Fair Face Verification by Using Non-Sensitive Soft-Biometric Attributes

ESTEBAN VILLALOBOS1, DOMINGO MERY1, (Member, IEEE), AND KEVIN BOWYER2, (Life Fellow, IEEE)
1Department of Computer Science, School of Engineering, Pontificia Universidad Catolica de Chile, Macul 7820436, Chile
2Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA
Corresponding author: Esteban Villalobos (egvillalobos@uc.cl)
This work was supported by Fondecyt from CONICYT–Chile under Grant 1191131.

ABSTRACT Facial recognition has been shown to have different accuracy for different demographic groups. When setting a threshold to achieve a specific False Match Rate (FMR) on a mixed demographic impostor distribution, some demographic groups can experience a significantly worse FMR. To mitigate this, some authors have proposed to use demographic-specific thresholds. However, this can be impractical in an operational scenario, as it would either require users to report their demographic group or the system to predict the demographic group of each user. Both of these options can be deemed controversial since the demographic group is a sensitive attribute. Further, this approach requires listing the possible demographic groups, which can become controversial in itself. We show that a similar mitigation effect can be achieved using non-sensitive predicted soft-biometric attributes. These attributes are based on the appearance of the users (such as hairstyle, accessories, and facial geometry) rather than how the users self-identify. Our experiments use a set of 38 binary non-sensitive attributes from the MAAD-Face dataset. We report results on the Balanced Faces in the Wild dataset, which has a balanced number of identities by race and gender. We compare clustering-based and decision-tree-based strategies for selecting thresholds. We show that the proposed strategies can reduce differential outcomes in intersectional groups twice as effectively as using gender-specific thresholds and, in some cases, are also better than using race-specific thresholds.

INDEX TERMS Facial recognition, fairness, differential outcomes, soft-biometric attributes.

I. INTRODUCTION
Recent studies have pointed to potential demographic biases in facial analysis [1]–[4] and facial recognition [1], [5]–[8]. In 2020, the Association for Computing Machinery (ACM) called for a suspension of facial recognition technologies as they produce “(...) results demonstrating clear bias based on ethnic, racial, gender, and other human characteristics recognizable by computer systems” [9]. The central concern is typically that different demographic groups experience different false match rates.

This has also become a concern in Facial Verification (FV), which consists in validating a person’s identity by comparing their captured biometric information with a biometric template stored in the system database [44]. Here, a false match occurs when the similarity between images of two different people is strong enough that the two images are assumed to be of the same person. False matches are of particular concern because they can lead to unnecessary encounters with law enforcement. There are multiple recent incidents of an incorrect lead provided by face recognition not being competently investigated by law enforcement and thereby resulting in a false arrest [10], [11].

To control the number of false matches, typically a threshold is set on the similarity value between two images, so that only pairs of images whose similarity exceeds that threshold are declared a match. The threshold is set based on training data referred to as an impostor distribution, which is the distribution of similarity values between pairs of images of different persons. A typical threshold value is
To compute group-specific thresholds, we compare the strategies of defining groups based on Demographic Attributes (orange), Non-sensitive Soft-biometric Attributes (green), and Facial Embeddings (blue). These groups will be defined based on the metadata (in the case of the demographic attributes), clustering (for facial embeddings and soft-biometric attributes), and using a decision tree (for soft-biometric attributes). During the training phase, a threshold will be computed for each of the corresponding groups, which will be later used to determine a match/non-match. All methods will compute the similarity scores using the same facial embeddings.

one that results in only 1 in 10,000 impostor image pairs being above threshold. This ratio of number of false matches over number of impostor pairings is called the False Match Rate (FMR). Unfortunately, it has been pointed out that setting a FMR on a mixed-demographic dataset does not ensure that all demographics actually experience an equal FMR [5], [7], [12]. The National Institute of Standards and Technology (NIST) showed in a recent Face Recognition Vendor Test (FRVT) [5] that, for many algorithms, the FMR of some demographic groups could be 10 or 100 times higher than for others.

To mitigate the problem of error rate varying across demographics, some authors have suggested using demographic-specific thresholds, i.e., to set a different threshold for each demographic group [12]–[14]. In [14], Cavazos et al., state that “it is clear that a uniform threshold is not adequate or equitable when the underlying sub-population distributions differ”, and therefore suggest that demographic-specific thresholds are more adequate. Unfortunately, this requires the system to explicitly use the demographic information of the user. In [5], Grother et al., pointed out that, if one trusts the self-reporting of the demographic group, then some malicious agent may try to impersonate someone of a low-threshold group to generate a false match. To prevent that from happening, one might be enticed to use a classifier of demographic groups. Still, in many cases, it may not be desirable to try to predict someone’s demographic group [15]. There has been an increased desire for privacy regarding facial analysis, and demographic data is usually considered a sensitive attribute. Facial analysis, such as gender classification, has also been seen to have high error rates in LGBTQ+, and non-binary individuals [16]. Moreover, in [1], Qiu et al., found that false classifications of gender correlate with a false rejection of a true matching.

Another problem with using demographic-specific thresholds is that there is no consensus on how many demographic groups should be considered. Most studies considered gender [6], race [8], and age [17]. However, it can be possible to define combinations of those demographic groups. Unfortunately, there is little literature on whether choosing a threshold for one demographic group (e.g., gender) decreases or increases the differential performance on another (e.g., race). If one wishes to go further, it could be possible to select a threshold for each country or continent of origin, as some studies have found differential performances when considering that variable [5], [18]. All-in-all, there are still some gaps in the possibility of using demographic-specific thresholds.

