referred to as the functional fairness measure criteria (FFMC). This measure was applied to identify shortcomings of the aforementioned fairness measurements based on which the same authors propose the Gini Aggregation Rate for Biometric Equitability (GARBE). When Grother [56] later published the “Face Recognition Vendor Test Part 3: Summarizing Demographic Differentials,” he added the FFMCs defined in Howard et al.’s [21] study and added two additional FFMCs. The mentioned FFMCs and fairness measurements will be detailed in the subsequent section.

In addition to the estimation of demographic bias, there are also numerous approaches that attempt to mitigate the bias. The approaches can be roughly divided into three categories. In the first category, there are approaches that focus on training [24, 30, 39, 57–63]. Some approaches focus on a training dataset that is as balanced as possible for the demographic covariates to be mitigated. Other approaches use specialized loss functions. For example, some algorithms are trained with more or fewer data from a particular covariate, depending on what results in the fairest outcome.

Another category of approaches dynamically selects the most appropriate recognition algorithm, decision threshold, or score normalization depending on the individual under consideration [42, 64–67].

Furthermore, some approaches try to obfuscate or remove an individual’s demographic information. Thus, the demographic covariate should not have any influence on the performance of the face recognition algorithms [68–71].

3. Fairness Metrics

Despite the biometric standardization community being working on standardizing fairness metrics [72], no final definitions are available for now. However, as mentioned before several metrics have been proposed by different researchers that will be described in detail as follows. Said metrics should fulfill five FFMCs [21, 56]:

(1) FFMC.1. The net contributions of FMR and FNMR differentials to the overall fairness measure should be intuitive when using a normal range of risk parameter weights and operationally relevant error rates.

(2) FFMC.2. There should be recognizable points of reference in the domain of the fairness measure, e.g., one bounded by known minimum and maximum possible values.

(3) FFMC.3. The fairness measure should be calculable when no recognition errors are observed for a demographic group. Given a finite image dataset partitioned into intersectional demographic groups, the likelihood that one group has zero FNMR rises with the number of groups.

(4) FFMC.4. The measure should reward more accurate algorithms if they distribute errors uniformly or in the same way as less accurate ones.

(5) FFMC.5. The measure should rank algorithms intuitively, correctly penalizing algorithms with the most nonuniform error rates.

Published fairness metrics, i.e., FDR, IR, and GARBE, have in common that they are composed of FNMR and FMR, as suggested in FFMC.1. In this regard, the formula is split into two terms each. An $A(\tau)$ term calculates the fairness concerning FMR, and a $B(\tau)$ term calculates the fairness with respect to FNMR. To flexibly and intuitively weigh the composition of the terms, a weighting parameter α (in the range $[0, 1]$) is used. A high α value means that FMR is strongly considered, and a low α value means that FNMR is more strongly considered. More specifically, for $\alpha = 0$ only the fairness concerning FNMR is computed, for $\alpha = 1$ only the FMR is considered, and for $\alpha = 0.5$ both rates are equally weighted.

3.1. Fairness Discrepancy Rate. The calculation of the FDR is shown in Equation (1) for two demographic groups d_i and d_j and a given decision threshold τ. The two fairness terms are determined by the largest difference in the FMRs and FNMRs of each demographic group. This means that fairness is generally lower when the system is more accurate, which partially contradicts FFMC.4. On the other hand, FDR can be computed, in case one error rate is 0, which fulfills FFMC.3. The main drawback of FDR is that while its theoretical range of values is between 0 and 1, as in FFMC.2 required, it uses only a small portion of that range in practice. In fact, the range is mostly narrowed between 0.9 and 1, as shown in Howard et al.’s [73] study. Since 1 means *fair* and 0 means *unfair*, this fact could lead to the impression that all systems are fair, even if it is not the case.

$$\begin{aligned}
A(\tau) &= \max\left(\left|\mathrm{FMR}_{d_i}(\tau) - \mathrm{FMR}_{d_j}(\tau)\right|\right), \forall d_i, d_j \in D \\
B(\tau) &= \max\left(\left|\mathrm{FNMR}_{d_i}(\tau) - \mathrm{FNMR}_{d_j}(\tau)\right|\right), \forall d_i, d_j \in D \\
\mathrm{FDR}(\tau) &= 1 - (\alpha A(\tau) + (1-\alpha)B(\tau)).
\end{aligned} \tag{1}$$

3.1.1. Inequity Rate. The IR is calculated based on ratio differences of max and min for FMR and FNMR separately. This is done for all demographic groups d_i and d_j as can be seen in Equation (2). A system with an IR close to 0 is considered fair and the higher the IR, the more unfair the system. The IR is not upper bounded, so it does not satisfy FFMC.2 and is difficult to classify alone without a reference system. In addition, the IR does not satisfy FFMC.3; if the error rate of a demographic group is 0, the metric is not defined since this leads to a division by 0:

$$\begin{aligned}
A(\tau) &= \frac{\max_{d_i}\mathrm{FMR}_{d_i}(\tau)}{\min_{d_i}\mathrm{FMR}_{d_i}(\tau)}, \forall d_i, d_j \in D \\
B(\tau) &= \frac{\max_{d_i}\mathrm{FNMR}_{d_i}(\tau)}{\min_{d_i}\mathrm{FNMR}_{d_i}(\tau)}, \forall d_i, d_j \in D \\
\mathrm{IR}(\tau) &= A(\tau)^{\alpha}B(\tau)^{1-\alpha}.
\end{aligned} \tag{2}$$

3.1.2. Gini Aggregation Rate for Biometric Equitability. The GARBE is inspired by the Gini coefficient and satisfies FFMC.1, FFMC.2, and FFMC.3. The GARBE can be calculated using Equation (3). The variable n represents the number of observations of the variable x, i.e., the number of demographic groups. x_i represents one observation from x, i.e., the FMR/FNMR of a demographic group, and $\overline{x}$ represents the mean of all observations x. The GARBE has a range of values of $[0, 1]$, where 0 is the fairest, and 1 is the most unfair system. Unlike the previous two fairness metrics, the GARBE considers the difference or ratio of the highest and lowest error rates of the demographic groups and includes all values in between. This matters when the fairness between more than two demographic groups is calculated, which is the case when combining, e.g., skin color and gender information:

$$\begin{aligned}
G_x &= \frac{n}{n-1}\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right|}{2n^2\overline{x}}, \forall d_i, d_j \in D \\
A(\tau) &= G_{\mathrm{FMR}_\tau} \\
B(\tau) &= G_{\mathrm{FNMR}_\tau} \\
\mathrm{GARBE}(\tau) &= \alpha A(\tau) + (1-\alpha)B(\tau).
\end{aligned} \tag{3}$$

4. Proposed System

While further discussion is required to standardize final definitions of fairness metrics, for this study, we follow the argumentation of Howard et al. [73] and use GARBE to compare the fairness of different systems since it satisfies the most FFMCs. Additionally, GARBE can differentiate very well between fair and unfair compared to FDR and its fixed range is easier to interpret than the unbound IR.

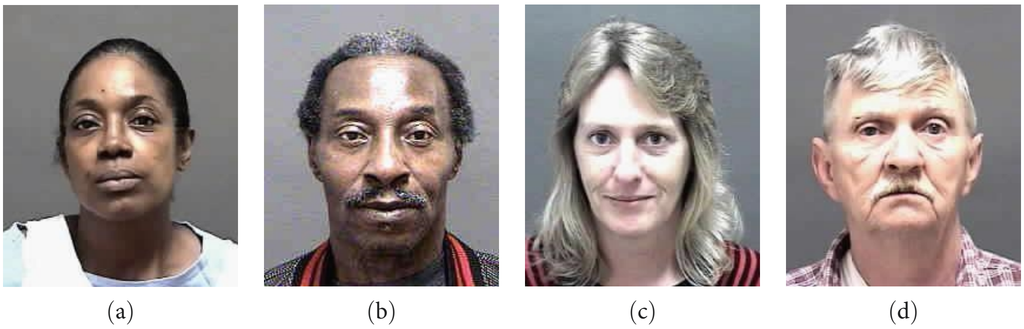
First, the facial images are processed by multiple face recognition systems and the biometric performance is reported. In addition, a fairness score is computed for each system. Finally, different fusion schemes are evaluated in terms of biometric performance as well as demographic fairness. The whole procedure is executed on the full database as well as on subsets for different demographic groups. More details for each step are provided in the following.

4.1. Face Recognition and Demographics. For our work, we want to use multiple face recognition systems. Each system is then evaluated in terms of biometric performance as defined in ISO/IEC 19795-1 [19] regarding FMR and FNMR. In this context, the FMR is fixed to 0.1%, as recommended, e.g., for border control [74], to benchmark the different systems regarding their FNMRs. Additionally, the biometric performance is monitored for separate demographic groups. With this, we can see how biased the different face recognition systems are toward specific demographics.

4.2. Pareto Efficiency. Pareto efficiency is an optimization method mainly used in economics. The idea of using Pareto efficiency for biometric systems originated from [75]. In this work, we can make use of Pareto efficiency to preselect biometric systems that lie on the Pareto curve. The Pareto-efficient systems are identified using FNMR and the GARBE with respect to FMR ($\alpha = 1$). With this setup, we combine all three inputs into a 2D Pareto curve. A system is Pareto efficient if no parameter (FNMR or GARBE) can be improved without worsening the other parameter.

4.3. Fusion Techniques. In general, there are many ways to fuse information in biometric systems [76]. In our approach, the two relevant fusion techniques are on decision level and on comparison score level. For the decision level fusion, the face recognition systems each compare their computed comparison score to the decision threshold. Subsequently, the decisions are fused using either AND/OR combinations or a majority voting. Especially the latter one requires an odd number of fused systems or a fallback strategy. For the score level fusion, each face recognition system computes one comparison score. Now, these scores need to be normalized to the same value range before fusing them. In our case, we use the min–max normalization to map all scores in the range of $[0, 1]$. The single comparison scores can then be weighted equally or a specific system can influence the final score to a larger amount. In any case, a new decision threshold is required for the fused system. This also implies a new calibration when different systems are fused or the weights are adjusted.

