PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA

FAIR FACE VERIFICATION BY USING NON-SENSITIVE SOFT-BIOMETRIC ATTRIBUTES

ESTEBAN VILLALOBOS

Thesis submitted to the Office of Research and Graduate Studies
in partial fulfillment of the requirements for the degree of
Master of Science in Engineering

Advisor:
DOMINGO MERY, PH. D.

Santiago de Chile, August 2021

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA

FAIR FACE VERIFICATION BY USING NON-SENSITIVE SOFT-BIOMETRIC ATTRIBUTES

ESTEBAN VILLALOBOS

Members of the Committee:
DOMINGO MERY, PH. D.
DENIS PARRA, PH. D.
KEVIN W. BOWYER, PH. D.
CARLOS BONILLA, PH. D.

Thesis submitted to the Office of Research and Graduate Studies
in partial fulfillment of the requirements for the degree of
Master of Science in Engineering

Santiago de Chile, August 2021

Gratefully to my parents, brother
and friends

ACKNOWLEDGEMENTS

This work was supported in part by Fondecyt grant 1191131 from CONICYT–Chile.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

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 MAADFace 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.

Keywords: Facial Recognition, Fairness, Differential Outcomes, FMR.

RESUMEN

Los algoritmos de reconocimiento facial han demostrado tener diferencias en los resultados entre los distintos grupos demográficos. Incluso cuando se establece un umbral global para obtener una tasa de falsas coincidencias (FMR) específica para todo el sistema, algunos grupos demográficos pueden obtener resultados significativamente peores que los indicados. Para mitigar esto, algunos autores han propuesto utilizar umbrales específicos por grupos demográficos. Sin embargo, esto es poco práctico en un entorno operativo, ya que requeriría que los usuarios informaran de su grupo demográfico o lo predijeran en el sistema. Ambas opciones son controversiales, debido a que el dato del grupo demográfico es sensible. Además, en el caso de utilizar umbrales basados en un grupo racial, requiere enumerar exhaustivamente todas las razas posibles para el sistema. Demostramos que se puede conseguir un efecto de mitigación similar utilizando atributos biométricos blandos predecibles no sensibles. Se trata de atributos basados en la apariencia de los sujetos que no dependen de cómo se identifican los usuarios (como el peinado, los accesorios y la geometría facial). Utilizamos 38 atributos binarios no demográficos del conjunto de datos MAADFace. Presentamos los resultados en el conjunto de datos BFW, que tiene un número equilibrado de identidades por raza y género. Comparamos las estrategias basadas en la agrupación y en los árboles de decisión como formas de seleccionar estos umbrales. Demostramos que estas estrategias pueden reducir los resultados diferenciales en los grupos interseccionales con el doble de eficacia que el uso de umbrales específicos de género y, en algunos casos, también son mejores que el uso de umbrales específicos de raza.

Palabras Claves: Reconocimiento Facial, Equidad, Resultados Diferenciales, Taza de Falsos Positivos.

1. INTRODUCTION

Recent studies have pointed to potential demographic biases in facial analysis (Buolamwini & Gebru, 2018; Muthukumar et al., 2018; Ngan, Grother, & Ngan, 2015; Qiu, Albiero, King, & Bowyer, 2021) and facial recognition (Albiero, KS, et al., 2020; Grother, Ngan, & Hanaoka, 2019; Howard, Sirotin, & Vemury, 2019; Krishnapriya, Albiero, Vangara, King, & Bowyer, 2020; Qiu et al., 2021). 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" (Committee, 2020). The central concern is typically that different demographic groups experience different false match rates. In facial verification, a false match occurs when the similarity between images of two different persons 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 (Anderson, July 10, 2020; Li, December 29, 2020).

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 one that results in only 1 in 10,000 impostor image pairs being above threshold. 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 (Grother et al., 2019; Howard et al., 2019; Robinson et al., 2020). The National Institute of Standards and Technology (NIST) showed in a recent Face Recognition Vendor Test (FRVT) (Grother et al., 2019) that, for many algorithms, the False Match Rate across demographic groups can vary by factors of 10 or 100.

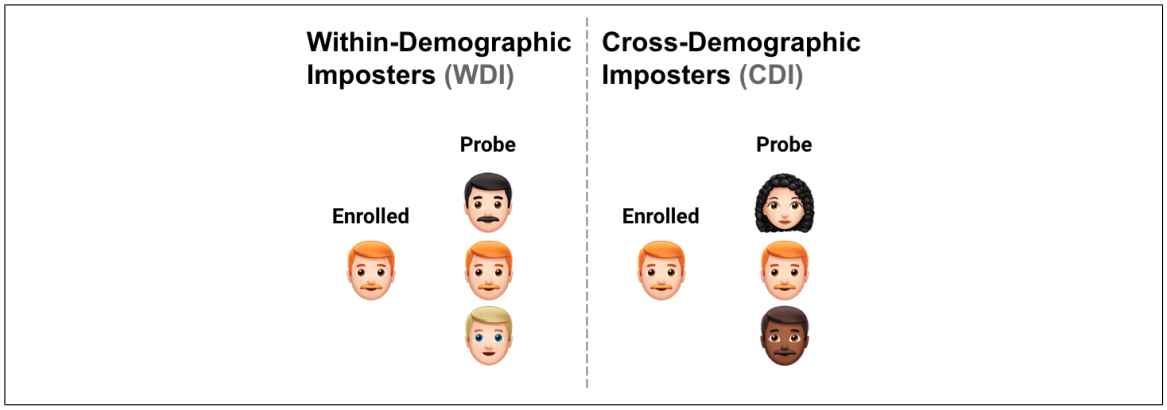

Figure 1.1. Scenarios when addressing demographic performance in facial verification. The Within-Demographic Imposters (WDI) Scenario consists on restricting comparisons to the same demographic group (in this work race and gender). The Cross-Demographic Imposters (CDI) Scenario allows imposter images to be from different groups.

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 (Cavazos, Phillips, Castillo, & O'Toole, 2020; Krishnapriya, Vangara, King, Albiero, & Bowyer, 2019; Robinson et al., 2020). This would ensure that each demographic respects the Policy FMR. Nonetheless, there are still some problems and unanswered questions concerning this approach. First, there is no consensus on how to choose the imposter images for a given demographic. To select the optimal thresholds and to evaluate the performance of the methods, one has to set one of two scenarios (see Fig. 1.1). 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 (Albiero, KS, et al., 2020; Krishnapriya et al., 2020, 2019). 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 (Cavazos et al., 2020; Grother et al., 2019; Howard et al., 2019), using WDI leads to overall higher impostor scores, because lower-similarity impostor pairs are not included in the distribution. Therefore, a higher threshold is required to ensure demographic groups

[1]Also known as demographic yoking (Cavazos et al., 2020).

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 but it has been been less explored to measure demographic performance. This means that a global threshold may be computed using CDI, but an analysis of bias may be performed using WDI. When doing this, it is plausible that all demographic groups will fall 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.

To select demographic thresholds, there is also the issue of how to assign the demographic label to an identity. Grother et al. (2019) 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 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 (Hamidi, Scheuerman, & Branham, 2018). 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 (Wu, Protopapas, Yang, & Michalatos, 2020). Moreover, Qiu et al. (2021) found that false classifications of gender correlate with a false rejection of a true matching.