In our work, we perform an in-depth analysis of the challenge of how to choose thresholds that mitigate differential outcomes in a fair FV context. We compare the current strategies of selecting a global threshold and demographic-specific thresholds with novel approaches that do not explicitly depend on demographic data (Fig. 1). Our proposed approaches consist on selecting a variable threshold based on (i) the clustering of the facial embedding features (ii) the clustering of non-sensitive soft-biometric features (such as hairstyle or accessories), and (iii) comparison-based decision trees to select the most informative soft-demographic attributes.

The main contributions of this paper are three-fold:

- To show the effect in FV of using demographic-thresholds based on a single demographic (i.e., gender or
Within-Demographic Imposters (WDI)

Cross-Demographic Imposters (CDI)

Enrolled

Probe

Enrolled

Probe

FIGURE 2. Scenarios when addressing demographic performance in facial verification. The Within-Demographic Imposters (WDI) Scenario consists on restricting comparisons to the same demographic group (in this work race and gender). The Cross-Demographic Imposters (CDI) Scenario allows imposter images to be from different groups.

race) and intersectional groups (i.e., race+gender) when testing on intersectional groups

• To compare automatic group-based thresholds strategies that do not depend on sensitive information (such as race and gender)

• To show that non-sensitive attributes can be an effective tool to mitigate differential outcomes across intersectional groups in FV

More details and experiments, like the performance analysis using other facial matchers, the effect dimensionality reduction on K-means performance using PCA in CDI scenario and WDI scenario can be found in the Master of Science Thesis [19].

II. RELATED WORK

In this section, we review the state of the art in two fields: studying the effects of soft-biometric attributes on facial verification (FV) and the efforts to achieve fairness in facial recognition.

A. EFFECTS OF SOFT-BIOMETRIC ATTRIBUTES ON FV

In [20], Dantcheva et al., defines soft-biometrics as the “physical, behavioral, or material accessories, which are associated with an individual, and which can be useful for recognizing an individual”. These include, but are not limited to, demographic attributes, hairstyle, face geometry, etc.

Several works have studied how restricting the demographic group of the imposter images affects the imposter distribution, selection of the optimal threshold, and the evaluation of the performance. This works can be summarized in two scenarios (see Fig. 2). On the one hand, the Within-Demographic Imposters (WDI) Scenario\(^1\) restricts comparisons to be between pairs of images of the same demographic group. This is a common approach to measure demographic performances of the methods [6], [8], [13] and to select demographic-specific thresholds [13]. However, this does not reflect any typical operational scenario, as it is not a common practice to restrict comparisons based on demographics. As noted in [5], [7], [14], using WDI leads to overall higher impostor scores because lower-similarity imposter pairs are not included in the distribution. Therefore, a higher threshold is required to ensure demographic groups fall below the desired FMR. This may lead to selecting thresholds that, in practice, produce a FMR much lower than the one reported but at the cost of producing a much higher FNMR. On the other hand, the Cross-Demographic Imposter (CDI) Scenario\(^2\) compares probe images to enrolled images from all different demographic groups. This is the standard approach to compute global thresholds [5], but it has been less explored as a scenario on which to measure the demographic performance of a system [12]. This means that a global threshold may be computed using an ‘easier’ distribution using CDI, but an analysis of bias may be performed using the more ‘difficult’ distribution of WDI [14]. When doing this, it is plausible that all demographic groups will have a FMR that falls above the Policy FMR, since using WDI tends to produce higher similarity scores. This may lead to wrong conclusions on whether or not certain demographics are observing the Policy FMR.

While demographic attributes are the most commonly studied [21], there are also studies on subject-specific attributes (e.g., hair style, expression and accessories) [22]–[25] and environmental context (e.g., illumination and resolution) [26]. In [24], Abate et al. made a comparison of different clustering algorithms on soft-biometric data. Their goal was to show that soft-biometric attributes can be clustered to provide sets of similar-looking subjects, which might help identify suspects in the presence of a challenging environmental context (e.g., occlusion). In [22] Terhörst et al., made a comprehensive study of the effect of 40 non-demographic attributes on differential outcomes. Their results found that many non-demographic attributes strongly affect the recognition performance of facial recognition models. They also show that, for ArcFace, the differential outcomes produced by specific attributes can vary significantly for different decision thresholds. Furthermore, in the context of FV, similarity scores can be influenced by whether both, one, or neither of the images present a soft-biometric attribute [7], [14], [27]. Overall, recent studies show that, even if facial embedding are trained with the goal of being robust to environmental and subject-specific attributes, they are currently still affected by non-demographic factors [25], [28], [29].

Our work proposes that subject-specific soft-biometric attributes can be used directly to select thresholds that mitigate differential performance. We propose clustering-based and decision tree-based strategies that do not depend on demographic information and show that they can reliably mitigate demographic differential performance under WDI and CDI scenarios.

B. FAIRNESS IN FACIAL RECOGNITION

There has been an increase in studies on bias in machine learning, such as hiring, recommendations, and facial

\(^1\) Also known as demographic yoking [14].