This leaves us with the question how to select the corresponding systems for the fusion. In this regard, we test three different approaches and evaluate how those improve the fairness metric as well as how they affect the biometric recognition performance:



(a) (b) (c) (d)

Figure 2: Example images from the used dataset: (a) dark female (df), (b) dark male (dm), (c) light female (lf), and (d) light male (lm).

(1) We select fusion candidates based on complementary FMRs. More specifically, focusing on one demographic characteristic (e.g., gender), we select the face recognition model with the lowest FMR for one group (e.g., female) and another model with the lowest FMR for the other group (e.g., male). By fusing both models, we hope that the strengths of both models combined can improve the fairness regarding this demographic characteristic (e.g., gender). The same selection process is applied for all demographic groups.

(2) The GARBE values are used to choose fusion candidates. Here, the idea is that the face recognition models with the best fairness scores are fused to hopefully complement each other, resulting in an even better fairness score.

(3) The Pareto efficiency for all models is computed. By visually inspecting the graph, the Pareto-efficient systems are identified and selected for the fusion. The computation of the Pareto efficiency relies on biometric performance as well as the fairness score, thus somehow combining both previous approaches.

5. Experimental Evaluation

This section provides information about the experimental setup including database preparation and selected face recognition models. Subsequently, the results for all selected demographic groups are presented and discussed.

5.1. Experimental Setup. The face image database used in this study is the UNCW-MORPH dataset [27], see Figure 2. More specifically, UNCW offers a free academic dataset (https://uncw.edu/oic/tech/morph_academic.html) and a commercial dataset (https://uncw.edu/oic/tech/morph.html), which comes in two parts and licensing options (we did only license the first part). We combined the first half of the commercial dataset with the free academic dataset to obtain a larger face database for our study. Both subsets do not contain identical images or subjects, which was checked using cryptographic hash functions on file side and face recognition systems for biometric comparisons.

Table 1: Number of subjects and images for the different overlapping demographic groups.

(a) Number of subjects

	Dark	Light	Total
Female	1,492	19,929	21,421
Male	10,085	4,127	14,212
Total	11,577	24,056	35,633

(b) Number of images

	Dark	Light	Total
Female	5,757	98,308	104,065
Male	70,875	71,084	141,959
Total	76,632	169,392	246,024

For the experiments, we split the UNCW database into smaller sets according to demographic attributes such as gender and skin color. It should be noted that we did not assign gender and ethnicity to the data subjects but used the available labels, coming with the database, as ground truth. In terms of gender, the database labels were binary, thus only distinguishing between female and male. In order to evaluate the influence of the skin color, we focused on the ethnicity labels *African* and *European* to have a clear separation in skin tones in these experiments. We are aware that these limitations do not represent all people, but we selected this setting to analyze bias reduction capabilities on clearly separable demographic subgroups. The idea is that those subgroups consist of a combination of two demographic attributes namely gender and skin color. In the following, the ethnic labels are discarded and the terms *dark* and *light* are used to separate the skin tones. The resulting demographic subgroups are therefore: dark female, dark male, light female, and light male. The resulting database comprises more than 246,000 images from 35,633 subjects, as can be seen in Table 1. This makes it one of the largest annotated databases providing demographic labels, which is captured in a controlled environment. When focusing on evaluating demographic fairness and bias, we do not want additional factors from unconstrained capture processes to influence the experimental results.

For the analysis of the demographic bias reduction capabilities, we need multiple face recognition systems in order to have a pool of possible fusion candidates to choose from. The selected face recognition systems should have state-of-the-art

Table 2: List of open source face recognition systems included in this study.

System	Model name	Backbone
ArcFace	af_casia	iResNet R100
	af_glint360k	iResNet R100
	af_ms1mv2	iResNet R100
	af_ms1mv3	iResNet R100
	af_mxnet	MXNet R100
	af_webface600k	iResNet R50
CurricularFace	curricularface	iResNet R101
ElasticFace	ef_arc	iResNet R100
	ef_arcplus	iResNet R100
	ef_cos	iResNet R100
	ef_cosplus	iResNet R100
MagFace	Magface	iResNet R100

For ArcFace (af) and ElasticFace (ef), multiple models were selected.

performance in terms of biometric recognition rates, thus only the leading open source models are used in this study.

The original ArcFace [77] is constantly updated [78] and retrained on new datasets. When looking at the different models (https://github.com/deepinsight/insightface/tree/master/model_zoo), that are made available by the authors, the reported performance increases for larger backbones (e.g., R100) compared to smaller ones (e.g., R50, R34, R18). Hence, we selected all available R100-models to be included in this study. However, for some of the pretrained models only R50 versions are available. Here, the *WebFace600K* model stands out since its reported performance is better than some of the previously selected R100 ones. Thus, this model is also included. The naming here mirrors the dataset, where the corresponding model was trained on, except for *mxnet*, which was trained on *MS1MV2* but builds upon a different backbone structure compared to the remaining models. From now on all ArcFace models are marked with the prefix `af_` followed by their original model name.

For the following open source face recognition systems, the selection process is more simple. The authors of CurricularFace [79] provide only one model (https://github.com/HuangYG123/CurricularFace) and MagFace [80] comes in multiple versions (https://github.com/IrvingMeng/MagFace), where again the R100-model is selected. Finally, ElasticFace [81] offers four pretrained models (https://github.com/fdbtrs/ElasticFace), which are all included. In the style of ArcFace, the ElasticFace models also get a prefix `ef_` to be discernible in the following. In addition to the 12 open source systems summarized in Table 2, one commercial off the shelf (COTS) face recognition system is also included in the benchmark. However, this system is not considered for the fusion approaches in order to grant full reproducibility of our results.

For more details on the specific pretrained face recognition models, the reader is referred to the descriptions of the original authors. For this study, we now focus on the demographic bias of each model and how to fuse them to improve the fairness.

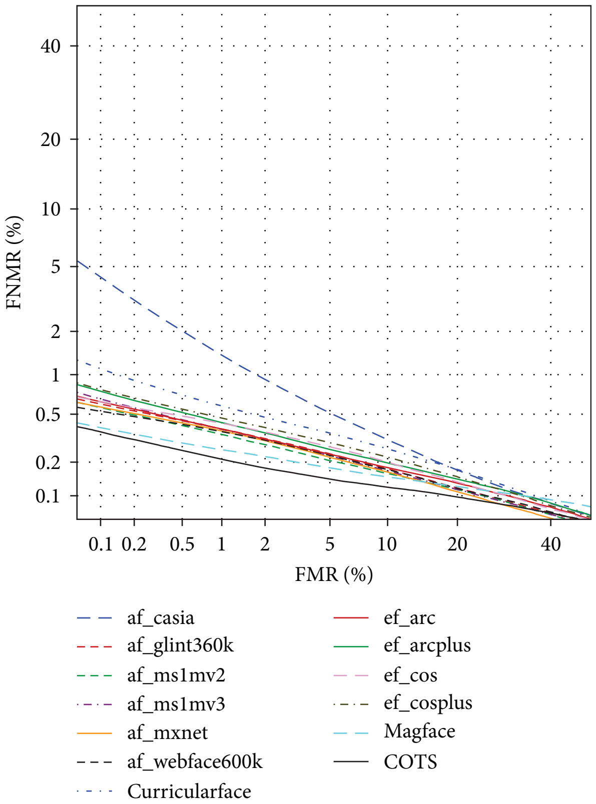


Figure 3: DET plot showing the FMRs and FNMRs for all selected models on the full database.

To retrieve comparable results of the different face recognition systems, RetinaFace [82] was used for face detection and alignment of the cropped face regions. Hence, all different face recognition models receive the same preprocessed face images as inputs.

5.2. *Benchmark Results.*

Figure 3 shows biometric performance on the full database for all selected models.

5.2.1. *False Non-Match Rate.*

Table 3 summarizes the FNMRs for all models and each demographic group. Since FMRs and FNMRs have a trade-off character, i.e., the higher the FMR is, the lower the FNMR is, we need to look at FNMRs in addition to FMRs. Magface is the best open-source model, and COTS is the best of all evaluated systems in terms of FNMR at a fixed FMR of 0.1%. The worst model with the highest FNMR is Casia. The statement that Magface is the best model and Casia the worst is feasible in this case, the comparison of all subjects with all subjects since the FMR is uniformly 0.1%. In the following observations regarding the FMRs of individual demographic groups, it must be noted that the FMR is not 0.1% for each individual group but varies. Thus, a statement regarding the improved accuracy or performance of the models cannot be made directly. However, it is still possible to say that a FNMR for certain groups is better or worse than others for a fixed FMR of 0.1% across all groups.

COTS performs best for each demographic group, but due to the smaller amount of comparison made with COTS

Table 3: FNMRs in percent for a fixed FMR of 0.1%.

Model	Demographic group							
	All	Female	Male	Dark	Light	df	dm	lf	lm
af_casia	4.341	5.539	4.093	2.525	4.987	2.874	2.517	5.632	4.799
af_glint360k	0.601	1.492	0.416	0.214	0.738	0.014	0.219	1.544	0.505
af_ms1mv2	0.561	1.449	0.377	0.211	0.686	0.014	0.216	1.499	0.450
af_ms1mv3	0.656	1.570	0.467	0.231	0.808	0.141	0.233	1.620	0.572
af_mxnet	0.568	1.523	0.369	0.207	0.696	0.014	0.211	1.576	0.440
af_webface600k	0.527	1.366	0.353	0.213	0.638	0.007	0.218	1.413	0.414
Curricularface	1.101	2.504	0.810	0.250	1.404	0.066	0.255	2.590	1.060
ef_arc	0.625	1.667	0.409	0.219	0.769	0.014	0.224	1.725	0.492
ef_arcplus	0.753	1.923	0.511	0.233	0.938	0.022	0.237	1.990	0.633
ef_cos	0.626	1.600	0.425	0.216	0.772	0.014	0.221	1.656	0.516
ef_cosplus	0.774	1.953	0.530	0.244	0.963	0.037	0.249	2.021	0.656
Magface	0.389	1.077	0.247	0.212	0.452	0.014	0.216	1.114	0.260
COTS	**0.359**	**0.971**	**0.232**	**0.205**	**0.415**	**0.007**	**0.210**	**1.004**	**0.242**

The lowest error rates are marked in bold.

and since it is not considered for fusion, only the FNMR values of the open-source models are described. First, we look at the gender-related columns. The highest FNMR among females has Casia with 5.539%; followed by the second worst model Curricularface with an FNMR of 2.504%. The best FNMR among females has Magface, with 1.077%. In the comparison among males, Casia is again the worst-performing model (FNMR = 4.093%). Again, Magface is the model with the lowest FNMR of 0.247%. These observations are the same as when looking at the comparisons of all subjects. In the direct comparison between the FNMR among males and females, it is clear that the FNMR among males is better than the FNMR among females for each model.