There is also no consensus on how many demographic groups should be considered. Most studies considered gender (Albiero, KS, et al., 2020), race (Krishnapriya et al., 2020), and age (Michalski, Yiu, & Malec, 2018). 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,

[2]Also known as zero-effort imposters (Grother et al., 2019).

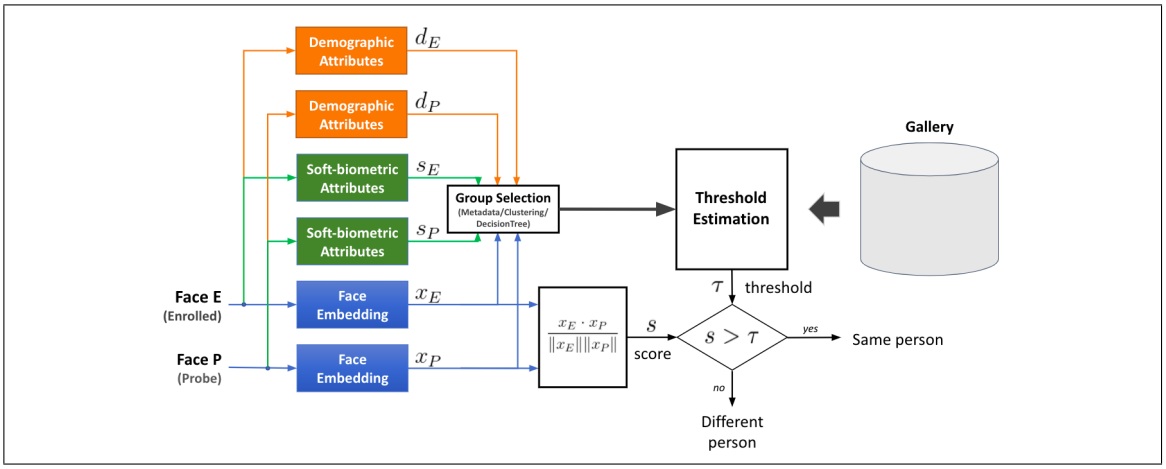

Figure 1.2. To compute group-specific thresholds, we compare the strategies of defining groups based on Demographic Attributes (orange), Non-sensitive Soft-biometric Attributes (green), and Facial Embeddings (blue). These groups will be defined based on the metadata (in the case of the demographic attributes), 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.

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 (Bruveris, Gietema, Mortazavian, & Mahadevan, 2020; Grother et al., 2019). All-in-all, there are still some gaps in the possibility of using demographic-specific thresholds.

In our work, we analyze in depth the problem of how to choose the threshold in a fair face verification context. We compare the strategy of selecting a global threshold and demographic-specific thresholds with others that do not explicitly depend on demographic data (Fig. 1.2). Since automatically predicting a demographic group may be undesirable, we will test by (i) selecting a variable threshold based on the clustering of the facial embedding features (ii) based on non-sensitive soft-biometric features (such as hairstyle or accessories) by clustering a binary soft-biometric attribute vector, and (iii) by building a strategy based on decision trees to select the most informative attributes.

The main contributions of this paper turned into thesis are four-fold:

- To show the effect of using demographic-thresholds based on a single demographic (i.e., gender or race) and intersectional groups (i.e., race+gender) when testing on intersectional groups
- To compare how different operational scenarios (WDI and CDI) affect the training of demographic thresholds and their reported effectiveness to mitigate differential outcomes
- To explore automatic group-based thresholds 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

2. RELATED WORK

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

2.1. Fairness in Machine Learning

There has been an increase in studies on bias in machine learning, such as hiring, recommendations, and facial analysis (Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2019). Facial recognition studies have focused on group fairness, which can be defined as 'treating similar groups similarly' (Mehrabi et al., 2019). Other areas of study have focused on individual fairness ('treating similar individuals similarly'), but this involves the non-trivial task of defining a similarity metric between individuals to measure fairness. In terms of group fairness criteria, most definitions are properties of a sensitive attribute A (e.g., gender), the target variable Y (e.g., whether a loan applicant will pay back), and a classifier R (e.g., credit score) (Barocas, Hardt, & Narayanan, 2019). Most criteria then fall into one of the following categories: independence ($R \perp A$)[1], separation ($R \perp A|Y$)[2] , and sufficiency ($Y \perp A|R$).

Facial recognition focuses on equalizing error rates. This falls in the category of separation, as the goal is to make the scores independent of the demographic group given the true label (match/non-match). Equalizing both false match and false non-match error rates across groups is rarely achievable in practice, and so proposed solutions focus on the error rate that is considered more important for the scenario. This is most often the false match rate, since a false match tends to be more undesirable for the subjects.

In order to meet these criterion, one can make changes before, during or after the training process of an algorithm. These are classified, respectively, as follows: a) preprocessing

[1] $A \perp B$ means A is independent of B
[2] $A \perp B|C$ means A is independent of B given C

(e.g., ensuring balanced datasets), b) in-processing (e.g., include bias regularization terms in the training process), or c) post-processing (e.g., normalizing the scores or varying the thresholds) (Mehrabi et al., 2019).

Both (Hardt, Price, & Srebro, 2016) and (Corbett-Davies, Pierson, Feller, Goel, & Huq, 2017) present post-processing methods to achieve equal error rates in classification problems by using demographic-specific thresholds. Unfortunately, these results are not directly applicable to facial verification. These approaches assume that a single input has a sensitive attribute (e.g., race) associated with it. In facial verification, both the enrolled and probe images each have their own sensitive attribute. While this is not a problem in a WDI Scenario, as both images are from the same group by design, it does pose a problem when considering a CDI Scenario. If the two images are from different demographic groups, it is unclear how one should select a threshold. One could argue that if both images are from different demographic groups, then it should be assumed to be an imposter, but this ignores the fact that some demographic attributes can change in time (e.g., age and gender) and, if the group is being predicted using an algorithm, the mismatch could be due to a classification error.

One can deal with the previous problems by only focusing on the enrolled image's attributes without considering the probe image's demographic group. Then, as Hardt et al. (2016) and Corbett-Davies et al. (2017) propose, we can use a different threshold to perform decisions in each demographic group. To further this idea, our work proposes the non-sensitive attributes to replace the use of demographic groups. In our work, the clustering-based solutions only look at the group of the enrolled image. This mitigates the effect of attributes that change over time, and, in the case of a classification error, it does not immediately classify the pair as an imposter.