\(^2\) Also known as zero-effort imposters [5].
analysis [30]. Facial recognition studies have focused on group fairness, which can be defined as ‘treating similar groups similarly’ [30]. To achieve this, one can make changes before, during, or after the training process of an algorithm. These are classified, respectively, as follows [30]: a) preprocessing (e.g., ensuring balanced datasets [6], [31]–[33]), b) in-processing (e.g., include bias regularization terms in the training process [34]–[36]), or c) post-processing (e.g., varying the thresholds [12] or normalizing the scores [37], [38]).

While most studies focused on mitigating biases in facial recognition focus on preprocessing and in-processing, this approaches may require many additional resources to acquire facial data or computational resources to retrain the networks. Furthermore, even if one dedicated time and resources to ensure balanced datasets, Albiero et al. in [33] showed that balanced training data does not imply that algorithms achieve balanced error rates. Post-processing approaches usually require fewer resources to develop. These methods usually rely on either normalizing the comparison scores, learning new similarity metrics, or varying the thresholds.

Bias in facial verification can be studied by analyzing genuine-imposter curves or analyzing error rates after applying a threshold. It is essential to make the distinction on which metrics are helpful for each case. In [7], Howard et al. introduced the terms differential performance, referring to differences in genuine and imposter distributions, and differential outcome, for differences in error rates given a decision threshold. Many studies have focused on differential performance [13], [35], [39]–[41]. Consequently, the comparison of ROC Curves and AUC-ROC became very popular metrics to use [39]. These studies tend to report demographic ROC curves that only use same-group comparisons; therefore, they fall into the WDI Scenario. As for differential feature, some studies have used differentials of FNMR at a given FMR [37], [38], while others have focused directly on differences in FMR [5], [7], [12], [14], [34].

When using ROC (or AUC) to compare across demographic groups, something to consider is that different demographics typically achieve a particular FMR at different thresholds [13], [14]. This means that a ROC analysis typically does not reflect a comparison that would be achieved in an operational scenario.

This motivates the use of demographic-specific thresholds to mitigate biases [8], [12]–[14]. In [14], Cavazos et al. say that threshold setting and controlling imposters are scenario-modeling factors relating to race bias that are “under control of the user”. They state that in order to achieve equitable error rates on a system it is important to consider using group-specific thresholds. In [13], Vangara et al. show that, even though African-American faces have better ROC curves than Caucasian faces, they also have a worse FMR for any given threshold. They do their analysis using WDI, comparing images with faces of the same demographic group. In [12], one of the few studies that explore differential performance and differential outcomes using CDI, query images from any demographic group are allowed. They show that these issues can be addressed using demographic-specific thresholds for the intersectional groups of race and gender.

Studies exploring the use of demographic-specific thresholds have suggested choosing thresholds based on one demographic group (e.g., gender) and report the results on the same groups. There has been little study on how setting a threshold based on only one demographic (e.g., gender) affects the intersectional subgroups (e.g., gender and race). In [42], Grother et al. showed the impact on different demographic groups of choosing a global threshold such that white males achieved a certain FMR, but they did not show how setting a threshold based on only one demographic group (e.g., male or white) would impact the intersectional groups. They also reported only the results of a global threshold strategy and did not use variable thresholds.

To the best of our knowledge, our work presented here is the first work that explicitly uses non-sensitive soft-biometric attributes to define group-specific thresholds to mitigate differential outcomes. This is presented in contrast to the approach of using demographic-specific thresholds, something that has been done both implicitly by equalizing AUC-ROC [13], [35], [39]–[41], and explicitly [12]–[14]. For this, we compute thresholds based on only one demographic attribute and compute their efficiency on mitigating differential outcomes on intersectional groups. We propose that using non-sensitive attributes can be more efficient than the previous approach of a single demographic-threshold.

III. THRESHOLD STRATEGIES IN FV

We will begin this section by formalizing the Facial Verification (FV) problem. For a given input feature $X_P$ from a probe claiming to be of an enrolled identity $I$ with a template feature of $X_E$, the null and alternative hypotheses of the FV problem are [44]:

- $H_0$: Input $X_P$ does not come from the same person as $X_E$
- $H_1$: Input $X_P$ comes from the same person as $X_E$

With this, the associated decisions are

- $D_0$: person is not who they claim (non-match)
- $D_1$: person is who they claim (match)

Where, given a threshold $\tau$ and similarity score $s$, we choose $D_1$ if $s > \tau$ and $D_0$ otherwise.

This allows to define the error rates as follows

- $FMR = P(D_1|H_0)$
- $FNMR = P(D_0|H_1)$

Then, for a given similarity function $s$ and global threshold $\tau_{global}$, the classical decision problem could be defined as

$$D_{global \ \text{thr.}} := s(X_E, X_P) > \tau_{global} \quad (1)$$

A variable threshold strategy would change this definition and consider the following problem

$$D_{variable \ \text{thr.}} := s(X_E, X_P) > \tau_f(X_E, X_P) \quad (2)$$

where $\tau_f$ is a function that depends on the facial features (or other attributes).
The first strategy, which is the standard approach, uses a fixed threshold for the whole dataset. The second strategy chooses a different threshold for each demographic group, as in [12]–[14]. We compare these strategies with others that also use varying thresholds without using demographic data. We use clustering-based strategies on facial embeddings and on soft-biometric features, and use a decision tree-based strategy, which tries to maximize the information gained on the false matches based on the soft-biometric features. The reader is referred to Fig. 1 as general overview of the five strategies, Fig. 2 to see the differences between the WDI and CDI Scenario, and Fig. 3 as a guide for the features used in each strategy.