The lowest and thus best FNMR within dark-skinned individuals could be achieved with the Mxnet model and an FNMR of 0.207%; many of the other models have a only slightly higher FNMR. Casia has the highest and, thus, worst FNMR. Casia also has the highest FNMR for light-skinned individuals. This time, however, Magface has the lowest FNMR with an FNMR of 0.452%. If we again compare the values of all models concerning dark-skinned and light-skinned people, the FNMR for dark people always turns out to be significantly better than for light people. In all models, the FNMR is lower for dark-skinned people than for light-skinned people. This is the opposite pattern of the FMRs. The obtained results could be caused by a difference in quality of facial images of certain subgroups. It has been observed that especially for light-skinned female subjects parts of the facial region may be occluded by hair. On the contrary, male subjects may have more distinct facial hair covering some of their facial region. This could be one hypothesis for the observed results. However, a more detailed investigation of causes of bias is out of scope for this work.

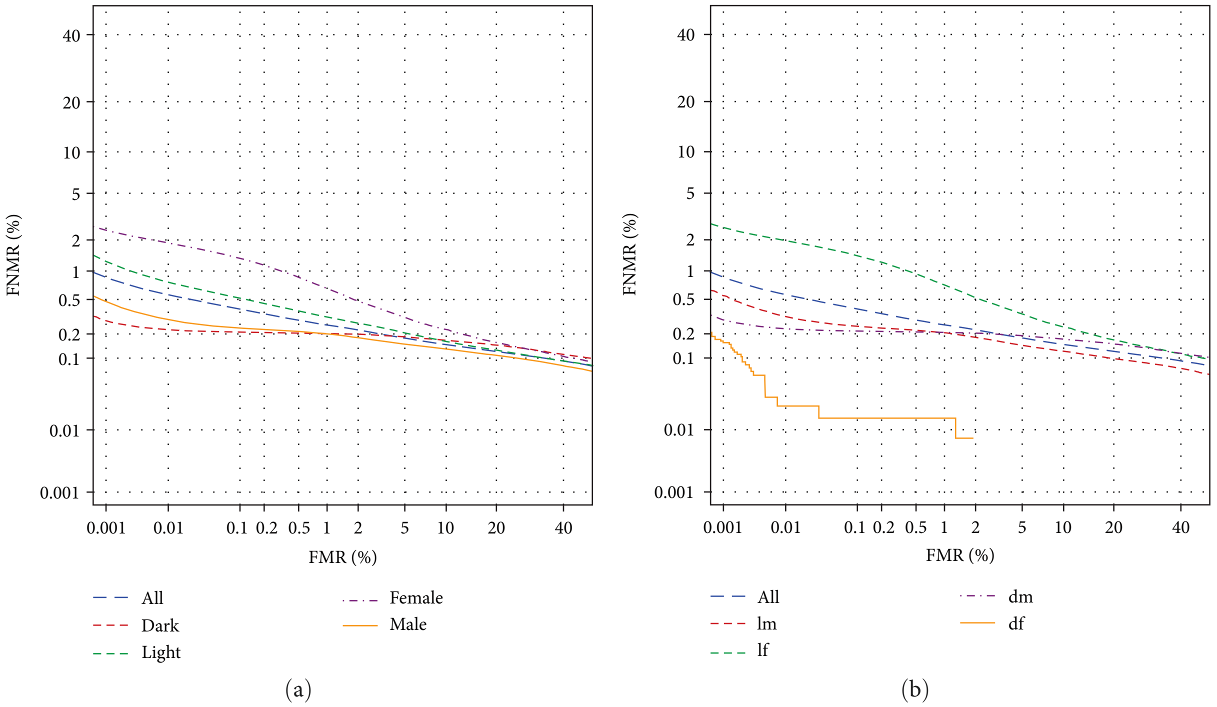
For the dark female subgroup, Webface600k provides the best FNMR with only 0.007%, and the FNMRs are quite close for all models except Casia, with an FNMR of 2.874%. Within the dark males subgroup, Mxnet has the lowest FNMR at 0.211%. Again, the values of all models are close to each other, except Casia. Magface has the best FNMR in the light females category, with an FNMR of 1.413%. In the light male subgroup, Magface also has 0.26%. Casia has the highest and, therefore, worst FNMR in each group. Comparing all subgroups with each other, for each model the FNMR is lowest or best for dark females, followed by dark males and light males. Light females have the worst FNMR in each model evaluated with an FMR across all groups of 0.1%. This observation is opposite to the behavior of FMRs. However, it is conclusive with the trade-off effect between FNMR and FMR: if FMR is higher, FNMR is lower, and conversely, if FMR is lower, FNMR is higher.

Figure 4 shows the biometric performance of the best open source model for each analyzed demographic group.

5.2.2. FMR Within Demographic Groups. Table 4 shows that the FMRs for all demographic groups are split into comparisons within each group and across different groups. Looking first at the FMR values within demographic groups in Table 4a, it is shown that with a global FMR of 0.1% across all demographic groups, the female group performs worse than 0.1% in most models. Only in Cosplus and Arcplus, the FMRs are below 0.1%. Also, in the male group, the FMR is higher than 0.1% in most models; only in Magface is the FMR significantly lower than 0.1% with 0.048%.

If we compare the male and female FMR values within their own demographic group, it is clear that in most models, the FMR of females is lower than the FMR of males. In 10 out of 13 models, females have a lower FMR than males, which means that a false match between females is less likely with these models than a false match between males. The two models that have a lower FMR among males are Webface600k and Magface.

Looking at the demographic groups dark and light, we notice that dark-skinned individuals have a higher FMR in most cases, as expected when comparing these individuals only among themselves. The FMR when comparing dark-skinned individuals is above 0.1% in all cases except in the



(a)

(b)

Figure 4: DET plot for the MagFace biometric recognition performance based on the particular demographic group considered: (a) main demographic groups and (b) demographic subgroups.

Table 4: FMRs in percent when using the global system threshold on demographic subsets.

(a) Comparisons only within the same demographic group

Model					Demographic group				
	All	Female	Male	Dark	Light	df	dm	lf	lm
af_casia	0.1	0.117	0.287	0.521	0.087	1.353	0.562	0.122	**0.017**
af_glint360k	0.1	0.101	0.307	0.450	0.086	0.748	0.531	0.108	0.155
af_ms1mv2	0.1	0.139	0.227	0.337	0.116	1.194	0.373	0.147	0.212
af_ms1mv3	0.1	0.162	0.167	0.267	0.129	0.890	0.296	0.174	0.069
af_mxnet	0.1	0.121	0.240	0.278	0.121	0.696	0.324	0.130	0.401
af_webface600k	0.1	0.184	0.138	0.215	0.153	0.673	0.239	0.203	0.114
Curricularface	0.1	0.123	0.191	0.319	0.093	1.630	0.315	0.115	0.080
ef_arc	0.1	0.103	0.300	0.471	0.081	1.161	0.538	0.105	0.111
ef_arcplus	0.1	0.086	0.327	0.519	**0.066**	1.360	0.578	**0.083**	0.108
ef_cos	0.1	0.116	0.255	0.386	0.096	1.177	0.421	0.118	0.183
ef_cosplus	0.1	**0.084**	0.315	0.436	0.079	0.908	0.496	0.084	0.254
Magface	0.1	0.246	**0.048**	**0.080**	0.195	**0.551**	**0.086**	0.278	0.047
COTS	0.1	0.188	0.170	0.269	0.150	1.025	0.306	0.207	0.122

(b) Comparisons only across demographic groups

Model	All	Gender	Skin	df–dm	df–lf	df–lm	dm–lf	dm–lm	lf–lm
af_casia	0.1	0.024	0.011	0.321	0.037	0.007	0.009	0.005	**0.009**
af_glint360k	0.1	0.030	0.029	0.156	0.030	0.026	0.022	0.064	0.026
af_ms1mv2	0.1	0.028	0.025	0.154	0.045	0.036	0.016	0.052	0.032
af_ms1mv3	0.1	0.030	0.029	0.125	0.057	0.025	0.025	0.031	0.025
af_mxnet	0.1	0.037	0.035	0.092	0.035	0.048	0.020	0.104	0.068
af_webface600k	0.1	0.023	0.016	0.098	0.043	0.019	0.012	0.020	0.037
curricularface	0.1	0.051	0.054	0.234	0.123	0.089	0.041	0.061	0.041
ef_arc	0.1	0.030	0.029	0.194	0.050	0.037	0.022	0.049	0.021
ef_arcplus	0.1	0.034	0.034	0.259	0.057	0.047	0.023	0.064	0.019
ef_cos	0.1	0.035	0.034	0.209	0.063	0.060	0.022	0.067	0.034
ef_cosplus	0.1	0.040	0.040	0.200	0.047	0.073	0.024	0.107	0.048
Magface	0.1	**0.006**	**0.005**	**0.023**	**0.020**	**0.001**	0.003	**0.002**	0.012
COTS	0.1	0.009	0.006	0.090	0.030	0.002	**0.002**	0.012	0.013

The lowest error rates are marked in bold.

case of Magface: In the case of Magface, the FMR is 0.08%. A different picture emerges here if we look at the comparison between light-skinned persons. For seven of 13 models, the FMR between light-skinned individuals is below 0.1%. Since one tends to expect higher values for comparisons within a demographic group than for comparisons between all groups, this is very striking. Unsurprisingly, 12 of 13 models have a better FMR of light-skinned versus dark-skinned individuals. The only exception is, again, MagFace.