2.2. Fairness in Facial Recognition

Most studies focused on mitigating biases in facial recognition can be classified as preprocessing, such as training on balanced datasets (Albiero, KS, et al., 2020; Albiero, Zhang, & Bowyer, 2020; Klare, Burge, Klontz, Bruegge, & Jain, 2012; Zhang & Deng, 2020), or in-processing, such as using adversarial networks (Gong, Liu, & Jain, 2020; Morales, Fierrez, Vera-Rodriguez, & Tolosana, 2020), reinforcement learning (Wang & Deng, 2020), or adding error rate penalties in the loss function (Xu et al., 2021). Nonetheless, there have also been a few studies on post-processing for bias mitigation (Terhörst, Kolf, Damer, Kirchbuchner, & Kuijper, 2020; Terhörst, Tran, Damer, Kirchbuchner, & Kuijper, 2020). While preprocessing and in-processing are the most common approaches, they also require many resources to acquire facial data or computational resources in re-training the networks. Furthermore, even if one dedicated time and resources to ensure balanced datasets, Albiero, Zhang, and Bowyer (2020) 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. Howard et al. (2019) 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 the differential performance (Gong et al., 2020; Krishnapriya et al., 2019; Serna et al., 2020; Sixta, Junior, Buch-Cardona, Vazquez, & Escalera, 2020; Wang, Deng, Hu, Tao, & Huang, 2019). Consequently, the comparison of ROC Curves and AUC-ROC became very popular metrics to use (Sixta et al., 2020). These studies tend to report demographic ROC curves that only use same-group comparisons; therefore, they fall into the WDI Scenario. As for differential outcomes, some studies have used differentials of FNMR at a given FMR (Terhörst, Kolf, et al., 2020; Terhörst et al., 2020), while others

have focused directly on differences in FMR (Cavazos et al., 2020; Grother et al., 2019; Howard et al., 2019; Robinson et al., 2020; Xu et al., 2021).

When using a ROC curve (or AUC) to compare across demographic groups, something to consider is that different demographics typically achieve a particular FMR at different thresholds (Cavazos et al., 2020; Krishnapriya et al., 2019). 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 (Cavazos et al., 2020; Krishnapriya et al., 2020, 2019; Robinson et al., 2020). Krishnapriya et al. (2019) shows 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 (Robinson et al., 2020), one of the few studies that explore differential performance and differential outcomes using CDI, query images from any demographic group are allowed.

Cavazos et al. (2020) says that threshold setting and controlling imposters are scenario-modeling factors relating to race bias that are under "control" of the user. They state that "it is clear that a uniform threshold is not adequate or equitable when the underlying sub-population distributions differ" and therefore using group-specific thresholds is more adequate. Even though they mention the issue of variable thresholds and enforcing imposter restrictions, they do not explore under which scenario these demographic thresholds should be set.

Studies exploring the use of demographic-specific thresholds have suggested choosing a threshold based on some demographic groups (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). Grother, Ngan, and Hanaoka (2018) showed the impact on different demographics 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. In our work, we not only compare the effect of controlling the imposters for the demographic thresholds but explore how setting thresholds according to non-sensitive attributes affects these intersectional groups.

2.3. Effects of Soft-Biometric Attributes on FV

Dantcheva, Elia, and Ross (2015) 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. While demographic attributes are the most commonly studied (Drozdowski, Rathgeb, Dantcheva, Damer, & Busch, 2020), there are also studies on subject-specific attributes (e.g., hair style, expression and accessories) (Albiero & Bowyer, 2020; Terhörst et al., 2021) and environmental (e.g., illumination and resolution) (Howard & Etter, 2013). Terhörst et al. (2021) 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 certain attributes can vary significantly for different decision thresholds.

To the best of our knowledge, our work presented here is the first work that uses 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 (Gong et al., 2020; Krishnapriya et al., 2019; Serna et al., 2020; Sixta et al., 2020; Wang et al., 2019), and explicitly (Cavazos et al., 2020; Krishnapriya et al., 2019; Robinson et al., 2020).

3. THRESHOLD STRATEGIES IN FV

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 verification problem are (Jain, Ross, & Prabhakar, 2004):

- 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 τ and similarity score s, we choose D_1 if $s > \tau$ and D_0 otherwise.

This allows to define the error rates as follows

- $FMR = \mathbb{P}(D_1|H_0)$
- $FNMR = \mathbb{P}(D_0|H_1)$

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

$$D_{\text{global thr.}} := s(X_E, X_P) > \tau_{\text{global}} \tag{3.1}$$

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

$$D_{\text{variable thr.}} := s(X_E, X_P) > \tau_{f(X_E,X_P)} \tag{3.2}$$

where τ_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 (Cavazos et al., 2020; Krishnapriya et al., 2019; Robinson et al., 2020). 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. 1.1 to see the differences between the WDI and CDI Scenario, Fig. 1.2 as general overview of the five strategies and Fig. 4.1 as a guide for the features used in each strategy.

3.1. Fixed Global Threshold

The global threshold will be chosen as the one that ensures a given FMR in the training set. While some authors suggest that this threshold should be set using WDI (Cavazos et al., 2020; Grother et al., 2019), it is usually computed using CDI (O'Toole, Phillips, An, & Dunlop, 2012). The Fixed Global Threshold strategy is the standard approach in facial verification, so we will use it as a baseline to compare against.

3.2. 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.

3.3. 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 (Das, Dantcheva, & Bremond, 2018; Morales et al., 2020; Ozbulak, Aytar, & Ekenel, 2016). Terhörst, Fährmann, Damer, Kirchbuchner, and Kuijper (2020) 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[1], 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. Later, when testing, we will use the cluster of the probe image to select the threshold.

3.4. 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. Terhörst et al. (2021) and Albiero and Bowyer (2020) showed that many non-demographic soft-biometric attributes strongly affect recognition performance. The MAADFace 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

[1]Clustering was performed using the MiniBatchKMeans implementation of scikit-learn version 0.24.2 with default parameters.

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% (Terhörst et al., 2019a) 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[2]. 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.

3.5. 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 (Cavazos et al., 2020; Howard et al., 2019; Terhörst et al., 2019a). To find attributes that might convey a lot of information on false matches, we will use an information-based decision tree model as suggested by (Howard & Etter, 2013) and (Howard et al., 2019).

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

$$E(Y) = -\sum_i p_i \log_2(p_i) \tag{3.3}$$

where $Y := D_{\text{global}}|H_0$ is a random variable representing the probability of a false match for a global threshold, and p_i is the probability of a pair of images being either $i \in$

[2]Experiments were also performed reduction the dimensionality of the embeddings before doing K-Means (Appendix B). In general, we did not see a major improvement on bias mitigation by reducing the dimensionality.

$\{$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

$$IG(Y, X) = E(Y) - E(Y|X) \tag{3.4}$$

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., $X =$(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.

4. EXPERIMENTAL METHODOLOGY

4.1. Datasets

Our work is based on the Balanced Faces in the Wild (BFW) (Robinson et al., 2020; Robinson, Qin, Henon, Timoner, & Fu, 2021) and MAAD-Face (Terhörst et al., 2019a, 2019b; Terhörst, Kolf, Damer, Kirchbuchner, & Kuijper, 2020) datasets. Both are datasets based on VGGFace2 (Cao, Shen, Xie, Parkhi, & Zisserman, 2018). 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 3.4. 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% (Terhörst et al., 2019a).

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)[1] and another 4 images that were not present in MAAD-Face[2]. 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.