**A. FIXED GLOBAL THRESHOLD**

The global threshold will be chosen as the one that ensures a given FMR in the training set. The Policy FMR will be $10^{-3}$ or lower, as recommended by the European Border Guard Agency Frontex [45]. While some authors suggest that this threshold should be set using WDI [5], [14], it is usually computed using CDI [46]. The Fixed Global Threshold strategy is the standard approach in facial verification, so we will use it as a baseline to compare against.

**B. DEMOGRAPHIC THRESHOLDS**

The most direct way to ensure that every demographic group follows the Policy FMR is to compute a different threshold for each demographic group. There is also a need to define which demographic group must be used to set the thresholds. In this work, we will compare the use of group-specific thresholds for a) gender (Male, Female), b) race (Asian, Black, White, Indian), and c) combinations of race and gender.

A problem of this approach is which threshold should be used when comparing imposters from different demographic groups. In this work, we choose the threshold based on the ground-truth demographic group of the enrolled image, without considering the demographic group of the probe image.

**C. EMBEDDING CLUSTERING THRESHOLDS**

It has been shown that facial embeddings encode information about demographic groups, even if they are not explicitly given that information in training [36], [47], [48]. In [49], Terhörst et al. found that it was possible to accurately predict 74 out of 113 soft-biometric attributes using facial embeddings. This suggests that facial embeddings encode more information than just identity. As so, we compare the use of demographics with directly clustering the feature embeddings.

In our work, training features will be clustered using K-Means, and for each cluster, we will choose a threshold such that the cluster achieves the probe FMR. To compute the thresholds for each cluster we will follow an approach similar to that of the CDI Scenario, in the sense that we will allow comparisons of images between clusters. This means that the threshold will be selected by choosing all probe images in that cluster but allowing query images from different clusters.

---

3 Clustering was performed using the MiniBatchKMeans implementation of scikit-learn version 0.24.2 with default parameters.
Later, when testing, we will use the cluster of the probe image to select the threshold.

D. SOFT-BIOMETRIC CLUSTERING THRESHOLDS

Even if facial embeddings carry more information than just identity, it could be a better alternative to cluster soft-biometric attributes directly. It has been shown that many non-demographic soft-biometric attributes strongly affect recognition performance [22], [23]. The MAAD-Face dataset includes 47 binary attributes, out of which 7 correspond to demographic information. Since this work aims to implement thresholds that do not depend on demographics, we will exclude these attributes from the clustering. We also removed the ‘Attractive’ and ‘Chubby’ attributes, as they could perpetuate standards of beauty associated with one culture. This means that each image in the training set will be associated with a feature vector of 38 binary (non-demographic) soft-biometric attributes, such as ‘is bald’, ‘has a mustache’, and ‘is wearing makeup’. We call these 38 attributes the non-sensitive soft-biometric attributes. All there attributes were predicted using a Massive Attribute Classifier (MAC). They have an average reported accuracy of 89.8% [27] and the worst reported attributes have 68% of accuracy (‘bags under eyes’ and ‘brown eyes’).

As with the facial embeddings, these features will be clustered using K-Means. We will use the threshold that achieves the Policy FMR using the facial embeddings for verification for each cluster. Training will be done allowing query images to belong to different clusters and, when testing, we will choose the threshold based on the cluster of the probe image.

E. DECISION TREE-BASED THRESHOLDS

While previous strategies focused on assigning individual images to a specific group, facial verification consists of classifying pairs of images. As such, the similarity score can be influenced by whether both, one, or neither of the images have a soft-biometric attribute or belong to a certain group [7], [14], [27]. To find attributes that might convey a lot of information on false matches, we will use an information-based decision tree model as suggested in [7], [26].

To measure the amount of information on false matches, we will use the Shannon Entropy

\[ E(Y) = - \sum_i p_i \log_2(p_i) \tag{3} \]

where \( Y := D_{global}|H_0 \) is a random variable representing the probability of a false match on a sub-group of comparisons using the global threshold, and \( p_i \) is the probability of a pair of images being either \( i \in \{ \text{false match, true no-match} \} \). To quantify the effect of knowing an attribute in the false matches, we will use the information gain of the error rate given the attribute

\[ IG(Y, X) = E(Y) - E(Y|X) \tag{4} \]

where \( E(Y|X) \) is the entropy of the false matches given that we know the variable \( X \). In our case, this variable is whether both, one, or neither of the images have a certain attribute (e.g., \( X = (\text{both are bald, only one has glasses, neither has black hair}) \)). In the case where one of the images presents the attribute (e.g., only one has glasses) it will be equivalent if the probe or the query image is the one presenting that attribute.

This technique allows us to create a decision tree model, where each branch is based on the attribute that gives the most information gain. At each leaf of the tree, we will compute a threshold such that the pairs of images that fall on that leaf achieve the desired FMR.

IV. EXPERIMENTAL METHODOLOGY

A. DATASETS

Our work is based on the Balanced Faces in the Wild (BFW) [12], [50] and MAAD-Face [27], [51] datasets. Both are datasets based on VGGFace2 [52]. BFW is a dataset balanced across race (i.e., Asian, Black, Indian, and White) and gender (i.e., Female and Male). It has an equal number of identities per subgroup (100 per subgroup) and faces per identity (25 faces), for a total of 20K images of 800 subjects. BFW has five pre-defined, person-disjoint folds for five-fold cross-validation to estimate the accuracy. MAAD-Face is an extension of VGGFace2 with annotations from 47 soft-biometric attributes. From them, we select the 38 non-sensitive soft-biometric attributes as explained in Section III-D. With 123.9M attribute annotations, MAAD-Face is currently the largest face annotation dataset. The selected non-sensitive attributes were predicted using a Massive Attribute Classifier (MAC) with a mean reported accuracy of 89.8% [27].