For the subgroup dark females, the FMRs are all higher than 0.1%. Similarly, in the dark-males subgroup, the FMRs of all models are above 0.1%. The only exception is Magface, with an FMR of 0.086% in the comparison between dark males. In the light-females subgroup, most FMRs are also above 0.1%, Arcplus and Cosplus being the two exceptions. When comparing within the light-males subgroup, nine of 13 models still have an FMR above 0.1%. If we compare the FMR values of the subgroups, the following picture emerges: In most models (7/13), dark females have the highest FMR, followed by dark males, and light males. Light females have the lowest FMR. Looking at the Magface model, the order of descending FMR is dark females, light females, dark males, and light males. It is noticeable that dark females have the worst FMR in every model. In most cases (11/13), dark males have the second-worst FMR, while light females and light males have the best or second-best FMR in almost equal proportions (4 : 6). Exceptions are Mxnet and Magface, although dark females still have the highest FMR, this time light females and light males have the second-highest FMR, respectively. In summary, FMRs are generally higher within dark-skinned subgroups. And the FMR within dark-skinned females is higher than within dark-skinned males.

5.2.3. *FMR across Different Demographic Groups.*

Table 4b shows the FMR across demographic groups. That is, subjects from one demographic group are only compared to subjects from other demographic groups and not their own group. Since the two subjects being compared do not belong to the same demographic group and thus do not share certain characteristics (gender, skin color), it is expected that the FMR should be lower compared to the average value of 0.1% [18, 41]. This is also true in most cases. Comparing subjects of different genders, any model has no FMR above 0.1%. The best model is Magface with 0.006%, and the worst model is CurricularFace with an FMR of 0.051%. When comparing different skin colors, the same pattern emerges. No model has an FMR of more than 0.1%. Magface is the best model with an FMR of 0.005%, and Curricularface is the worst model with an FMR of 0.054%. Comparing the FMRs across gender with those across skin color, the values are very similar in magnitude. Only Casia can distinguish skin color with an FMR of 0.011%, significantly better than Gender with an FMR of 0.024%.

Looking at the FMRs between the subgroups, it is noticeable that especially the FMRs between dark females and dark males are relatively high. Only three of the 13 models have an FMR of less than 0.1% when comparing dark females and dark males. Besides COTS, one is Magface with 0.023% and only very close Webface600k with 0.098%. This observation aligns with the findings of Kolberg et al. [54], where the error rates across dark-skinned subgroups were also significantly higher than for other demographic subgroups.

In the comparison between dark females and light females, the FMR of 12 of 13 models is below 0.1%, only Curriuclarface performs worse in this category with 0.123%. Magface again performs best with 0.02%. The FMRs between dark females and light males are relatively low. Magface distinguishes best with an FMR of 0.001%, and Curricularface distinguishes worst with an FMR of 0.89%. When comparing dark males with light females, the same picture emerges: COTS distinguishes best with an FMR of 0.002%, followed by Magface with an FMR of 0.003%, and Curricularface distinguishes worst with an FMR of 0.041%.

When comparing dark males and light males, the FMR values are slightly higher for most models compared to dm–lf. For Mxnet and Cosplus, they are above 0.1%. The best differentiator is again Magface with an FMR of 0.002%.

The last comparison is between the groups light females and light males. No model has an FMR of more than 0.1%. The best model is Casia with an FMR of 0.009%. The worst model is Mxnet, with an FMR of 0.068%.

Table 5 lists the GARBE fairness scores for all different models and demographics.

For each model and metric, α was varied to include only FNMR fairness ($\alpha = 0$), only FMR fairness ($\alpha = 1$), and both equally combined ($\alpha = 0.5$). The GARBE is 0 for a very fair system and 1 for a very unfair system.

5.3. *Subgroups.*

Finally, we consider the fairness values with respect to the subgroups. The values can also be found in Table 5. As with the categories gender and skin color, the Casia model has the best fairness values with regard to FNMR, with a GARBE of 0.237. According to the GARBE, the Arc model is the least fair, with 0.732. The fact that Casia is the fairest model with respect to GARBE in all three considered categories regarding FNMR could again be related to the fact that Casia has generally higher FNMR values than all other models, which lowers the chance of a high ratio between the considered values (FNMR-male and FNMR-female, etc.) and thus makes a fairer impression than if the considered FNMR values are low. According to GARBE, Mxnet provides the best fairness between subgroups in terms of FMR. The Mxnet model has a GARBE of 0.38. Further, Curricularface is the least fair model with 0.755. Looking at both error rates combined ($\alpha = 0.5$), the fairest model with respect to GARBE is Casia with 0.479. The least fair model is Curricularface with 0.728.

5.4. *Summary—Individual Algorithms.*

We observe that GARBE, as mentioned earlier in the discussion of metrics, is a fairness metric with good predictive power with respect to equal treatment of different demographic groups. In contrast to other metrics, e.g., IR or FDR, the GARBE metric considers all error values, which becomes clear when assessing fairness between subgroups, since there are four demographic groups to consider there, instead of two. However, the results also show the weakness of GARBE, since it does not fulfill

Table 5: GARBE fairness evaluations for different demographics and different α factors, where $\alpha = 1$ emphasizes on FNMR fairness and $\alpha = 0$ emphasizes on FMR fairness.

Model	Gender			Skin color			Subgroups		
	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$
af_casia	**0.150**	0.284	0.419	**0.327**	0.520	0.713	**0.237**	**0.479**	0.720
af_glint360k	0.563	0.533	0.503	0.549	0.613	0.677	0.711	0.603	0.495
af_ms1mv2	0.586	0.414	0.241	0.528	0.507	0.487	0.716	0.643	0.571
af_ms1mv3	0.540	**0.277**	**0.014**	0.555	0.452	0.349	0.620	0.611	0.602
af_mxnet	0.609	0.469	0.329	0.541	0.467	0.349	0.730	0.555	**0.380**
af_webface600k	0.588	0.365	0.142	0.498	**0.333**	**0.168**	0.716	0.590	0.464
Curricularface	0.510	0.362	0.214	0.696	0.622	0.547	0.702	0.728	0.755
ef_arc	0.605	0.547	0.489	0.555	0.630	0.705	0.732	0.679	0.625
ef_arcplus	0.580	0.580	0.581	0.602	0.688	0.773	0.728	0.700	0.672
ef_cos	0.580	0.476	0.373	0.561	0.580	0.599	0.722	0.660	0.598
ef_cosplus	0.573	0.576	0.579	0.595	0.643	0.691	0.715	0.616	0.518
Magface	0.626	0.648	0.669	0.361	0.389	0.418	0.693	0.641	0.588
COTS	0.614	0.332	0.050	0.336	0.311	0.285	0.688	0.625	0.562

The best fairness values are marked in bold.

FFMC.4. When a system performs equally bad for all demographic groups, the GARBE fairness score is better than for other systems (c.f., af_casia GARBE scores).

5.5. Fusion Results. In the following evaluations of the fusions, we again only consider the GARBE fairness metric. Additionally, only FMR fairness ($\alpha = 1$) is evaluated, as the scope of this paper does not allow for further analysis of FNMR fairness values. The combined fairness of FMR and FNMR is also left out, for these values another circumstance is that the influence of the initial values is not directly comprehensible, since one does not know to what extent FMR or FNMR are causal for the result. Table 6 shows the fusion results when selecting the candidates based on their FMR performance. Table 7 shows the fusion results when selecting the candidates based on their GARBE scores. Table 8 shows the fusion results when selecting the candidates based on their Pareto efficiency. Figure 5 plots the Pareto efficiency for all tested systems based on the particular demographics. The Pareto curve combining all systems in the most lower left corner is called Pareto efficient and marked in green. These systems are then used for fusion and further evaluation.

5.5.1. Skin Color. Arcplus and Magface were chosen with the intention of aligning the two FMR values within light and dark subjects. Arcplus has the best FMR value for light–light comparisons, while Magface has the best FMR value for dark–dark comparisons. Subsequently, we expect a better fairness score over the demographic characteristic skin color from a fusion of these two models. In Table 6a, we first compare the new FMR values for dark–dark and light–light for the different fusions AND-, OR-, and Score-fusion with the values of the baseline models.

For the AND-Decision fusion, the FMR value of the fusion improves to 0.030% compared to the baseline models for dark–dark comparisons, and the FMR value of light–light also improves to 0.013% compared to both baseline models. Since these two values are relatively closer to each other than the values of the initial models, the GARBE measure also improves: the fairness score with respect to the fusion is 0.398, while the fairness score for the initial models is 0.418 for Magface and 0.773 for Arcplus. Thus, this fusion was able to improve fairness with respect to the considered fairness score. With the OR-Decision fusion, the fairness value changes to 0.392 and in the case of the Score-fusion to 0.302. Based on these values, the selection criterion of the initial models seems to make sense. Furthermore, the Score-fusion seems to have the best effect on fairness. The other values (FMR and FNMR within genders and subgroups) are not compared, since they were irrelevant for the selection of the initial models and the inclusion and discussion of these parameters exceed the scope of the thesis.