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

[2] 'n009142/0501_01.jpg', 'n001555/0307_02.jpg', 'n009142/0501_01.jpg', 'n001555/0307_02.jpg'

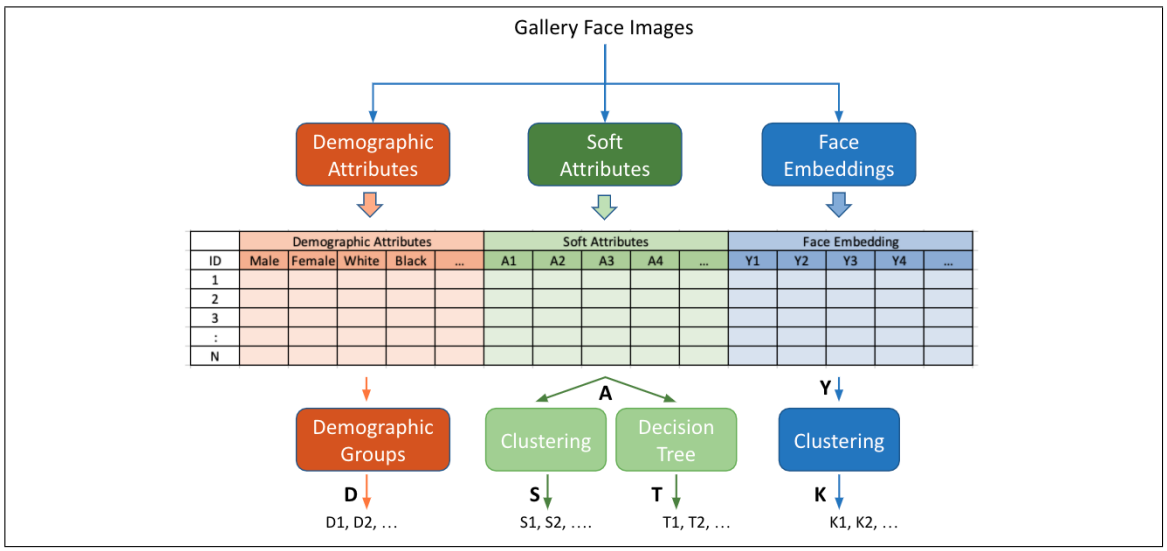

Figure 4.1. For training, the groups are created based on Demographic Attributes (D), Soft-biometric Attributes (A), and Face Embeddings (Y). On this work, the Demographic Attributes are based on the ground-truth labels of BFW Robinson et al. (2020). We will use labels for race and gender. We use 39 Soft-biometric Attributes come from MAAD-Face Terhörst et al. (2019a) (e.g., 'is bald', 'has a beard'), these attributes were predicted using a Massive Attribute Classifier. The Facial Embeddings are 512-dimensional vectors computed using ArcFace Deng et al. (2019).

4.2. Training process

The training process consisted of two steps: defining groups (Fig. 4.1) and computing thresholds for each group (Fig. 4.2). For all methods, we computed the facial embeddings using ArcFace (Resnet-101) (Deng et al., 2019), which provides a 512-dimensional vector for each facial image[3]. The Demographic Groups are created based on the metadata provided on BFW. We will use all possible combinations of race and gender. For 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

[3]We also present results using SphereFace (Liu et al., 2017) in Appendix A.

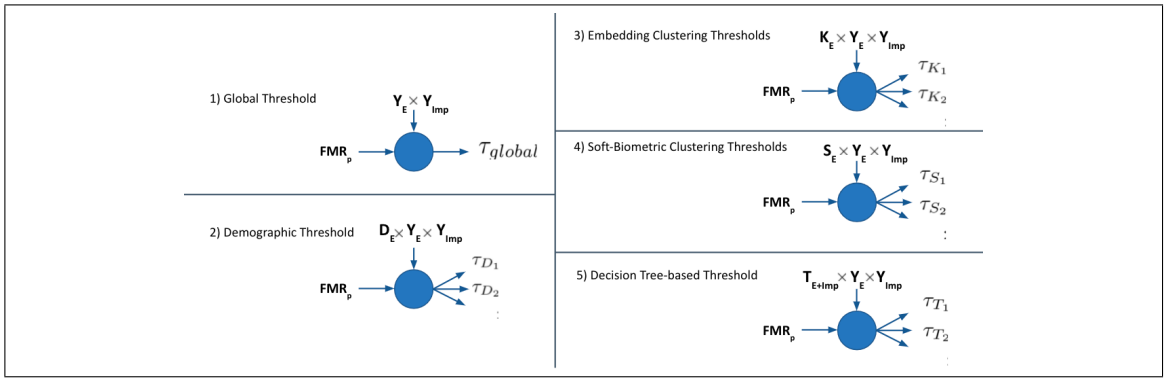

Figure 4.2. Diagram showing information required to compute each threshold in the training phase. All methods require a Policy FMR (FMR_p), the features of the enrolled images Y_E and the corresponding features of the set of imposters for each enrolled image Y_{imp}. When training on a WDI Scenario, these imposters would be from the same demographic group as the enrolled image. On the other hand, when training on a CDI Scenario there is no restriction on the demographic group of the imposters. Strategies (2), (3) and (4) receive the demographic group (D_E), facial embedding cluster (K_E) and soft-demographic cluster (S_E) respectively from the enrolled image. The Decision Tree-based strategy receives comparisons between the soft-biometric attributes of the enrolled and imposter images (T_{E+Imp}), and produces a threshold for each leaf of the tree.

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 (Frontex, 2015).

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.

4.3. 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 $\mathcal{G}$ (e.g., race, gender), we will

measure the differential outcome of each strategy using the Skewed Error Ratio (SER) (Wang & Deng, 2020):

$$SER = \frac{\max\{FMR_g \; \forall g \in \mathcal{G}\}}{\min\{FMR_g \; \forall g \in \mathcal{G}\}} \tag{4.1}$$

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 1 every 100 people). Therefore, the SER has a very intuitive explanation of how many times is the error in the worst demographic compared to the best one. A high value means that method has a high disparity in the error rates, while a value of 1 means that all groups have the same FMR. Grother (2021) declared that the NIST would start reporting this metric on their FRVT on a recent EAB event.

Robinson et al. (2020) reported the percentage error in order to measure the deviation from the Policy FMR for the system (FMR_p):

$$\text{Percentage Error}_g = \frac{FMR_g - FMR_p}{FMR_p} \tag{4.2}$$

We will use the Mean Absolute Percentage Error (MAPE) to quantify how much the groups differ on average from the Policy FMR. Given a desired FMR_p, the MAPE of the error rates is:

$$MAPE = \frac{100}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \left| \frac{FMR_g - FMR_p}{FMR_p} \right| \tag{4.3}$$

We will use the MAPE instead of the Mean Percentage Error (MPE) for two reasons. First, we do not want the low error rate of one demographic group to cancel out the high error rate of another. Second, a high deviation to a lower value of FMRs can be detrimental to the system, as it almost always comes accompanied by an increase in FNMR. If this metric is zero, then all demographic groups achieve the desired FMR.

5. RESULTS

5.1. 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. 5.1, 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 (Cavazos et al., 2020; Grother et al., 2019; Howard et al., 2019), 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.

Figure 5.1. The problem with selecting thresholds and reporting metrics for different scenarios. Left: If thresholds are selected for WDI and metrics are reported for Scenario 2 (zero-effort imposters) then all groups are over the Policy FMR for global and demographic-specific thresholds. Right: If thresholds are selected for CDI and metrics are reported for WDI then all groups fall bellow the Policy FMR by almost an order of magnitude for global and demographic-specific thresholds.