Accuracy is reported as the average across 5-fold cross-validation. For each image, 475 imposters were selected from the same fold and all genuine pairs were used. We also removed 3 images that contained bugs reported by the authors of BFW (wrong identity and cartoon faces) and another 4 images that were not present in MAAD-Face. In total, we used 239,880 pairs of genuine faces and 9,497,625 imposter pairs separated into 5 folds. In WDI Scenario we will restrict query images to be from the same race and gender of the probe image. In CDI Scenario we will have no such restriction, so imposters can be of any demographic group. In both cases we will sample the same amount of images, so we will have the same amount of imposters.

B. TRAINING PROCESS

The training process consisted of two steps: defining groups (Fig. 3) and computing thresholds for each group.

---

4Even though our research involves human beings (i.e., recognition of human faces), in our paper, we don’t provide a statement to confirm that informed consent was obtained because we use public datasets of faces. Details of how the images were acquired can be obtained in the corresponding references.

5https://github.com/visionjo/facerec-bias-bfw/blob/master/data/README.md#reported-bugs

6"n009142/0501_01.jpg", "n001555/0307_02.jpg", "n009142/0501_01.jpg", "n001555/0307_02.jpg"
For all methods, we computed the facial embeddings using ArcFace (Resnet-101) [43], which provides a 512-dimensional vector for each facial image. The demographic groups are created based on the metadata provided on BFW. We will use all possible combinations of race and gender. In the Embedding Clustering strategy, we will use K-Means to cluster the facial embeddings in the training set given by ArcFace. We will also use K-Means to cluster the non-sensitive soft-biometric attributes provided by MAAD-Face. To compute the thresholds, we will select the 475 imposters for each image in the group and compute a threshold that achieves the policy FMR on the imposter distribution. The Policy FMR is usually set by policymakers who measure the risk of the system. We will use a Policy FMR of $10^{-3}$, as recommended by the European Border Guard Agency Frontex [45].

For the Decision Tree, we set the threshold according to the comparisons between images rather than assigning groups to specific faces. Each threshold is set based on each leaf that the pairs of images fall into. We will adjust the number of leaves by setting a high number for the depth of the tree and then selecting the N most relevant leaves. This will allow us only to select the most informative comparisons.

### C. EVALUATION METRICS

There is currently no consensus on which is the best metric to measure differential outcomes in facial recognition. Nonetheless, most works try to focus on achieving equal error rates across demographics. Given a set of groups $G$ (e.g., race, gender), we will measure the differential outcome of each strategy using the Skewed Error Ratio (SER) [53]:

$$ SER = \frac{\max\{FMR_g : \forall g \in G\}}{\min\{FMR_g : \forall g \in G\}} $$

This is a pessimistic metric, as it focuses on the worst-case scenario. We would also like to measure the dispersion of the error rates. In facial verification, we often think of errors in ratios (e.g., an FMR of 0.01 means falsely accepting in
FIGURE 6. Distribution of FMR for race and gender groups calculated using WDI. Reported FMR is the average performance of the folds using 5-fold cross validation. Red line is the desired FMR for the system.

### V. RESULTS

#### A. IMPORTANCE OF NOT MIXING SCENARIOS

It is important to select the thresholds and report group metrics using the same scenario, and not to use CDI for one and WDI for the other. As seen in Fig. 5, if the thresholds are trained using CDI and reported on WDI, then all demographic groups will be above the Policy FMR. This happens because, as seen by several studies [5], [7], [14], choosing similar subjects increases the similarity of the imposters’ distribution. On the other hand, if one wishes to report metrics using CDI (since this scenario is the most similar to an operational setting) but chooses the demographic thresholds based on WDI (which is the most common approach) then the results for all demographic groups fall almost an order of magnitude below the Policy FMR. This happens because the thresholds were chosen with a distribution that had harder examples than the ones that it is being tested on.

#### B. MITIGATING DIFFERENTIAL OUTCOMES

When training and testing the methods using WDI (Fig. 6) we see that, while the strategy of clustering facial embeddings does not give a significant improvement, clustering soft-demographic attributes reduces the gap between the error rates and gets them closer to the desired FMR. As seen in Table 1, these results are even better than using thresholds based on gender or race by themselves. Using a global threshold, the worst demographic group performs over 11 times worse than the best one. Compared to the global threshold, all methods show an improvement on the mitigation of differential outcomes, with the exception of the decision tree-based threshold. Using the non-sensitive soft-biometric clusters, this disparity is reduced to less than 3 times. Even if a difference exists, the result is better than the $6 \times$ and $4 \times$ disparity produced by the gender and race thresholds respectively. This means that differential outcomes on demographic groups can be mitigated without explicitly using said demographics. Decision Tree-based strategies lower the FMR, but it lowers it so much that it affects the FNMR metrics, making it an undesirable result.
FIGURE 7. Distribution of FMR for race and gender groups calculated using CDI. Reported FMR is the average performance of the folds using 5-fold cross validation. Red line is the desired FMR for the system.

