# Fairness in Biometrics: a figure of merit to assess biometric verification systems

Tiago de Freitas Pereira, Sébastien Marcel

Idiap Research Institute

{tiago.pereira,sebastien.marcel}@idiap.ch

**Abstract**—Machine learning-based (ML) systems are being largely deployed since the last decade in a myriad of scenarios impacting several instances in our daily lives. With this vast sort of applications, aspects of fairness start to rise in the spotlight due to the social impact that this can get in some social groups. In this work aspects of fairness in biometrics are addressed. First, we introduce a figure of merit that is able to evaluate and compare fairness aspects between multiple biometric verification systems, the so-called Fairness Discrepancy Rate (FDR). A use case with two synthetic biometric systems is introduced and demonstrates the potential of this figure of merit in extreme cases of demographic differentials. Second, a use case using face biometrics is presented where several systems are evaluated compared with this new figure of merit using three public datasets exploring gender and race demographics.

**Index Terms**—Biometrics, Fairness, Face Recognition

---◆---

## 1 INTRODUCTION

THe pipeline from research to deployment of an ML-based system can assume several shapes with different steps. In abstract terms (and allow us to do such simplification), such pipeline is composed of i-) **Data Collection**: where the "state of the world" is reduced to a set of rows and columns of data (e.g. face images, bank transactions, medical data, etc...); ii-) **Modelling**: where the "model" is supposed to summarize the patterns of the data and be able to make generalizations (via supervised/unsupervised learning, etc..); iii-) **Benchmarking**: where the model is evaluated with respect to some figure of merit (e.g. accuracy, f1-score, etc..); iv-) **Feedback**: where it is decided if the model is "good" for deployment or not; if not, steps (i) and/or (ii) needs to be redone; v-) **Deployment**: ML-System goes to production[1]. During the benchmarking stage, it is common to use reference databases. Such reference databases are supposed to represent somehow operational conditions and it is hypothesized that ML-based systems that presents high accuracy, high f1 score, low false-positive rate, low false-negative rate, etc in such benchmarks is a proxy to have the same figures of merit in operational conditions. Once this is achieved (by any criteria ML engineers decide), ML is "safe" to be deployed.

Fairness issues arise from the analysis of these figures of merit in specific demographics groups (e.g., gender, ethnicity, race, revenue levels, or any covariate in general) and the observation that operational conditions estimated initially cannot be reproduced in those. In biometrics, this is coined as demographic differentials. The large-scale deployment of such systems in so many different scenarios raises the debate about how these differentials impact people's lives.

For instance, the book Weapons of Math Destruction [1] presents several cases where unfair decision-making tools based on ML impacted the life of city populations

---

1. Usually, feedbacks are also done after deployment, but let's keep this simplification as is because it is enough for our purposes.
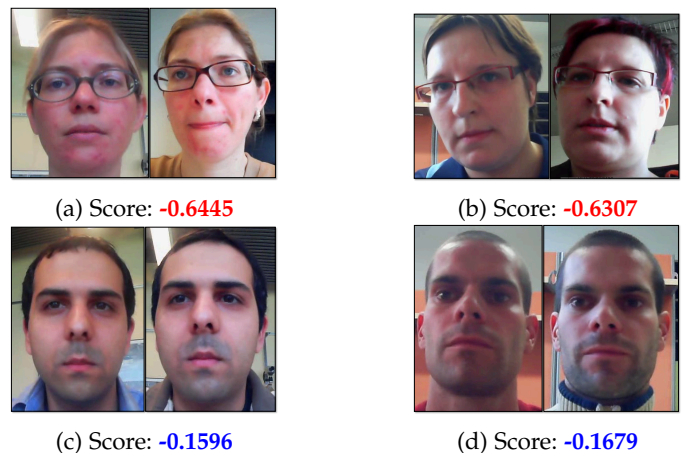


Fig. 1: Inception Resnet v2 FR system [2]: Genuine comparison pairs extracted MOBIO Database [3]. The decision threshold is: $\tau = -0.5298$. Hence, score $\geq \tau$, the comparison pair is accepted; otherwise it is rejected.

negatively. Such cases cover different applications, such as online advertising tools, automatic resumé evaluations for HR, and credit score tools for different purposes (e.g., bank or insurance companies).