Figure 5.2. 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.

Figure 5.3. 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 5.1. 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
DecisionTree Thr. (Leafs=32)	13.07	92.53%	0.000075	0.381199

Table 5.2. 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
DecisionTree Thr. (Leafs=32)	**1.48**	**12.30%**	0.001016	0.164454

5.2. Mitigating Differential Outcomes

When training and testing the methods using WDI (Fig. 5.2) 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 A.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.

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 ($5\times$ vs. $11\times$). In this scenario all methods show at least a slight advantage over the global threshold (Table A.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. B.2). Nonetheless, this is done at the expense of having the worst global FNMR.

6. DISCUSSION

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 them, which could lead to correlations in the comparisons made by the decision tree. However, the approach of clustering soft-biometric attributes was consistent in reducing differential outcomes in both scenarios.

This could be a great advantage in operational settings, as it reduces demographic disparities without using demographic data. To use a demographic threshold on an operational setting one either has to ask for (and trust) a self-reported demographic group or try to predict it. The latter is the most controversial, as some people could consider it a privacy violation to detect a sensitive attribute through facial features. Predicting the demographic category also forces the system developers to formalize race as a categorical variable, which by itself can become controversial. Whether controversial or not, it can be problematic as there is no clear consensus on how many races should be considered. For example, on the one hand, the MAAD-Face dataset classifies each image in VGGFace2 as White, Asian and/or Black. On the other hand, the BFW dataset uses, for the same dataset, White, Asian, Indian or Black. One can then ask what would happen with other demographic groups, such as Hispanic or a mixture of them. This problem is not limited to these two datasets. Khan and Fu (2021) notes that many datasets can have badly defined and incongruent definitions of races. The FRVT (Grother et al., 2019) offers analysis separating images by country of origin, but getting to this level of granularity becomes impractical when thinking on storing demographic thresholds. While it is important to measure the

performance of algorithms according to demographic groups, using them explicitly on an operational setting seems impractical.

To avoid the problems of using demographic thresholds explicitly, we show that differential outcomes can be mitigated using non-sensitive soft-biometric data, which can be predicted with fairly good accuracy (Terhörst, Fährmann, et al., 2020; Terhörst et al., 2019a). This also shifts the focus of using labels relating to how the subjects identify themselves (like race or gender) to attributes related to the subject's appearance (like hairstyle or accessories). While some attributes may be correlated with demographic attributes, not making this relation explicit means that there is no hard boundary between demographics (e.g., while men correlate with bald, the method does not reject the existence of bald women). This means that attributes related to gender expression, for example, will be taken into account when selecting the threshold without making assumptions on gender identity.

The proposed approaches are also robust to errors in the classification of soft-biometric attributes, since even if an attribute is wrongly predicted the image is not immediately classified as genuine or imposter. The proposed methods use the predicted soft-biometric information as a guideline on how high (or low) the threshold should be set for the comparison. The results presented in this paper, for example, use the imperfect information of the classifier. Future work could analyze the sensibility of these methods to noisy predictions.

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.

7. CONCLUSION

Our work showed the effects of using demographic-thresholds based on a single demographic (i.e., gender or race) when testing intersectional groups. In our experiments we saw that using a race-specific threshold is better at mitigating differential outcomes than a gender-specific threshold.

We compared how different operational scenarios (WDI and CDI) may affect the threshold selection process and the reporting of differential outcomes metrics. We also highlated the importance of keeping these scenarios consistent.

We explored different techniques of group-based thresholds that do not depend on sensitive information. We implemented clustering and decision tree-based strategies to define group-specific thresholds. The proposed methods can be easily integrated with any black-box model.

Finally, our work showed that non-sensitive soft-biometric attributes can be an effective tool to mitigate differential outcomes. Soft-biometric attributes have already been shown to be easily predictable from facial images and to have an impact on facial verification performance. Our work shows that these attributes can be as effective as demographic-based threshold in mitigating differential outcomes. This moves the focus away from identity-based attributes (i.e., gender and race), which are considered sensitive information and more controversial to use.

Future work could focus on defining thresholds based on other soft-biometric attributes, such as lighting and image quality. This may lead to systems that are more robust to different conditions and appearances of the subjects. This approach could be implemented, supervised, and modified during the deployment of facial recognition systems, as it required minimal training and can be easily extended to many biometric algorithms.

REFERENCES

Albiero, V., & Bowyer, K. W. (2020). Is face recognition sexist? no, gendered hairstyles and biology are. *2020 British Machine Vision Conference (BMVC)*.

Albiero, V., KS, K., Vangara, K., Zhang, K., King, M. C., & Bowyer, K. W. (2020). Analysis of gender inequality in face recognition accuracy. In *Proceedings of the ieee/cvf winter conference on applications of computer vision workshops* (pp. 81–89).

Albiero, V., Zhang, K., & Bowyer, K. W. (2020). How does gender balance in training data affect face recognition accuracy? In *2020 ieee international joint conference on biometrics (ijcb)* (pp. 1–10).

Anderson, E. (July 10, 2020). Controversial detroit facial recognition got him arrested for a crime he didn't commit. *Detroit Free Press*. (`https://www.freep.com/story/news/local/michigan/detroit/2020/07/10/facial-recognition-detroit-michael-oliver-robert-williams/5392166002/`)

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning*. fairmlbook.org. (`http://www.fairmlbook.org`)

Bruveris, M., Gietema, J., Mortazavian, P., & Mahadevan, M. (2020). Reducing geographic performance differentials for face recognition. In *Proceedings of the ieee/cvf winter conference on applications of computer vision workshops* (pp. 98–106).

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

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *Face and gesture recognition.*

Cavazos, J. G., Phillips, P. J., Castillo, C. D., & O'Toole, A. J. (2020). Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE Transactions on Biometrics, Behavior, and Identity Science*.

Committee, A. U. T. P. (2020). *Statement on principles and prerequisites for the development, evaluation and use of unbiased facial recognition technologies.* Online. Retrieved from `https://www.acm.org/binaries/content/assets/public-policy/ustpc-facial-recognition-tech-statement.pdf`

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 797–806).

Dantcheva, A., Elia, P., & Ross, A. (2015). What else does your biometric data reveal? a survey on soft biometrics. *IEEE Transactions on Information Forensics and Security*, *11*(3), 441–467.

Das, A., Dantcheva, A., & Bremond, F. (2018, September). Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In *Proceedings of the european conference on computer vision (eccv) workshops.*

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 4690–4699).

Drozdowski, P., Rathgeb, C., Dantcheva, A., Damer, N., & Busch, C. (2020). Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, *1*(2), 89–103.

Frontex. (2015). *Best practice technical guidelines for automated border control (abc) systems.* European Agency for the Management of Operational Cooperation at the ....

Gong, S., Liu, X., & Jain, A. K. (2020). Jointly de-biasing face recognition and demographic attribute estimation. In *European conference on computer vision* (pp. 330–347).

Grother, P. (2021). *Demographic differentials in face recognition algorithms.* (EAB VIRTUAL EVENTS SERIES – DEMOGRAPHIC FAIRNESS IN BIOMETRIC SYSTEMS)

Grother, P., Ngan, M., & Hanaoka, K. (2018). Ongoing face recognition vendor test (frvt) part 1: Verification. *National Institute of Standards and Technology*.