TABLE 1. Evaluation metrics training and reporting using WDI on the BFW dataset.

| WDI Scenario                      | Differential Outcome | Global Performance |
|-----------------------------------|----------------------|--------------------|
|                                   | SER                  | MAPE               | Global FMR | Global FNMR |
| **Baseline**                      |                      |                    |            |             |
| Global Thr.                       | 11.70                | 51.48%             | 0.001003   | 0.188759    |
| **Demographic-groups Thresholds**|                      |                    |            |             |
| Gender Thr.                       | 6.11                 | 46.91%             | 0.001024   | 0.188488    |
| Race Thr.                         | 4.03                 | 40.79%             | 0.001025   | 0.186091    |
| Race+Gender Thr.                  | 1.14                 | 7.88%              | 0.001079   | 0.184953    |
| **Non-sensitive Groups Thresholds**|                     |                    |            |             |
| Embedding Clustering Thr. (K=32)  | 8.46                 | 54.20%             | 0.001105   | 0.187873    |
| Soft-Biometric Clustering Thr. (K=16)  | 2.81               | 32.60%             | 0.001034   | 0.186889    |
| Decision Tree Thr. (Leafs=32)     | 13.07                | 92.53%             | 0.000075   | 0.381199    |

TABLE 2. Evaluation metrics training and reporting using CDI on the BFW dataset.

| CDI Scenario                      | Differential Outcome | Global Performance |
|-----------------------------------|----------------------|--------------------|
|                                   | SER                  | MAPE               | Global FMR | Global FNMR |
| **Baseline**                      |                      |                    |            |             |
| Global Thr.                       | 4.94                 | 36.10%             | 0.001002   | 0.154333    |
| **Demographic-groups Thresholds**|                      |                    |            |             |
| Demographic Thr. (gender)         | 3.44                 | 32.35%             | 0.001008   | 0.154078    |
| Demographic Thr. (race)           | 2.18                 | 24.27%             | 0.001007   | 0.153566    |
| Demographic Thr. (race+gender)    | 1.04                 | 1.31%              | 0.001013   | 0.153307    |
| **Non-sensitive Groups Thresholds**|                     |                    |            |             |
| Embedding Clustering Thr. (K=32)  | 4.06                 | 29.95%             | 0.000967   | 0.156431    |
| Soft-Biometric Clustering Thr. (K=16)  | 2.65               | 21.20%             | 0.000949   | 0.155873    |
| Decision Tree Thr. (Leafs=32)     | 1.48                 | 12.30%             | 0.001016   | 0.164454    |

The CDI scenario is the one most commonly used currently in operational settings. In this scenario, the disparity of the global threshold is lower than when using WDI (5 × vs. 11 ×). In this scenario all methods show at least a slight advantage over the global threshold (Table 2). The soft-biometric clusters reduce the disparity to almost half of the global threshold. This strategy is still better than using a gender threshold, but in this scenario, it is more comparable to a threshold based on race. After the threshold based on race and gender, the best strategy is to use a Decision Tree-based threshold. We see that the disparity is reduced considerably for different hyperparameters (see Fig. 7). Nonetheless, this is done at the expense of having the worst global FNMR.

VI. DISCUSSION

In our work, the global threshold is presented as a baseline in the presented results since it is the standard approach in FV. The demographic thresholds are presented as the current state-of-the-art in bias mitigation using variable thresholds. While there have been a lot of other state-of-the-art works that
attempt to mitigate bias using different strategies (e.g., pre-processing or in-processing), these were not included as not to divert the focus away from the variable-thresholds strategies. Therefore, the performance non-sensitive group thresholds will be analyzed with respect to the global thresholds and demographic thresholds.

We saw that differential outcomes can be mitigated using groups that do not use the demographic groups explicitly, sometimes even surpassing the thresholds based on demographic groups. In both the CDI and WDI scenarios, the non-sensitive approach proves to be twice as effective at reducing differential outcomes in intersectional groups. The decision tree-based approach could mitigate a lot of the bias in the CDI scenario but performed poorly on the WDI scenario. The reason for this could be due to the decision tree overfitting the data. This is supported by the fact that the literature presents WDI as more similar between demographic groups. Another limitation for using the proposed strategies in an operational context is the additional time required to run the soft-biometric attribute classifier. Our work directly uses the outputs proposed in the MAAD-Face dataset. Future work could explore the efficiency and scalability of including these models into a classification pipeline.

The proposed method shows a path to mitigating observed differential outcomes for demographic groups, “bias”, by defining variable thresholds without asking for or explicitly predicting demographic groups. This is a new approach for how to apply variable thresholds. Furthermore, this method can be applied to any black-box facial recognition system, requiring minimal training to achieve results that effectively mitigate bias.