Decision-making tools based on biometrics, as part of this Machine Learning wave, have been largely deployed in the recent decade. For instance, it is present in our daily lives for data protection (e.g., unlock mobile phones or computers), law enforcement, airport e-gates, and other applications. The public media has been reported several cases where biometric systems, especially Face Recognition (FR) technology, present recognition disparities between different demographic groups. For instance, a simple test executed in 2018 using a Commercial-Off-The-Shelf (COTS) system comparing face images from members of the US

parliament found 28 false matches, and most of them occur with "people of color"[2]. In 2020, the Association for Computing Machinery in New York City recommended the suspension of private and government use of FR technology due to "clear bias based on ethnic, racial, gender, and other human characteristics"[3].

This work addresses demographic differential aspects in biometric systems, and its contributions are twofold. First, it discusses the factors to consider a biometric verification system as fair and introduces a figure of merit called Fairness Discrepancy Rate. Second, a case study of this figure of merit is presented using face recognition as a biometric trait. We aim to make this reproducible: all the source code, trained models, and scores are made publicly available. Details on how to reproduce this work can be found on the provided link[4].

## 2 RELATED WORK

In this section, we present the related work by first discussing the Machine Learning community's efforts to suppress demographic differentials, and then we move to actions made by the biometrics community in this direction.

### 2.1 Machine Learning Background

Many criteria to assess and address differentials in pattern recognition problems have been proposed over the years, each one phrasing the problem in different ways. The recent work from [4] hypothesizes that most of these criteria described in the machine learning literature fall into three major categories of conditional independence, and they are: Independence, Separation, and Sufficiency.

To illustrate these criteria, let's consider $X \in \mathbb{R}^n$ a random variable denoting the input data, $D = \{d_1, d_2...d_n\}$ a random variable denoting a set of sensitive attributes (e.g. gender, demographics, etc), $Y \in \{0,1\}$ (for simplicity) a random variable denoting the target variable (representing a binary classifier) and $F : f(X, D)$ the trained predictor (that can be possibly thresholded). The first non-discrimination criteria, and the most simplistic one, is **independence** which requires that the classifier $F$ must be independent of the sensitive attributes $D$, or $F \perp D$. This is also addressed as *demographic parity* or *statistical parity*. For our binary classification case, this can be rewritten as:

$$P\{F = 1|D = d_i\} = P\{F = 1|D = d_j\} \forall \{i, j\} \in D. \quad (1)$$

This criterion is largely used in ML, in general, to mitigate differentials either via regularization [5], [6], representation learning constraints [7], [8], [9], or post-processing mechanisms [10]. The independence criteria, although popular, have some limitations. For instance, at training-time, a classifier might trade false positives to false negatives in a group to match independence. Other limitations of independence are largely discussed in [11] and more recently in [4].

The second criterion is **separation** (or equalized odds criteria), where it explicitly acknowledges that the target variable $Y$ might be correlated with $D$. Conditioning in $Y$ might be desirable in some scenarios. For instance, a medical doctor might argue that a particular disease is more likely to be developed in one demographic group than others, and a "disease" prediction function $F$ must take this into account. The following condition independence summarizes this: $F \perp D|Y$. For our binary classification case, this is equivalent to:

$$P\{F = 1|Y = y, D = d_i\} = P\{F = 1|Y = y, D = d_j\}$$
$$\forall \{i, j\} \in D \qquad (2)$$
$$y \in \{0, 1\}$$

The work from [11] studied different formulations of separation (e.g., Equalized Odds and Equality of Opportunity) and proposed different constrained optimization approaches to reach separation. Using FICO scores as a case study demonstrates that such criterion can "remediate" some of the problems from methods based on independence. The authors from [12] followed a similar path. In [6] the authors propose different regularization strategies to reach independence by making risk scores threshold independent.

The third criterion is **sufficiency**. $F$ is sufficient to predict $Y$ independently of the demographic attribute $D$ if the following conditional probability is matched: $Y \perp D|F$.

In this case:

$$P\{Y = 1|F = s, D = d_i\} = P\{Y = 1|F = s, D = d_j\}$$
$$\forall \{i, j\} \in D. \qquad (3)$$

Most of the literature that approaches sufficiency estimators does so via calibration of the risk scores by group. One of the most popular methods to perform such calibration is Platt scaling [13], which, roughly, consists of fitting a sigmoid function on uncalibrated scores. More recently, in [14] the authors propose different optimization mechanisms to achieve sufficiency.

Those three essential fairness criteria support most of what was published in the Machine Learning literature explicitly or implicitly.