Grother, P., Ngan, M., & Hanaoka, K. (2019). *Face recognition vendor test (frvt) part 3: Demographic effects*. Retrieved from `http://dx.doi.org/10.6028/NIST.IR.8280` doi: 10.6028/nist.ir.8280

Hamidi, F., Scheuerman, M. K., & Branham, S. M. (2018). Gender recognition or gender reductionism? the social implications of embedded gender recognition systems. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1–13).

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Nips*.

Howard, J. J., & Etter, D. (2013). The effect of ethnicity, gender, eye color and wavelength on the biometric menagerie. In *2013 ieee international conference on technologies for homeland security (hst)* (pp. 627–632).

Howard, J. J., Sirotin, Y. B., & Vemury, A. R. (2019). 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).

Jain, A. K., Ross, A., & Prabhakar, S. (2004). An introduction to biometric recognition. *IEEE Transactions on circuits and systems for video technology*, *14*(1), 4–20.

Khan, Z., & Fu, Y. (2021). One label, one billion faces: Usage and consistency of racial

categories in computer vision. In *Proceedings of the 2021 acm conference on fairness, accountability, and transparency* (pp. 587–597).

Klare, B. F., Burge, M. J., Klontz, J. C., Bruegge, R. W. V., & Jain, A. K. (2012). Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, *7*(6), 1789–1801.

Krishnapriya, K., Albiero, V., Vangara, K., King, M. C., & Bowyer, K. W. (2020). Issues related to face recognition accuracy varying based on race and skin tone. *IEEE Transactions on Technology and Society*, *1*(1), 8–20.

Krishnapriya, K., Vangara, K., King, M. C., Albiero, V., & Bowyer, K. (2019, June). Characterizing the variability in face recognition accuracy relative to race. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr) workshops.*

Li, D. K. (December 29, 2020). Black man in new jersey misidentified by facial recognition tech and falsely jailed, lawsuit claims. *NBC News*. (`https://www.nbcnews.com/news/us-news/black-man-new-jersey-misidentified-facial-recognition-tech-falsely-jailed-n1252489`)

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., & Song, L. (2017). Sphereface: Deep hypersphere embedding for face recognition. In *The ieee conference on computer vision and pattern recognition (cvpr).*

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Michalski, D., Yiu, S. Y., & Malec, C. (2018). 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).

Morales, A., Fierrez, J., Vera-Rodriguez, R., & Tolosana, R. (2020). Sensitivenets: Learning agnostic representations with application to face images. *IEEE Transactions on Pattern*

Analysis and Machine Intelligence.

Muthukumar, V., Pedapati, T., Ratha, N., Sattigeri, P., Wu, C.-W., Kingsbury, B., ... Varshney, K. R. (2018). Understanding unequal gender classification accuracy from face images. *arXiv preprint arXiv:1812.00099*.

Ngan, M., Grother, P. J., & Ngan, M. (2015). *Face recognition vendor test (frvt) performance of automated gender classification algorithms*. US Department of Commerce, National Institute of Standards and Technology. Retrieved from `https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8052.pdf`

O'Toole, A. J., Phillips, P. J., An, X., & Dunlop, J. (2012). Demographic effects on estimates of automatic face recognition performance. *Image and Vision Computing*, *30*(3), 169–176.

Ozbulak, G., Aytar, Y., & Ekenel, H. K. (2016). How transferable are cnn-based features for age and gender classification? In *2016 international conference of the biometrics special interest group (biosig)* (pp. 1–6).

Qiu, Y., Albiero, V., King, M. C., & Bowyer, K. W. (2021). Does face recognition error echo gender classification error? *2021 IEEE International Joint Conference on Biometrics, IJCB 2021*.

Robinson, J. P., Livitz, G., Henon, Y., Qin, C., Fu, Y., & Timoner, S. (2020). Face recognition: too bias, or not too bias? In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops.*

Robinson, J. P., Qin, C., Henon, Y., Timoner, S., & Fu, Y. (2021). Balancing biases and preserving privacy on balanced faces in the wild. *arXiv preprint arXiv:2103.09118*.

Serna, I., Morales, A., Fierrez, J., Cebrian, M., Obradovich, N., & Rahwan, I. (2020). Sensitiveloss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. *arXiv preprint arXiv:2004.11246*.

Sixta, T., Junior, J. C. J., Buch-Cardona, P., Vazquez, E., & Escalera, S. (2020). Fairface challenge at eccv 2020: analyzing bias in face recognition. In *European conference on computer vision* (pp. 463–481).

Terhörst, P., Fährmann, D., Damer, N., Kirchbuchner, F., & Kuijper, A. (2020). Beyond identity: What information is stored in biometric face templates? In *2020 ieee international joint conference on biometrics (ijcb)* (pp. 1–10).

Terhörst, P., Huber, M., Kolf, J. N., Zelch, I., Damer, N., Kirchbuchner, F., & Kuijper, A. (2019a). Reliable age and gender estimation from face images: Stating the confidence of model predictions. In *10th IEEE international conference on biometrics theory, applications and systems, BTAS 2019, tampa, fl, usa, september 23-26, 2019* (pp. 1–8). IEEE. Retrieved from https://doi.org/10.1109/BTAS46853.2019.9185975 doi: 10.1109/BTAS46853.2019.9185975

Terhörst, P., Huber, M., Kolf, J. N., Zelch, I., Damer, N., Kirchbuchner, F., & Kuijper, A. (2019b). Reliable age and gender estimation from face images: Stating the confidence of model predictions. In *10th IEEE international conference on biometrics theory, applications and systems, BTAS 2019, tampa, fl, usa, september 23-26, 2019* (pp. 1–8). IEEE. Retrieved from https://doi.org/10.1109/BTAS46853.2019.9185975 doi: 10.1109/BTAS46853.2019.9185975

Terhörst, P., Kolf, J. N., Damer, N., Kirchbuchner, F., & Kuijper, A. (2020). Post-comparison mitigation of demographic bias in face recognition using fair score normalization. *Pattern Recognition Letters*, *140*, 332–338.

Terhörst, P., Kolf, J. N., Damer, N., Kirchbuchner, F., & Kuijper, A. (2020). SER-FIQ: unsupervised estimation of face image quality based on stochastic embedding robustness. In *2020 IEEE/CVF conference on computer vision and pattern recognition, CVPR 2020, seattle, wa, usa, june 13-19, 2020* (pp. 5650–5659). IEEE. Retrieved from https://doi.org/10.1109/CVPR42600.2020.00569 doi:

10.1109/CVPR42600.2020.00569

Terhörst, P., Kolf, J. N., Huber, M., Kirchbuchner, F., Damer, N., Morales, A., ... Kuijper, A. (2021). A comprehensive study on face recognition biases beyond demographics. *arXiv preprint arXiv:2103.01592*.

Terhörst, P., Tran, M. L., Damer, N., Kirchbuchner, F., & Kuijper, A. (2020). Comparison-level mitigation of ethnic bias in face recognition. In *2020 8th international workshop on biometrics and forensics (iwbf)* (pp. 1–6).

Wang, M., & Deng, W. (2020). Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 9322–9331).