For GARBE, the candidates are the Webface600k, MS1MV3 and Mxnet models. The results of this fusion are shown in Table 7a. As previously observed, the AND-Decision fusion lowers the FMR across all demographic groups, in this case, reducing the FMR to 0.011%. The FNMR increases to 0.75% with the AND-Decision, as expected. The FMR between dark-skinned subjects decreases to 0.04% by AND-fusion, and for light-skinned subjects, the FMR decreases to 0.014%. As a result, the GARBE between the two FMRs is 0.477. Thus, the GARBE of the AND-fusion is not better than the GARBE of Webface600k, but it is better than the GARBE of MS1MV3 and Mxnet. An improvement over all the initial models could not be achieved with the AND-fusion. In contrast to AND-fusion, OR-fusion increases FMR across all groups to 0.242%, while FNMR improves to 0.441%, as expected. The FMR between dark-skinned subjects increases to 0.572%, and the FMR between light-skinned subjects increases to 0.328% compared to all baseline models. But GARBE deteriorates to 0.271 compared to all baseline models. The Majority-Vote-fusion improves FMR to 0.047%, while FNMR settles between baseline models at 0.562%. The FMR between dark-skinned subjects improves to 0.149%, while the

Table 6: Fusion results based on FMR candidates.

(a) Fusions of ef_arcplus and Magface to improve skin color fairness

	FMR All	FNMR All	FMR Dark	FMR Light	GARBE ($\alpha = 1$)
ef_arcplus	0.1	0.753	0.519	0.066	0.773
Magface	0.1	0.389	0.080	0.195	0.418
AND-fusion	**0.009**	0.782	**0.030**	**0.013**	0.398
OR-fusion	0.191	**0.361**	0.570	0.249	0.392
Score-fusion	0.1	0.393	0.275	0.147	**0.302**

(b) Fusions of ef_cosplus and Magface to improve gender fairness

	FMR All	FNMR All	FMR Female	FMR Male	GARBE ($\alpha = 1$)
ef_cosplus	0.1	0.774	0.084	0.315	0.579
Magface	0.1	0.389	0.246	0.048	0.669
AND-fusion	**0.007**	0.804	**0.014**	**0.014**	**0.006**
OR-fusion	0.193	**0.360**	0.317	0.351	0.051
Score-fusion	0.1	0.395	0.194	0.158	0.102

(c) Fusions of af_casia, ef_rcplus, and Magface to improve fairness of demographic subgroups

	FMR All	FNMR All	FMR df	FMR dm	FMR lf	FMR lm	GARBE ($\alpha = 1$)
af_casia	0.1	4.341	1.353	0.562	0.122	0.017	0.720
ef_arcplus	0.1	0.753	1.360	0.578	0.083	0.108	0.672
Magface	0.1	0.389	0.551	0.086	0.278	0.047	**0.588**
AND-fusion	**0.002**	4.431	**0.068**	**0.010**	**0.004**	**0.001**	0.842
OR-fusion	0.277	**0.349**	2.758	1.114	0.447	0.166	0.628
Majority-Vote	0.021	0.704	0.441	0.104	0.034	0.008	0.778
Score-fusion	0.1	0.408	1.400	0.357	0.201	0.044	0.702

Bold numbers mark best results.

FMR between light-skinned subjects drops to 0.062%. GARBE deteriorates to 0.416 compared to the baseline models. Score-fusion can again use the FMR of 0.1%. The FNMR across all groups improves to 0.486%. The FMR among dark-skinned subjects deteriorates to 0.296%, and among light-skinned subjects, it is in between the scores of the initial models with 0.091%. Thus, the Score-fusion achieves a GARBE value of 0.377, which is only better than the fairness value of Mxnet, but worse than that of MS1MV3 and Webface600k.

None of the tested fusions could improve the fairness of Webface600k, and all fusions perform worse than the fairest initial model in terms of GARBE.

For the skin color characteristic, only two models form the Pareto curve, i.e., Webface600k and Magface, as can be seen in Figure 5(a). Since only two models are on the Pareto curve, only these two are fused, and there is no opportunity for a Majority-Vote-fusion. The results are shown in Table 8a, and the AND-fusion again improves all FMR values and worsens the FNMR value across all groups to 0.573%. The GARBE improves to 0.076. With OR-fusion, the FNMR value improves compared to the baseline models, and the FMRs worsen. GARBE also improves with OR-fusion, this time to 0.084. With Score-fusion, we normalize the FMR across all groups to 0.1%. The FNMR improves slightly to 0.383%. The FMR between dark-skinned subjects is 0.141%, lower than that of Webface600k but higher than that of Magface. The FMR between light-skinned subjects is 0.182%, lower than that of Magface but higher than the FMR of Webface600k. The GARBE here is lower than that of the baseline model but higher than the GARBE of the other fusions with 0.127. The OR-fusion and AND-fusion are the only systems forming a new Pareto-curve and offer a Pareto-efficient trade-off between FNMR and GARBE.

5.5.2. Gender. To improve fairness between the female and male demographic groups, Cosplus and Magface were chosen. Cosplus has the lowest FMR for females, and Magface has the lowest FMR for males. Table 6b shows the fusion results. The AND-fusion results in the FMR of females and males being almost equal, both around 0.014%. This results in a fairness score of 0.006 for the AND-fusion, compared to the baseline models' fairness scores of 0.579 for Cosplus and 0.669 for Magface. The OR-fusion yields similar fairness scores: the GARBE value is decreased to 0.051. And the Score-fusion approach can also improve the fairness score to 0.102. The selection criterion appears to be appropriate for improving fairness, at least for these two models, for each fusion tested.

The fairest three models in the gender category are MS1MV3, Webface600k, and Curricularface. Table 7 summarizes the results. With the AND-fusion, the FMR across

Table 7: Fusion results based on GARBE candidates.

(a) Fusions of af_ms1mv3, af_mxnet, and af_webface600k to improve skin color fairness

	FMR	FNMR	FMR	FMR	GARBE
	All	All	Dark	Light	($\alpha = 1$)
af_ms1mv3	0.1	0.656	0.267	0.129	0.349
af_mxnet	0.1	0.568	0.278	0.121	0.393
af_webface600k	0.1	0.527	0.215	0.153	**0.168**
AND-fusion	**0.011**	0.750	**0.040**	**0.014**	0.477
OR-fusion	0.242	**0.441**	0.572	0.328	0.271
Majority-Vote	0.047	0.562	0.149	0.062	0.416
Score-fusion	0.1	0.486	0.296	0.134	0.377

(b) Fusions of af_ms1mv3, af_webface600k, and curricularface to improve gender fairness

	FMR	FNMR	FMR	FMR	GARBE
	All	All	Female	Male	($\alpha = 1$)
af_ms1mv3	0.1	0.656	0.162	0.167	0.014
af_webface600k	0.1	0.527	0.184	0.138	0.142
Curricularface	0.1	1.101	0.123	0.191	0.214
AND-fusion	**0.005**	1.197	**0.007**	**0.016**	0.430
OR-fusion	0.258	**0.447**	0.407	0.401	**0.007**
Majority-Vote	0.037	0.641	0.058	0.080	0.159
Score-fusion	0.1	0.510	0.145	0.230	0.227

(c) Fusions of af_glint360k, af_mxnet, and af_webface600k to improve fairness of demographic subgroups

	FMR	FNMR	FMR	FMR	FMR	FMR	GARBE
	All	All	Df	Dm	Lf	Lm	($\alpha = 1$)
af_glint360k	0.1	0.601	0.748	0.531	0.108	0.155	0.495
af_mxnet	0.1	0.568	0.696	0.324	0.130	0.401	0.380
af_webface600k	0.1	0.527	0.673	0.239	0.203	0.114	0.464
AND-fusion	**0.011**	0.698	**0.161**	**0.055**	**0.018**	**0.020**	0.608
OR-fusion	0.242	**0.445**	1.470	0.826	0.355	0.560	**0.375**
Majority-Vote	0.046	0.554	0.487	0.214	0.070	0.091	0.532
Score-fusion	0.1	0.487	0.888	0.404	0.156	0.210	0.479

Bold numbers mark best results.

all groups improves to 0.005%, while the FNMR across all groups deteriorates to 1.197%. The FMR among females improves to 0.007% and that among males to 0.016%. The GARBE of these two values is 0.430, which is worse than all GARBE values of the baseline models. In OR-fusion, the FMR increases to 0.258%, and the FNMR across all groups improves to 0.447%. The FNMR within female subjects increases to 0.407% and that between male subjects increases to 0.401%. The GARBE improves to 0.007, an improvement compared to all baseline models. The Majority-Vote again lowers the FMR among all demographic groups, and the FNMR is intermediate to the FNMRs of the baseline models. The FMR among female subjects drops to 0.058%, while it drops to 0.08% among male subjects. The GARBE fails to improve over all the baseline models and is 0.159. In Score-fusion, the FNMR improves to 0.51%. The FMR among female subjects is between all baseline models with 0.145%, and that of male subjects deteriorates to 0.230%. Thus, the GARBE worsens compared to the baseline models and amounts to 0.227%.

For the gender characteristic, MS1MV3, Webface600k, and Magface form the Pareto curve, as can be seen in Figure 5(b) and are selected for fusion. The results of the fusion are summarized in Table 8b. The AND-fusion again improves the FMRs, both within subjects of all groups and within the specific groups, female and male. In return, the FNMR worsens to 0.737% across all groups. The GARBE for the FMR among females and the FMR among males is 0.217 for the AND-fusion, which is fairer than Magface, but still more unfair than MS1MV3 and Webface600k. With OR-fusion, again, the opposite effect can be seen. The FNMR across all groups decreases, and the FMRs increase compared to all initial models. The GARBE is 0.27 and is slightly worse than the GARBE of the AND-fusion. The Majority-Vote-fusion also behaves like the Majority-Vote-fusions before: FMRs decrease across all groups and in the male and female demographic groups. The FNMR is within the range of the baseline models. GARBE improves to 0.145 compared to the AND- and OR-fusion, but the fusion is still more unfair than the MS1MV3 and Webface600k baseline models. Score-fusion improves the FNMR across all groups slightly to 0.388%. The FMR between females is 0.213%. The FMR within males is lower than that of MS1MV3 and Webface600k with 0.070% but higher than that of Magface. The GARBE is the highest compared to the

Table 8: Fusion results based on Pareto-efficient candidates.