REFERENCES

[1] Y. Qiu, V. Albiero, M. C. King, and K. W. Bowyer, “Does face recognition error echo gender classification error?” in Proc. IEEE Int. Joint Conf. Biometrics, Aug. 2021, pp. 1–8, doi: 10.1109/BCB52358.2021.9484346.
[2] J. Buolamwini and T. Gebru, “Gender shades: Intersectional accuracy disparities in commercial gender classification,” in Proc. Conf. Fairness, Accountability Transparency, 2018, pp. 77–91. [Online]. Available: http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf
[3] V. Muthukumar, T. Pedapati, N. Ratha, P. Sattigeri, C.-W. Wu, B. Kingsbury, A. Kumar, S. Thomas, A. Mojsilovic, and K. R. Varshney, “Understanding unequal gender classification accuracy from face images,” 2018, arXiv:1812.00099.
[4] M. Ngan, P. J. Grother, and M. Ngan, Face Recognition Vendor Test (FRVT) Performing Automated Gender Classification Algorithms. Gaithersburg, MD, USA: National Institute of Standards and Technology, 2015, doi: 10.6028/NIST.IR.8052.
[5] P. Grother, M. Ngan, and K. Hanaoka, Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. Gaithersburg, MD, USA: NIST, 2019, doi: 10.6028/NIST.IR.8280.
[6] V. Albiero, K. S. Krishnapriya, K. Vangara, K. Zhang, M. C. King, and K. W. Bowyer. “Analysis of gender inequality in face recognition accuracy,” in Proc. IEEE Winter Appl. Comput. Vis. Workshops (WACVW), Mar. 2020, pp. 81–89, doi: 10.1109/wacwv50321.2020.9096947.
[7] J. J. Howard, Y. B. Sirotin, and A. R. Vemury, “The effect of broad and specific demographic homogeneity on the impostor distributions and false match rates in face recognition algorithm performance,” in Proc. IEEE Int’l Conf. Biometrics Theory, Appl. Syst. (BTAS), Sep. 2019, pp. 1–8, doi: 10.1109/btas46853.2019.9186002.
[8] K. 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 Trans. Technol. Soc., vol. 1, no. 1, pp. 8–20, Oct. 2020, doi: 10.1109/itts.2020.2974996.
P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch, “E. Villalobos, ‘’Fair-face verification by using non-sensitive soft-biometric S. Yucer, F. Tektas, N. A. Moubayed, and T. P. Breckon, ‘’Mea-
J. J. Howard and D. Etter, ‘’The effect of ethnicity, gender, eye
differences in face recognition accuracy explained by hairstyles, makeup,
and facial morphology.’’ IEEE Trans. Inf. Forensics Security, vol. 17, no. 6, pp. 1789–1801, Dec. 2012, doi: 10.1109/TIFS.2012.2214212.
V. Alibiero, K. Zhang, and K. W. Bowyer, ‘’How does gender bal-
ace in training data affect face recognition accuracy?’’ in Proc. IEEE Int. Joint Conf. Biometrics (IJCB), Sep. 2020, pp. 1–10, doi: 10.1109/IJCB48548.2020.9304924.
X. Xu, Y. Huang, P. Shen, S. Li, L. Li, F. Huang, Y. Li, and Z. Cui, ‘’Consistent instance false positive improves fairness in face recognition,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 578–586, doi: 10.1109/CVPR46437.2021.00064.
A. Morales, J. Fierrez, R. Vera-Rodriguez, and R. Tolosana, ‘’Sensi-
tiveNets: Learning agnostic representations with application to face images,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 6, pp. 2158–2164, Jun. 2021, doi: 10.1109/TPAMI.2020.3015420.
P. Terhorst, J. N. Kolf, N. Damer, F. Kirchbuechner, and A. Kuijper, ‘’Post-
comparison mitigation of demographic bias in face recognition using fair score normalization,’’ Pattern Recognit. Lett., vol. 140, pp. 332–338, Dec. 2020, doi: 10.1016/j.patrec.2020.11.007.
P. Terhorst, M. L. Tran, N. Damer, F. Kirchbuechner, and A. Kuijper, ‘’Comparison-level mitigation of ethnic bias in face recognition,’’ in Proc. 8th Int. Workshop Biometrics Forensics (IWBFW), Apr. 2020, pp. 1–6, doi: 10.1109/iwbfw49977.2020.907956.
T. Sixta, J. C. J. Junior, P. Buch-Cardona, E. Vazquez, and S. Escalera, ‘’Facefair challenge at ECCV 2020: Analyzing bias in face recognition,’’ in Proc. Eur. Conf. Comput. Vis. Springer, 2020, pp. 463–481, doi: 10.1007/978-3-030-65414-6_32.
I. Serna, A. Morales, J. Fierrez, M. Cebrian, N. Obradovich, and I. Rahwan, ‘’SensitiveLoss: Improving accuracy and fairness of face representations with discrimination-aware deep learning,’’ 2020, arXiv:2004.11246.
M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang, ‘’Racial faces in the wild: Reducing racial bias by information maximization adaptation network,’’ in Proc. IEEE/CVF Int. Conf. Comput. Vis. Conf., Mar. 2019, pp. 692–702, doi: 10.1109/CVPR.2019.00078.
P. Grother, M. Ngan, and K. Hanaoka, ‘’Ongoing face recognition vendor test (FRVT) part 1: Verification,’’ Nat. Inst. Standards Techn-
nol., Gaithersburg, MD, USA, Tech. Rep., 2019. [Online]. Available: https://www.nist.gov/document/frvtreport20190731pdf.
J. Deng, J. Guo, N. Xue, and S. Zafeiriou, ‘’ARCFace: Additive angular margin loss for deep face recognition,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Oct. 2019, pp. 4600–4609, doi: 10.1109/CVPR.2019.00482.
A. K. Jain, A. Ross, and S. Prabhakar, ‘’An introduction to biometric recognition,’’ IEEE Trans. Circuits Syst. Video Technol., vol. 14, no. 1, pp. 4–20, Jan. 2004.
Frontex. (2015). Best Practice Technical Guidelines for Automated Border Control (ABC) Systems. [Online]. Available: https://frontex.europa.eu/assets/Publications/Research/Best_Practice_Technical_Guidelines_ABC.pdf.
[46] A. J. O’Toole, P. J. Phillips, X. An, and J. Dunlop, “Demographic effects on estimates of automatic face recognition performance,” Image Vis. Comput., vol. 30, no. 3, pp. 169–176, Mar. 2012, doi: 10.1016/j.imavis.2011.12.007.