Wang, M., Deng, W., Hu, J., Tao, X., & Huang, Y. (2019). Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 692–702).

Wu, W., Protopapas, P., Yang, Z., & Michalatos, P. (2020). Gender classification and bias mitigation in facial images. In *12th acm conference on web science* (pp. 106–114).

Xu, X., Huang, Y., Shen, P., Li, S., Li, J., Huang, F., ... Cui, Z. (2021, June). Consistent instance false positive improves fairness in face recognition. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 578-586).

Zhang, Y., & Deng, W. (2020). Class-balanced training for deep face recognition. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops* (pp. 824–825).

APPENDIX

A. SPHEREFACE RESULTS

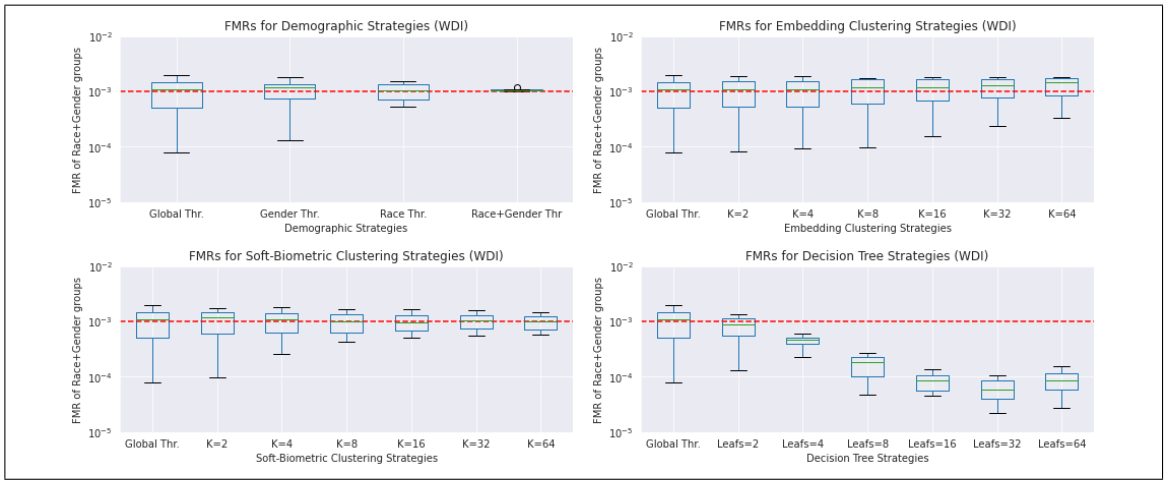

Figure A.1. Distribution of FMR for race and gender groups calculated using WDI. Facial embeddings computed using SphereFace. Reported FMR is the average performance of the folds using 5-fold cross validation. Red line is the desired FMR for the system.

Figure A.2. Distribution of FMR for race and gender groups calculated using CDI. Facial embeddings computed using SphereFace. Reported FMR is the average performance of the folds using 5-fold cross validation. Red line is the desired FMR for the system.

Table A.1. Evaluation metrics training and reporting on the WDI Scenario. Facial embeddings computed using SphereFace.

WDI Scenario	**Differential Outcome**		**Global Performance**	
	SER	**MAPE**	**Global FMR**	**Global FNMR**
Baseline				
Global Thr.	25.57	55.51%	0.001001	0.222274
Demographic-groups Thresholds				
Gender Thr.	14.05	46.58%	0.001028	0.220569
Race Thr.	2.83	34.62%	0.001031	0.213937
Race+Gender Thr.	**1.20**	**6.83%**	0.001067	0.213704
Non-sensitive Groups Thresholds				
Embedding Clustering Thr. (K=32)	7.58	52.55%	0.001199	0.213321
Soft-Biometric Clustering Thr. (K=16)	**3.32**	**36.79%**	0.001035	0.216991
DecisionTree Thr. (Leafs=32)	4.84	93.66%	0.000063	0.441907

Table A.2. Evaluation metrics training and reporting on the CDI Scenario. Facial embeddings computed using SphereFace.

CDI Scenario	**Differential Outcome**		**Global Performance**	
	SER	**MAPE**	**Global FMR**	**Global FNMR**
Baseline				
Global Thr.	11.88	44.27%	0.001002	0.151529
Demographic-groups Thresholds				
Demographic Thr. (gender)	5.69	35.73%	0.001007	0.150691
Demographic Thr. (race)	2.29	29.67%	0.001008	0.149332
Demographic Thr. (race+gender)	**1.05**	**1.41%**	0.001014	0.148903
Non-sensitive Groups Thresholds				
Embedding Clustering Thr. (K=32)	5.04	33.50%	0.001054	0.149736
Soft-Biometric Clustering Thr. (K=16)	3.20	26.64%	0.000971	0.151875
DecisionTree Thr. (Leafs=32)	**2.11**	**23.40%**	0.001031	0.170985

B. EFFECT DIMENSIONALITY REDUCTION ON K-MEANS PERFORMANCE

This section contains the results of the experiments applying a dimensionality reduction on the facial embeddings before performing the clustering using K-Means. The dimensionality reduction was performed using Principal Component Analysis (PCA). In general, we did not see a major improvement on bias mitigation by reducing the dimensionality.

B.1. CDI Scenario

Table B.1. Comparison of SER using PCA for dimensionality reduction before K-Means. Training and testing performed using CDI Scenario

	K=2	K=4	K=8	K=16	K=32	K=64
No PCA	4.95	5.00	4.74	4.46	4.06	4.38
PCA (dim=10)	4.90	4.94	4.76	4.94	5.24	3.97
PCA (dim=20)	4.96	4.85	5.04	4.72	4.51	4.53
PCA (dim=40)	4.94	4.88	4.79	4.83	4.52	4.32
PCA (dim=60)	4.92	4.93	4.74	4.51	4.54	4.38
PCA (dim=80)	4.99	4.94	4.80	4.43	4.82	4.12

Table B.2. Comparison of MAPE using PCA for dimensionality reduction before K-Means. Training and testing performed using CDI Scenario

	K=2	K=4	K=8	K=16	K=32	K=64
No PCA	36.24%	35.90%	35.99%	35.28%	29.95%	31.07%
PCA (dim=10)	36.13%	36.31%	35.74%	34.59%	31.47%	50.11%
PCA (dim=20)	36.06%	35.53%	36.17%	34.79%	29.34%	43.45%
PCA (dim=40)	36.11%	35.92%	35.29%	35.51%	29.80%	28.70%
PCA (dim=60)	35.94%	35.81%	35.57%	35.32%	32.53%	29.81%
PCA (dim=80)	36.10%	35.85%	36.09%	36.12%	34.11%	30.30%