### 2.2 Fairness in Biometrics

In the biometrics literature, some recent work addresses demographic differentials for some biometric traits. For instance, the Face Recognition Vendor Test (FRVT) has a special report addressing demographic effects in FR [15]. This report presents several analyses observing the effect, mostly of race and gender in terms of False Matches and False non Matches with more than 100 COTS FR systems. More recently, a comprehensive survey on demographics differentials on biometrics [16] demonstrated efforts in this direction on the face, fingerprint, palm vein, iris, and palm print recognition research. In March 2021, the European Association for Biometrics (EAB) organized an event[5] dedicated to such a topic. Researchers from industry, governments, and academia from all over the globe shared their

---

2. https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28

3. https://www.acm.org/binaries/content/assets/public-policy/ustpc-facial-recognition-tech-statement.pdf

4. https://gitlab.idiap.ch/bob/bob.paper.fdr

5. Demographic Fairness in Biometric Systems - https://eab.org/events/program/237
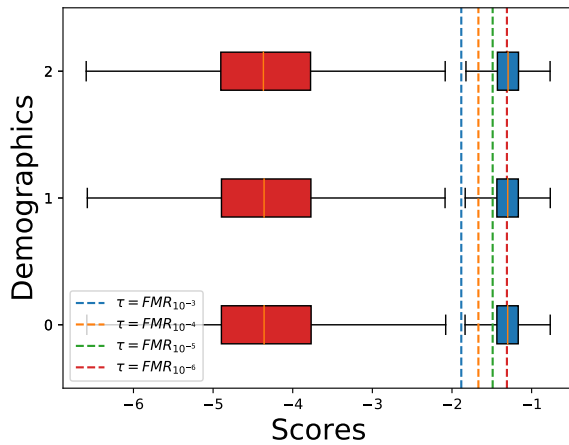
Fig. 2: Example of a canonical fair biometric verification system with three demographics (0, 1, 2) and four operational thresholds (depicted with the dashed lines). Performance measures in terms of $FMR(\tau)$ and $FNMR(\tau)$ can be found in TABLE 1. The zero-effort impostor scores are represented in red; genuine scores are represented in blue.

TABLE 1: Canonical fair biometric verification system: $FNMR(\tau)$, $FMR(\tau)$, and $FDR(\tau)$ per demographic (Demog.) where the operational points are defined as $\tau = FMR_x{}^{*}$

| $\tau = FMR_{10^x}$ | $10^{-3}$ | $10^{-4}$ | $10^{-5}$ | $10^{-6}$ |
|---|---|---|---|---|
| Demog. | $FMR(\tau)$ | | | |
| 0 | 0.001 | 0.000 | 0.000 | 0.000 |
| 1 | 0.001 | 0.000 | 0.000 | 0.000 |
| 2 | 0.001 | 0.000 | 0.000 | 0.000 |
| | $FNMR(\tau)$ | | | |
| 0 | 0.003 | 0.036 | 0.179 | 0.484 |
| 1 | 0.003 | 0.039 | 0.178 | 0.481 |
| 2 | 0.003 | 0.036 | 0.168 | 0.469 |
| $FDR(\tau)$ | 0.999 | 0.999 | 0.994 | 0.992 |

${}^{*}\tau$ is set using an independent zero-effort impostor score distribution with scores from all demographics. It can be seen as a development set.

expertise in such an effort to measure and mitigate possible demographic differentials in biometric systems.

The recent work from [17] describes some underlying factors that bias COTS face recognition systems concerning **race**. For instance, they observed that the "Other Race Effect", well known in humans, [18] can also be observed in FR algorithms; FR systems developed in Asia are more accurate with Asian subjects than with Caucasians, and vice-versa. Furthermore, they observed that demographic differentials are frequently observed where low-quality probe samples are used. The FRVT report also raised such observations about image quality. Studying race, the work from [19] observed consistently higher False Match Rates (FMR) with African American cohorts compared with Caucasians using two COTS systems. Furthermore, this work extended its analysis with ICAO face checker[6]. They observed that ICAO SDKs work better with Caucasian subjects than with African Americans. The work from [20] made an extensive study analyzing several **age** cohorts using one COTS system. Among several observations made, the most impacting one was the high FMR and high False Non-Match Rates (FNMR) in pairs of images where age is lower than four years old.

Focusing on face biometrics, the work from [5] introduces the Racial Faces in the Wild dataset. Such dataset is a subset of the MSCeleb-1M [21] whose identities are organized in four different races (Caucasians, Black, Indian and Asian). Using such data, the authors, at training-time, regularized different deep neural networks by minimizing the Mutual Information between the face recognition classifier and the demographic attributes. In [22] the authors leverage from synthetically generated data to train fairer DCNNs for FR. The authors from [23] propose a framework to automatically calibrate the weight of samples at training-

time of a DCNN based on protected attributes. In [24] the authors followed a similar path in the task of facial attributes classification. The authors from [25] propose a post-processing mechanism based on score normalization to deal with demographic discrepancies.