[47] A. Das, A. Dantcheva, and F. Bremond, “Mitigating bias in gender, age and ethnicity classification: A multi-task convolution neural network approach,” in Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops, Sep. 2018, doi: 10.1007/978-3-030-11009-3_35.

[48] G. Ozbulak, Y. Aytar, and H. K. Ekenel, “How transferable are CNN-based features for age and gender classification?” in Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG), Sep. 2016, pp. 1–6, doi: 10.1109/biosig.2016.7736925.

[49] P. Terhorst, M. Huber, J. N. Kolf, I. Zelch, N. Damer, F. Kirchbuchner, and A. Kuijper, “Reliable age and gender estimation from face images: Stating the confidence of model predictions,” in Proc. Int. Conf. Biometrics Theory, Appl. Syst., Tampa, FL, USA, Sep. 2019, pp. 1–8, doi: 10.1109/BTAS46853.2019.9185975.

[50] J. P. Robinson, C. Qin, Y. Henon, S. Timoner, and Y. Fu, “Balancing biases and preserving privacy on balanced faces in the wild,” 2021, arXiv:2103.09118.

[51] P. Terhorst, D. Fahrmann, N. Damer, F. Kirchbuchner, and A. Kuijper, “Beyond identity: What information is stored in biometric face templates?” in Proc. IEEE Int. Joint Conf. Biometrics (IJCB), Sep. 2020, pp. 1–10, doi: 10.1109/IJCB48548.2020.9304874.

[52] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, “VGGFace2: A dataset for recognising faces across pose and age,” in Proc. Face Gesture Recognit., May 2018, pp. 67–74, doi: 10.1109/fgr.2018.00020.

[53] M. Wang and W. Deng, “Mitigating bias in face recognition using skewness-aware reinforcement learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 9322–9331, doi: 10.1109/CVPR42600.2020.00934.

[54] P. Grother, “Demographic differentials in face recognition algorithms,” in Virtual Events Series–Demo-Graphic Fairness in Biometric Systems. 2021.

[55] Z. Khan and Y. Fu, “One label, one billion faces: Usage and consistency of racial categories in computer vision,” in Proc. ACM Conf. Fairness, Accountability, Transparency, Mar. 2021, pp. 587–597, doi: 10.1145/3442188.3445920.

---

ESTEBAN VILLALOBOS received the M.Sc. degree in computer science from the Universidad Católica de Chile, in 2021. He is currently pursuing the Ph.D. degree with the Institute de Recherche Informatique de Tolouse (IRIT), Université Paul Sabatier Tolouse III, France. His research interests include biometrics, fairness, computer vision, learning analytics, and technology enhanced learning.

DOMINGO MERY (Member, IEEE) received the M.Sc. degree in electrical engineering from the Technical University of Karlsruhe, in 1992, and the Ph.D. degree (Hons.) from the Technical University of Berlin, in 2000. In 2001, he worked as an Associate Researcher with the Department of Computer Engineering, Universidad de Santiago, Chile. In 2014, he was a Visiting Professor with the University of Notre Dame. He is currently a Full Professor with the Department of Computer Science, Pontificia Universidad Católica de Chile, where he served as the Chair, from 2005 to 2009, and the Director of Research and Innovation with the School of Engineering, from 2015 to 2018. His research interests include image processing, computer vision applications, and biometrics. He received the Ron Halmshaw Award, in 2005, 2012, and 2017, and the John Green Award, in 2013, from the British Institute of Non-Destructive Testing, which was established to recognize the best papers published in the Insight Journal on Industrial Radiography. He received the Best Paper Award at the International Workshop on Biometrics in conjunction with the European Conference on Computer Vision (ECCV 2014). He is currently serving as an Associate Editor for IEEE TRANSACTION ON INFORMATION, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, and IEEE TRANSACTIONS ON BIOMETRICS, BEHAVIOR, AND IDENTITY SCIENCE.

KEVIN BOWYER (Life Fellow, IEEE) is currently the Schubmehl-Prein Family Professor in computer science and engineering with the University of Notre Dame. He has served as an EIC for the IEEE TRANSACTIONS ON BIOMETRICS, BEHAVIOR, AND IDENTITY SCIENCE and the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. He was elected as a fellow of the American Academy for the Advancement of Science with the citation “for distinguished contributions to the field of computer vision and pattern recognition, biometrics, object recognition and data science.” He received the IEEE Biometrics Council Meritorious Service Award and the Technical Achievement Award from the IEEE Computer Society with the citation “for pioneering contributions to the science and engineering of biometrics.” He was elected as a fellow of the IEEE with the citation “for contributions to algorithms for recognizing objects in images,” and a fellow of the IAPR with the citation “for contributions to computer vision, pattern recognition and biometrics.” He received the Award of Excellence from the Society for Technical Communication, Philadelphia Chapter, for the paper “Face recognition technology and the security versus privacy tradeoff,” published in IEEE Technology and Society.