(a) Fusions of af_webface600k and Magface to improve skin color fairness

	FMR	FNMR	FMR	FMR	GARBE
	All	All	Dark	Light	($\alpha = 1$)
af_webface600k	0.1	0.527	0.215	0.153	0.168
Magface	0.1	0.389	0.080	0.195	0.418
AND-fusion	**0.017**	0.573	**0.026**	**0.030**	0.076
OR-fusion	0.183	**0.344**	0.270	0.319	**0.084**
Score-fusion	0.1	0.383	0.141	0.182	0.127

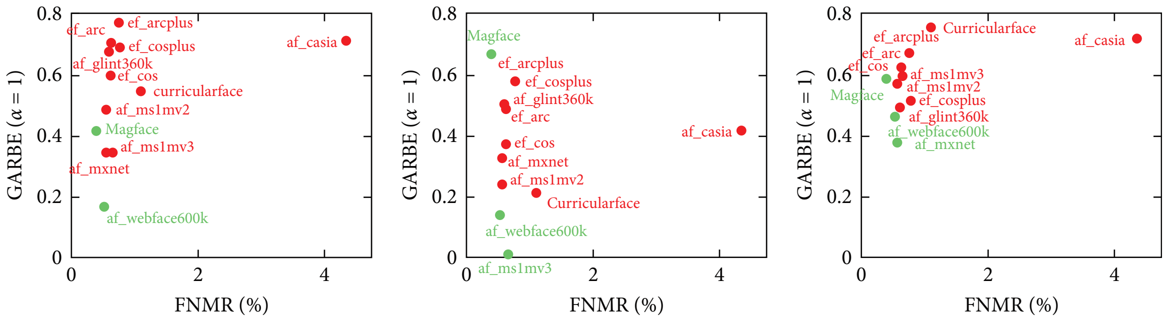
(b) Fusions of af_ms1mv3, af_webface600k, and Magface to improve gender fairness

	FMR	FNMR	FMR	FMR	GARBE
	All	All	Female	Male	($\alpha = 1$)
af_ms1mv3	0.1	0.656	0.162	0.167	**0.014**
af_webface600k	0.1	0.527	0.184	0.138	0.142
Magface	0.1	0.389	0.246	0.048	0.669
AND-fusion	**0.006**	0.737	**0.014**	**0.009**	0.217
OR-fusion	0.257	0.323	0.507	0.292	0.270
Majority-Vote	0.036	0.514	0.073	0.054	0.145
Score-fusion	0.1	0.388	0.213	0.119	0.280

(c) Fusions of af_mxnet, af_webface600k, and Magface to improve fairness of demographic subgroups

	FMR	FNMR	FMR	FMR	FMR	FMR	GARBE
	All	All	df	dm	lf	lm	($\alpha = 1$)
af_mxnet	0.1	0.568	0.696	0.324	**0.130**	0.401	0.380
af_webface600k	0.1	0.527	0.673	0.239	0.203	0.114	0.464
Magface	0.1	0.389	0.551	0.086	0.278	0.047	0.588
AND-fusion	**0.006**	0.656	**0.110**	**0.018**	0.014	**0.005**	0.724
OR-fusion	0.259	**0.317**	1.424	0.535	0.523	0.508	**0.308**
Majority-Vote	0.035	0.512	0.387	0.098	0.075	0.051	0.562
Score-fusion	0.1	0.391	0.922	0.239	0.229	0.111	0.541

Bold numbers mark best results.



(a)

(b)

(c)

Figure 5: Pareto efficiency plots for different demographics. The Pareto-efficient systems are marked in green, while Pareto-inefficient systems are shown in red. The efficiency level always depends on the other systems in comparison. (a) Pareto efficiency for skin color. (b) Pareto efficiency for gender. (c) Pareto efficiency for subgroups.

other fusions at 0.280%; compared to the baseline models, only Magface is more unfair. The OR-fusion and the Majority-Vote-fusion both form the new Pareto curve and are therefore pareto efficient.

5.5.3. Demographic Subgroups. In the comparison of fairness between subgroups, the Casia, Arcplus, and Magface models were used as initial models for the fusion. The models were chosen so that the selected models each included the lowest FMR for all subgroups. The results of the fusion are documented in Table 6b. The baseline models have a GARBE fairness score of 0.72 for Casia, 0.672 for Arcplus, and 0.588 for Magface. For the AND-fusion, all FMR scores of the subgroups improve compared to all baseline models. The GARBE fairness score is 0.842 for the AND-Fusion. For the OR-fusion, the FMR values certainly deteriorate as expected.

The fairness score also declines to 0.628. A deterioration in fairness also occurs with the Majority-Vote-fusion, with the GARBE score deteriorating to 0.778. With the Score-fusion, a mean GARBE value of 0.702 is reached. Fusions of these three models chosen by the selection criterion do not improve fairness between demographic subgroups compared to the initial models.

The three fairest models in terms of subgroups are Mxnet, Webface600k, and Glint360k. The results of the fusion are summarized in Table 7b. In the AND-fusion, the FMR decreases to 0.11%, while the FNMR between all subjects increases to 0.698%. The FMR of each subgroup decreases to 0.161% within dark females, to 0.055% within dark males, to 0.018% within light females, and to 0.02% within light males. The GARBE for the AND-fusion is 0.608, which is worse than that of the baseline models. For the OR-fusion, the FNMR worsens across all groups, as do the FMRs for each demographic group. The FNMR improves to 0.445%. GARBE also improves compared to the baseline models to 0.375. In the Majority-Vote-fusion, as in the AND-fusion, all FMR values improve. The FNMR is between the values of the initial models at 0.554%. GARBE is worse than the baseline models at 0.532. In Score-fusion, the FNMR again improves to 0.487%. The FMR of dark females worsens, and the FMR of dark males, light females, and light males is between the initial FMRs. The GARBE of the Score-fusion with 0.479 is only better than that of Glint360k but still worse than that of Mxnet and Webface600k.

The models Mxnet, Webface600k, and Magface form the Pareto curve concerning subgroups, as can be seen in Figure 5(c), are merged. The results are shown in Table 8. The AND-fusion again reduces all FMRs, while the FNMR increases to 0.656% within all subjects. The GARBE of 0.724 is higher than the GARBE of the initial models, making the AND-decision more unfair. For the OR-decision, the FMRs increase within all subjects and the subjects of the specific demographic groups. In return, the FNMRs are at 0.317%. The GARBE also drops to 0.308 in this case, making the OR-fusion fairer than all the baseline models. The Majority-Vote-fusion lowers the FMR across all subjects to 0.035%. The FNMR is between the values of the baseline models at 0.512%. The FMR among dark females decreases to 0.387%, as does the FMR among light females to 0.075%. The FMR within dark males and light males is between the values of the initial models. The GARBE of 0.562 is lower than that of Magface but still higher than that of Mxnet and Webface600k. The Score-fusion is again standardized to an FMR of 0.1%. The FNMR is slightly worse than the FNMR of Magface with 0.391%. FMRs within demographic subgroups lie between the baseline models, except dark females perform worse. The GARBE is 0.541, which is fairer than Magface but unfairer than Mxnet and Webface600k.

The OR-fusion is the only pareto-efficient system, when comparing the baseline systems and the other fusions.

5.6. *Summary—Algorithm Fusion*

5.6.1. Effect on Fairness. We conclude that 12 out of 33 mergers improved fairness under GARBE. Six of these are accounted for by the characteristics of gender and skin color and the selection criterion of the lowest FMR of each covariate considered. In those cases, every fusion improved the fairness of the baseline models. The OR-fusion accounts for two improvements in fairness regarding the selection criterion of the best fairness values. Three are accounted for by the fusions with the selection criterion of the Pareto curve, with only two models on the Pareto curve. And the last one is also accounted for by the OR-fusion with the selection criterion of the Pareto curve. The summary is visualized in Table 9.

Table 9: Summary of evaluated fusions showing whether they improved the fairness score or not.

Measure	Fusion	Gender	Skin color	Subgroups
FMR	AND	✓	✓	×
	OR	✓	✓	×
	Majority	–	–	×
	Score	✓	✓	×
GARBE	AND	×	×	×
	OR	✓	×	✓
	Majority	×	×	×
	Score	×	×	×
Pareto	AND	×	✓	×
	OR	×	✓	✓
	Majority	×	–	×
	Score	×	✓	×

5.6.2. Effect of the AND-Fusion. In every fusion performed, the AND-fusion led to a reduction in FMR between all subjects. This means that the overall probability of a false match can be significantly reduced with the AND-fusion. However, the AND-fusion significantly increases the FNMR between all subjects. This means the probability of a false nonmatch occurring is higher in systems with an AND-fusion.

It could also be shown that the FMR of the individual covariates decreases in each case due to the AND-fusion. More interesting is the effect of the AND-fusion on the fairness of the GARBE for FMR. In three out of nine cases, the AND-fusion improved fairness relative to all baseline models. This was the case for all the fusions using only two models. The AND-fusion improved fairness twice for the covariate skin color and once for gender. No improvement in subgroup fairness was possible with the AND-fusion. The AND-fusion combined with the selection criterion of the lowest FMR of a covariate appears promising when only two covariates (male and female or dark- and light-skinned) are considered. In both cases, fairness could be improved.

5.6.3. Effect of the OR-Fusion. The OR-fusion behaves opposite to the AND-fusion concerning FMR and FNMR. The FMR within all subjects and each covariate increases significantly compared to the baseline value of the merged models (0.1%). In turn, the FNMR decreases significantly. This secures the OR-fusion a place on the Pareto curve for the selection criterion every time. The OR-fusion can improve the fairness of the initial models in six out of nine cases. The OR-fusion could improve the fairness of all combinations created by the selection criterion of the lowest FMR per variate.

For the selection criterion based on the lowest fairness scores, OR-fusion improved fairness in two out of three cases. For the Pareto curve, fairness was only improved in one case.

5.6.4. Effect of the Majority-Vote-Fusion. The Majority-Vote-fusion could only be applied in six of the nine fusions. The FMR within all subjects could be reduced in every case. The FNMR, on the other hand, is always a value between the initial values of the baseline models. The FMRs of the individual covariates are mostly between the initial values and, in some cases, below.

The fairness could not be improved compared to all baseline models. In some cases, the fairness even worsened compared to all initial models.