Demographic differentials towards gender were observed using periocular biometrics. For instance, the work from [26] demonstrates that several periocular recognition systems perform better with male subjects than with female ones.

The NIST SRE[7] is the most relevant benchmark for speaker recognition, and along with last editions, it consistently evaluates error rates looking at gender cohorts.

To the best of our knowledge, the works from biometric literature that addresses somehow demographic differentials by either analyzing COTS systems or proposing a strategy to mitigate it do so using different criteria. However, the trend seems to achieve somehow the *statistical parity* (or independence), even if this detail is not explicitly mentioned. Even if this is the trend, a figure of merit to directly address it is nonexistent. For instance, the work from [27] uses the Area Under the ROC curve as a figure of merit to measure demographic differentials. ROC curves measure the True and False Positive Rates (TPR and FPR, respectively) trade-offs. Although this seems sensible to assess demographic discrepancies, it has a serious flaw; it assumes that the verification decision threshold (let us call it $\tau$; we will formally define this further) is demographic-specific. Hence, $TPR(\tau)$ and $FPR(\tau)$ is computed under different decision thresholds depending of the demographic and can give a false impression that a biometric verification system is fair (this problem is further discussed in section 3). Furthermore, this does not represent operational conditions where normally one single $\tau$ is set, and this operational point has to be fair for different demographics. This issue with ROC curves was recently raised in [28]. This problem can be observed also in several works that refers to biometric verification; for instance, in [5], [17], [26], [29], [30].

Some works in the biometrics literature **explicitly** analyses the possibility of $\tau$ being demographic-specific, such as in [31], [32]. Even in [4, (sec.2, p.14)] (covering a general case of pattern recognition), the authors work with the possibility of one $\tau$ per demographic. For biometric recognition, strategies in this direction might not be ethical, since it imposes

6. https://www.icao.int/Security/FAL/TRIP/Documents/TR - Portrait Quality v1.0.pdf
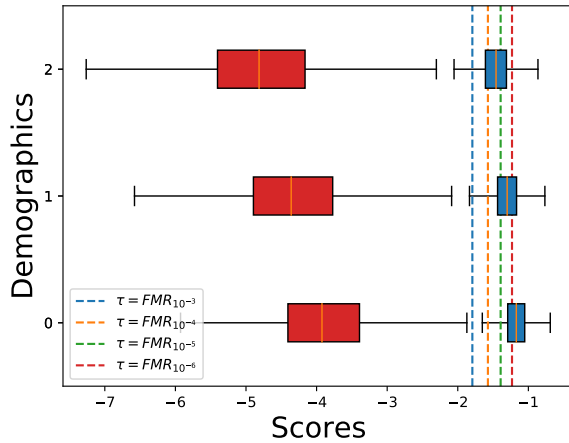
7. https://sre.nist.gov/

Fig. 3: Example of a canonical UNfair biometric verification system with three demographics (0, 1, 2) and four operational thresholds (depicted with the dashed lines). Performance measures in terms of $FMR(\tau)$ and $FNMR(\tau)$ can be found in TABLE 2. The zero-effort impostor scores are represented in red; genuine scores are represented in blue.

TABLE 2: Canonical UNfair biometric verification system: $FNMR(\tau)$, $FMR(\tau)$, and $FDR(\tau)$ per demographic (Demog.) where the operational points are defined as $\tau = FMR_x{}^*$

| $\tau = FMR_{10^x}$ | $10^{-3}$ | $10^{-4}$ | $10^{-5}$ | $10^{-6}$ |
|---|---|---|---|---|
| Demog. | $FMR^{\text{test}}(\tau)$ | | | |
| 0 | 0.002 | 0.002 | 0.000 | 0.000 |
| 1 | 0.000 | 0.000 | 0.000 | 0.000 |
| 2 | 0.000 | 0.000 | 0.000 | 0.000 |
| | $FNMR^{\text{test}}(\tau)$ | | | |
| 0 | 0.001 | 0.016 | 0.114 | 0.369 |
| 1 | 0.009 | 0.092 | 0.321 | 0.637 |
| 2 | 0.071 | 0.304 | 0.614 | 0.855 |
| $FDR(\tau)$ | 0.963 | 0.855 | 0.750 | 0.756 |

$^*\tau$ is set using an independent zero-effort impostor score distribution
with scores from all demographics. It can be seen as a development set.

a disparate treatment on deployed systems. Furthermore, in practical terms, this can involve another classification task in the biometric recognition pipeline, which might be error-prone and subject to disparities.