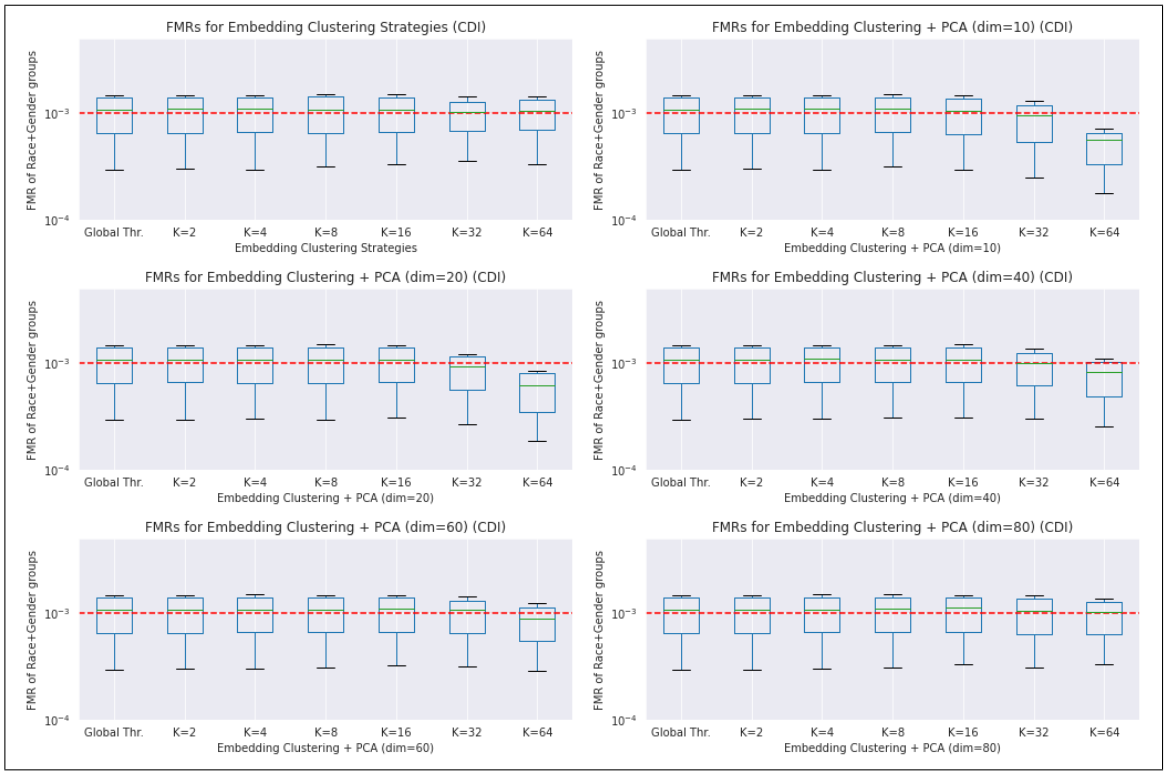

Figure B.1. Distribution of FMR for race and gender groups calculated using CDI. Reported FMR is the average performance of the folds using 5-fold cross validation. Red line is the desired FMR for the system.

B.2. WDI Scenario

Table B.3. Comparison of SER using PCA for dimensionality reduction before K-Means. Training and testing performed using WDI Scenario

	K=2	K=4	K=8	K=16	K=32	K=64
No PCA	12.02	11.53	12.13	10.63	8.46	9.65
PCA (dim=10)	11.69	11.99	11.67	12.09	11.39	10.96
PCA (dim=20)	11.83	11.91	12.02	11.49	11.35	10.56
PCA (dim=40)	12.04	11.96	11.89	11.18	10.57	9.41
PCA (dim=60)	11.60	11.98	10.90	11.81	10.75	9.25
PCA (dim=80)	11.98	11.66	11.76	9.54	9.93	9.51

Table B.4. Comparison of MAPE using PCA for dimensionality reduction before K-Means. Training and testing performed using WDI Scenario

	K=2	K=4	K=8	K=16	K=32	K=64
No PCA	51.56%	51.08%	53.24%	54.98%	54.20%	53.02%
PCA (dim=10)	50.79%	51.63%	52.41%	52.53%	52.90%	51.31%
PCA (dim=20)	51.19%	52.00%	52.22%	52.95%	52.24%	56.01%
PCA (dim=40)	51.29%	50.94%	53.64%	54.75%	56.41%	59.88%
PCA (dim=60)	51.30%	51.59%	50.90%	53.29%	54.36%	55.96%
PCA (dim=80)	51.76%	50.74%	52.46%	52.43%	53.44%	60.27%

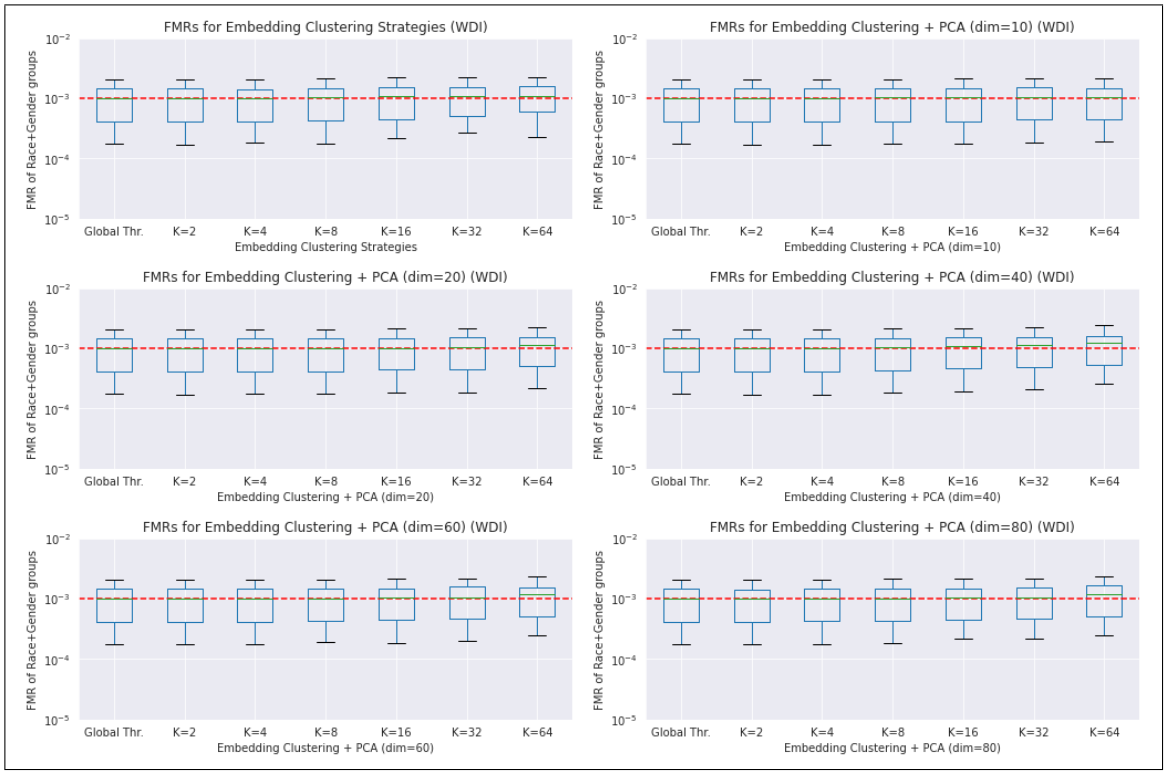

Figure B.2. Distribution of FMR for race and gender groups calculated using WDI. Reported FMR is the average performance of the folds using 5-fold cross validation. Red line is the desired FMR for the system.