5.6.5. Effect of the Score-Fusion. Score-fusion normalized the FMR to 0.1%, as in the baseline models. In five out of nine cases, the FNMR was reduced by Score-fusion. In the other four cases, it is comparable to the FNMR of the best baseline model. The FMR of the individual covariates is mostly between those of the baseline models, but in some cases, it is higher.

In three out of nine cases, fairness could be improved. Fairness was specifically improved when only two models were merged with the AND-fusion. In the other cases, the GARBE is between or even above that of the baseline models.

6. Conclusions

This work presented a benchmark of 12 open source face recognition systems on a common database. The overall biometric performance as well as the performance for specific demographic groups is evaluated as a baseline to inspect the raw system bias.

The main contribution of the work was to analyze whether fairness can be improved by fusing face recognition models. Since all possible combinations of models would have been a too large number to evaluate effectively, three selection criteria for models to be fused were formulated. The first selection criterion chooses the models with the lowest FMR for their demographic group. The second selection criterion selects the three models with the best fairness in terms of FMR. The last criterion selects the models based on the Pareto curve. For the selection criterion based on FMR, improvements were achieved for fairness concerning gender and skin color for all types of fusions (AND-, OR-, Majority-Vote-, and Score-fusion). However, fairness between subgroups could not be improved. For the selection criterion GARBE, the fairness could only be improved in two cases: in each case for the OR-fusion with respect to gender and demographic subgroups. For the Pareto efficiency, fairness could be improved with AND-, OR-, and Score-fusion for skin color demographics, while gender fairness could not be improved. Fairness for demographic subgroups could only be improved with the OR-fusion. In addition, we tested whether the fusions were Pareto efficient relative to the baseline models, thus adding better points in the Pareto curve. The OR-fusion is always Pareto efficient, while the AND- and Majority-Vote-fusion were Pareto efficient only in individual cases.

Based on these results, the following trends could be identified. The OR-fusion was most successful in improving fairness, while the Majority-Vote-fusion failed to improve fairness in any case. Fairness was best improved for skin color and gender, while the fairness of demographic subgroups could only be improved in two of 12 cases. The combination of two models seems to give better results regarding fairness than the combination of three models. The selection criterion of the lowest FMR seems to be the most effective to improve fairness.

The question of how the fusions influence the general performance, i.e., the FNMR, must be answered for each individual fusion. The OR-fusion always improves the FNMR for the cost of a worst FMR. The AND-fusion worsens the FNMR in every case, but on the other hand improves the FMR. The Majority-Vote-fusion mostly achieves an intermediate FNMR of the initial models, while the FMR can be significantly improved. The Score-fusion is the only one where we can maintain a fixed FMR of 0.1% with the effect that the FNMR also depends on the new threshold, thus varying increase and decrease. Accordingly, the choice of fusion is closely related to the application scenario and whether security or user convenience is preferred.

A general recommendation on how systems should be fused cannot be made from the above trends. This would require a statistical study for each criterion, fusion type, and demographics. However, the trends can be used to examine the different types of fusions, selection criteria, and demographics more closely and individually to avoid a flood of combinations.

Data Availability

The image data used to support the findings of this study were supplied by the University of North Carolina Wilmington under license and cannot be made freely available.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE. Open Access funding enabled and organized by Projekt DEAL.

References

[1] L. Pascu, “Global biometrics-as-a service to surpass $10B by 2030, contactless biometrics to top $18B by 2026,” 2020, [Online]. Available: https://www.biometricupdate.com/202008/global-biometrics-as-a-service-to-surpass-10b-by-2030-contactless-biometrics-to-top-18b-by-2026.

[2] F. Pasquale, *Black Box Society*, Harvard University Press, 2015.

[3] A. L. Washington, "How to argue with an algorithm: lessons from the COMPAS ProPublica debate," 2019, Rochester, NY [Online]. Available: https://papers.ssrn.com/abstract=3357874.

[4] K.-H. Yu and I. S. Kohane, "Framing the challenges of artificial intelligence in medicine," *BMJ Quality & Safety*, vol. 28, no. 3, pp. 238–241, 2019.

[5] M. Hurley and J. Adebayo, "Credit scoring in the era of big data," 2017, [Online]. Available: http://hdl.handle.net/20.500.13051/7808.

[6] Directorate-General for Parliamentary Research Services (European Parliament), C. Castelluccia, and D. Le Métayer, "Understanding algorithmic decision-making—opportunities and challenges," Publications Office of the European Union, 2019, LU: Publications Office of the European Union [Online]. Available: https://data.europa.eu/doi/10.2861/536131.

[7] K. Hill, "Wrongfully accused by an algorithm," 2020, The New York Times [Online]. Available: https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html.

[8] K. Hill, "Another arrest, and jail time, due to a bad facial recognition match," 2020, The New York Times [Online]. Available: https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html.

[9] E. Anderson, "Controversial detroit facial recognition got him arrested for a crime he didn't commit," 2020, Last Accessed: August 18, 2022 [Online]. Available: https://eu.freep.com/story/news/local/michigan/detroit/2020/07/10/facial-recognition-detroit-michael-oliver-robert-williams/5392166002/.

[10] J. F. C. Garvie and A. Bedoya, "The perpetual line-up," 2016, [Online]. Available: https://www.perpetuallineup.org/.

[11] P. J. Phillips, F. Jiang, A. Narvekar, J. Ayyad, and A. J. O'Toole, "An other-race effect for face recognition algorithms," *ACM Transactions on Applied Perception*, vol. 8, no. 2, pp. 1–11, 2011.

[12] A. Ross, S. Banerjee, C. Chen et al., "Some research problems in biometrics: the future beckons," 2019, [Online]. Available: http://arxiv.org/abs/1905.04717.

[13] C. Rathgeb, P. Drozdowski, N. Damer, D. C. Frings, and C. Busch, "Demographic fairness in biometric systems: what do the experts say?" 2021, [Online]. Available: http://arxiv.org/abs/2105.14844.

[14] A. K. Jain, D. Deb, and J. J. Engelsma, "Biometrics: trust, but verify," 2021, [Online]. Available: http://arxiv.org/abs/2105.06625.

[15] A. for Computing Machinery, "Acm us technology policy committee urges suspension of private and governmental use of facial recognition technologies," 2020, Last Accessed: August 18, 2022 [Online]. Available: https://www.acm.org/media-center/2020/june/ustpc-issues-statement-on-facial-recognition-technologies.

[16] E. Jillson, "Aiming for truth, fairness, and equity in your company's use of AI," 2021, Last Accessed: August 18, 2022, [Online]. Available: https://www.ftc.gov/business-guidance/blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai.

[17] European Commission, "New rules for artificial intelligence—questions and answers," 2021, Last Accessed: August 18, 2022 [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/QANDA_21_1683.

[18] J. J. Howard, Y. B. Sirotin, and A. R. Vemury, "The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance," in *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pp. 1–8, IEEE, 2019.

[19] ISO/IEC JTC1 SC37 Biometrics, ISO/IEC 19795-1:2021, "Information technology–biometric performance testing and reporting–part 1: principles and framework," International Organization for Standardization, 2021.

[20] R. T. Freitas, K. R. T. Aires, A. de Paiva, R. de M. S. Veras, and P. L. M. Soares, "A cnn-based multi-level face alignment approach for mitigating demographic bias in clinical populations," 2022, Rochester, NY [Online]. Available: https://papers.ssrn.com/abstract=4154044.

[21] J. J. Howard, E. J. Laird, Y. B. Sirotin, R. E. Rubin, J. L. Tipton, and A. R. Vemury, "Evaluating proposed fairness models for face recognition algorithms," 2022, [Online]. Available: http://arxiv.org/abs/2203.05051.

[22] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch, "Demographic bias in biometrics: a survey on an emerging challenge," *IEEE Transactions on Technology and Society*, vol. 1, no. 2, pp. 89–103, 2020.

[23] G. Boesch, "Deep face recognition: an easy-to-understand overview," 2022, [Online]. Available: https://viso.ai/deep-learning/deep-face-recognition/.

[24] M. Wang and W. Deng, "Mitigate bias in face recognition using skewness-aware reinforcement learning," 2019, [Online]. Available: http://arxiv.org/abs/1911.10692.

[25] A. Morales, J. Fierrez, R. Vera-Rodriguez, and R. Tolosana, "SensitiveNets: learning agnostic representations with application to face images," 2020, [Online]. Available: http://arxiv.org/abs/1902.00334.

[26] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang, "Racial faces in-the-wild: reducing racial bias by information maximization adaptation network," 2019, [Online]. Available: http://arxiv.org/abs/1812.00194.

[27] K. Ricanek and T. Tesafaye, "MORPH: a longitudinal image database of normal adult age-progression," in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pp. 341–345, IEEE, Southampton, UK, 2006.

[28] J. R. Beveridge, G. H. Givens, P. J. Phillips, and B. A. Draper, "Factors that influence algorithm performance in the face recognition grand challenge," *Computer Vision and Image Understanding*, vol. 113, no. 6, pp. 750–762, 2009.

[29] P. Grother, G. Quinn, and P. Phillips, "Report on the evaluation of 2d still-image face recognition algorithms," 2010.

[30] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain, "Face recognition performance: role of demographic information," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1789–1801, 2012.

[31] J. R. Beveridge, H. Zhang, B. A. Draper et al., "Report on the FG. 2015 video person recognition evaluation," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–8, IEEE, Ljubljana, Slovenia, 2015.

[32] H. Khiyari and H. Wechsler, "Face verification subject to varying (age, ethnicity, and gender) demographics using deep learning," *Journal of Biometrics & Biostatistics*, vol. 07, 2016.

[33] L. Best-Rowden and A. K. Jain, "Longitudinal study of automatic face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 148–162, 2018.

[34] S. H. Abdurrahim, S. A. Samad, and A. B. Huddin, "Review on the effects of age, gender, and race demographics on automatic face recognition," *The Visual Computer*, vol. 34, no. 11, pp. 1617–1630, 2018.

[35] B. Lu, J.-C. Chen, C. D. Castillo, and R. Chellappa, "An experimental evaluation of covariates effects on unconstrained face verification," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 1, pp. 42–55, 2019.

[36] N. Srinivas, M. Hivner, K. Gay, H. Atwal, M. King, and K. Ricanek, "Exploring automatic face recognition on match performance and gender bias for children," in *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pp. 107–115, IEEE, Waikoloa, HI, USA, 2019.

[37] C. M. Cook, J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury, "Demographic effects in facial recognition and their dependence on image acquisition: an evaluation of eleven commercial systems," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 1, pp. 32–41, 2019.

[38] I. Hupont and C. Fernández, "DemogPairs: quantifying the impact of demographic imbalance in deep face recognition," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pp. 1–7, IEEE, Lille, France, 2019.

[39] R. Vera-Rodriguez, M. Blazquez, A. Morales, E. Gonzalez-Sosa, J. C. Neves, and H. Proença, "FaceGenderID: exploiting gender information in DCNNs face recognition systems," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2254–2260, IEEE, 2019.

[40] I. Serna, A. Morales, J. Fierrez, M. Cebrian, N. Obradovich, and I. Rahwan, "Algorithmic discrimination: formulation and exploration in deep learning-based face biometrics," 2019, [Online]. Available: http://arxiv.org/abs/1912.01842.

[41] P. Grother, M. Ngan, and K. Hanaoka, "Face recognition vendor test part 3: demographic effects," 2019.

[42] J. P. Robinson, G. Livitz, Y. Henon, C. Qin, Y. Fu, and S. Timoner, "Face recognition: too bias, or not too bias?" 2020, [Online]. Available: http://arxiv.org/abs/2002.06483.

[43] V. Albiero, K. K. S., K. Vangara, K. Zhang, M. C. King, and K. W. Bowyer, "Analysis of gender inequality in face recognition accuracy," 2020, [Online]. Available: http://arxiv.org/abs/2002.00065.

[44] K. S. Krishnapriya, V. Albiero, K. Vangara, M. C. King, and K. W. Bowyer, "Issues related to face recognition accuracy varying based on race and skin tone," *IEEE Transactions on Technology and Society*, vol. 1, no. 1, pp. 8–20, 2020.

[45] Y. M. Lui, D. Bolme, B. A. Draper, J. R. Beveridge, G. Givens, and P. J. Phillips, "A meta-analysis of face recognition covariates," in *2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems*, pp. 1–8, IEEE, Washington, DC, 2009.

[46] J. Buolamwini and T. Gebru, "Gender shades: intersectional accuracy disparities in commercial gender classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pp. 77–91, PMLR, 2018.

[47] I. D. Raji and J. Buolamwini, "Actionable auditing: investigating the impact of publicly naming biased performance results of commercial AI products," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 429–435, Association for Computing Machinery, New York, NY, USA, 2019.

[48] V. Muthukumar, T. Pedapati, N. Ratha et al., "Color-theoretic experiments to understand unequal gender classification accuracy from face images," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2286–2295, IEEE, 2019.

[49] G. H. Givens, J. R. Beveridge, P. J. Phillips, B. Draper, Y. M. Lui, and D. Bolme, "Introduction to face recognition and evaluation of algorithm performance," *Computational Statistics & Data Analysis*, vol. 67, pp. 236–247, 2013.

[50] J. G. Cavazos, P. J. Phillips, C. D. Castillo, and A. J. O'Toole, "Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?" *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 1, pp. 101–111, 2021.

[51] D. Deb, L. Best-Rowden, and A. K. Jain, "Face recognition performance under aging," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 548–556, IEEE, Honolulu, HI, USA, Jul. 2017.

[52] S. Nagpal, M. Singh, R. Singh, and M. Vatsa, "Deep learning for face recognition: pride or prejudiced?" 2019, [Online]. Available: http://arxiv.org/abs/1904.01219.

[53] P. Drozdowski, C. Rathgeb, and C. Busch, "The watchlist imbalance effect in biometric face identification: comparing theoretical estimates and empiric measurements," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3750–3758, IEEE, Montreal, BC, Canada, 2021.

[54] J. Kolberg, C. Rathgeb, and C. Busch, "The influence of gender and skin colour on the watchlist imbalance effect in facial identification scenarios," in *Pattern Recognition, Computer Vision, and Image Processing*, vol. 13643 of *Lecture Notes in Computer Science*, pp. 1–13, Springer, Cham, August 2022.

[55] T. de Freitas Pereira and S. Marcel, "Fairness in biometrics: a figure of merit to assess biometric verification systems," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 1, pp. 19–29, 2022.

[56] P. Grother, "Face recognition vendor test (frvt) part 8: summarizing demographic differentials," National Institute of Standards and Technology (NIST), vol. 8429, 2022.

[57] Y. Guo and L. Zhang, "One-shot face recognition by promoting underrepresented classes," 2018, [Online]. Available: http://arxiv.org/abs/1707.05574.

[58] H. J. Ryu, H. Adam, and M. Mitchell, "InclusiveFaceNet: improving face attribute detection with race and gender diversity," 2018, [Online]. Available: http://arxiv.org/abs/1712.00193.

[59] M. A. Hasnat, J. Bohné, J. Milgram, S. Gentric, and L. Chen, "von mises-fisher mixture model-based deep learning: application to face verification," 2017, [Online]. Available: http://arxiv.org/abs/1706.04264.

[60] D. Deb, N. Nain, and A. K. Jain, "Longitudinal study of child face recognition," in *2018 International Conference on Biometrics (ICB)*, pp. 225–232, IEEE, 2018.

[61] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter, "Analyzing and reducing the damage of dataset bias to face recognition with synthetic data," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2261–2268, IEEE, 2019.

[62] M. Bruveris, P. Mortazavian, J. Gietema, and M. Mahadevan, "Reducing geographic performance differentials for face recognition," in *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pp. 98–106, IEEE, 2020.

[63] P. Smith and K. Ricanek, "Mitigating algorithmic bias: evolving an augmentation policy that is non-biasing," in *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pp. 90–97, IEEE, 2020.

[64] G. Guo and G. Mu, "Human age estimation: what is the influence across race and gender?" in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pp. 71–78, IEEE, 2010.

[65] D. Michalski, S. Y. Yiu, and C. Malec, “The impact of age and threshold variation on facial recognition algorithm performance using images of children,” in *2018 International Conference on Biometrics (ICB)*, pp. 217–224, IEEE, Gold Coast, QLD, Feb. 2018.

[66] K. K. S., K. Vangara, M. C. King, V. Albiero, and K. Bowyer, “Characterizing the variability in face recognition accuracy relative to race,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2278–2285, IEEE, 2019.

[67] P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, “Post-comparison mitigation of demographic bias in face recognition using fair score normalization,” *Pattern Recognition Letters*, vol. 140, pp. 332–338, 2020.

[68] M. Alvi, A. Zisserman, and C. Nellaker, “Turning a blind eye: explicit removal of biases and variation from deep neural network embeddings,” 2018, [Online]. Available: http://arxiv.org/abs/1809.02169.

[69] A. Acien, A. Morales, R. Vera-Rodriguez, I. Bartolome, and J. Fierrez, “Measuring the gender and ethnicity bias in deep models for face recognition,” in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, R. Vera-Rodriguez, J. Fierrez, and A. Morales, Eds., Lecture Notes in Computer Science, pp. 584–593, Springer International Publishing, Cham, 2019.

[70] P. Terhörst, N. Damer, F. Kirchbuchner, and A. Kuijper, “Suppressing gender and age in face templates using incremental variable elimination,” in *2019 International Conference on Biometrics (ICB)*, pp. 1–8, IEEE, Crete, Greece, Jun. 2019.

[71] P. Terhörst, N. Damer, F. Kirchbuchner, and A. Kuijper, “Unsupervised privacy-enhancement of face representations using similarity-sensitive noise transformations,” *Applied Intelligence*, vol. 49, no. 8, pp. 3043–3060, 2019.

[72] ISO/IEC JTC1 SC37 Biometrics, ISO/IEC 19795-10, “Information technology–biometric performance testing and reporting–part 10: quantifying biometric system performance variation across demographic groups,” International Organization for Standardization

[73] J. J. Howard, E. J. Laird, Y. B. Sirotin, R. E. Rubin, J. L. Tipton, and A. R. Vemury, “Evaluating proposed fairness models for face recognition algorithms,” arXiv preprint arXiv: 2203.05051, 2022.

[74] eu-LISA, “Best practice technical guidelines for automated border control (ABC) systems,” European Agency for the Management of Operational Cooperation at the External Borders of the Member States of the European Union, Tech. Rep.TT-02-16-152-EN-N, 2015.

[75] S. Wei and M. Niethammer, “The fairness-accuracy Pareto front,” *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 15, no. 3, pp. 287–302, 2022.

[76] M. Singh, R. Singh, and A. Ross, “A comprehensive overview of biometric fusion,” *Information Fusion*, vol. 52, pp. 187–205, 2019.

[77] J. Deng, J. Guo, and S. Zafeiriou, “ArcFace: additive angular margin loss for deep face recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4690–4699, IEEE, 2019.

[78] J. Deng, J. Guo, T. Liu, M. Gong, and S. Zafeiriou, “Sub-center ArcFace: boosting face recognition by large-scale noisy web faces,” in *Computer Vision–ECCV 2020*, vol. 12356 of *Lecture Notes in Computer Science*, Springer, Cham, 2020.

[79] Y. Huang, Y. Wang, Y. Tai et al., “CurricularFace: adaptive curriculum learning loss for deep face recognition,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5901–5910, IEEE, Seattle, WA, USA, 2020.

[80] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, “MagFace: a universal representation for face recognition and quality assessment,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14225–14234, IEEE, 2021.

[81] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, “ElasticFace: elastic margin loss for deep face recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1578–1587, IEEE, 2022.

[82] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, “RetinaFace: single-shot multi-level face localisation in the wild,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5203–5212, IEEE, 2020.