FRVT goes in the right direction concerning the afore-mentioned threshold problem by discussing the impact of demographics in terms of $FMR(\tau)$ and $FNMR(\tau)$ for one decision threshold only. Such a decision threshold is picked from an independent zero-effort score distribution, where the demographic does not play a role. This is the most sensible evaluation if the goal is to assess demographic differentials in operational conditions. However, FRVT discussess the impact of $FMR(\tau)$ and $FNMR(\tau)$ **separately**. Hence, the trade-off between them is not considered. Furthermore, only one decision threshold is analyzed, limiting the perception of demographic differentials under different operational points. In [6] a similar direction was taken where risk distributions among the different demographic groups were equalized via different approximation methods, introducing then threshold invariant classifiers. However, no analysis in terms $FMR(\tau)$ and $FNMR(\tau)$ was carried out.

Our work tries to fill these evaluation gaps for biometric verification systems by: (i) - taking into consideration the above-mentioned threshold problems (ii) - considering the trade-off between $FMR(\tau)$ and $FNMR(\tau)$ in the demographic differential assessment, and (iii) - taking into account different operation points (decision thresholds).

## 3 FAIRNESS DISCREPANCY RATE: A FIGURE OF MERIT TO ASSESS FAIRNESS IN BIOMETRIC VERIFICATION

Biometric verification is the task of verifying if a given sample is from a claimed identity or not. This decision is made based on a scoring function $s(e, p)$ and a decision

threshold $\tau$, where $e$ is the claimed identity, $p$ is a probe sample (test sample). If $s(e, p) \geq \tau$ it is said that $e$ and $p$ are from the same identity. Conversely, if $s(e, p) < \tau$ it is said that $e$ and $p$ are not from the same identity. There are two possible types of errors that biometric verification systems can make, and they are the False Match Rate (**FMR**) and False Non-Match Rate (**FNMR**). Worth noting that these two errors are functions of a decision threshold $\tau$, whose impact is further discussed.

The value of $\tau$ plays a decisive role in these two errors. Its value is usually set by targeting a specific FMR value in a reference impostor score distribution set[8]. Some examples of such operational points are: $\tau = FMR_{10^{-1}}$ corresponds to the $\tau$ where FMR reaches 0.1 (or 10%) in the impostor distribution scores; $\tau = FMR_{10^{-3}}$ corresponds to the $\tau$ where FMR reaches 0.001 (or 0.1%) in the impostor distribution scores; $\tau = FMR_{10^{-6}}$ corresponds to the $\tau$ where FMR reaches $10^{-6}$ (or 0.001%) in the impostor distribution scores[9].

Given a test set, "good" biometric recognition systems should present $FMR_x(\tau)$ around the operational point given by $x$ and the lowest value as possible for $FNMR(\tau)$. Furthermore, for a "good" biometric system to be considered fair, it should present $FMR_x(\tau)$ around the operational point $x$ for all observed demographic groups and approximately "same" $FNMR(\tau)$ for all observed demographic groups. The impact of the decision thresholds is illustrated in Figure 1. In this example, we chose two comparison scores from male and female subjects of the MOBIO dataset using one of our tested Deep Convolutional Neural Network (DCNN) (see section 4 for further details). Those genuine pairs were cherry-picked by looking at the score values around the average genuine scores for each demographic group. $\tau$, in this case, equals $-0.5298$. It can be noticed that both comparisons using female subjects are rejected using this operational point, and the two male subjects are accepted.

In this work, a biometric verification system is considered fair if statistical parity between groups is reached in terms of both $FMR$ and $FNMR$ for a given decision threshold $\tau$. More formally, given a set of demographic

8. Impostor score distribution is made of $s(e, p)$ values where $e$ and $p$ are not from the same identity

9. Examples on how this is set in practice can be seen in this report https://pages.nist.gov/frvt/reports/11/frvt_11_report.pdf section 1.1

groups $\mathcal{D} = \{d_1, d_2, ..., d_n\}$, and $\tau = \mathrm{FMR}_x$[10], a biometric verification system is considered fair with respect to FMR if the following premisse holds:

**Premisse 1.** $\mathrm{FMR}^{d_i}(\tau) \geq \mathrm{FMR}^{d_j}(\tau) - \epsilon \ \forall d_i, d_j \in D$.

Such premisse can be written with the following equation:

$$A(\tau) = \max(|\mathrm{FMR}^{d_i}(\tau) - \mathrm{FMR}^{d_j}(\tau)|) \leq \epsilon \ \forall d_i, d_j \in \mathcal{D}, \quad (4)$$

where $\epsilon$ is a relaxation constraint.

Conversely, in terms of FNMR, a biometric verification system is considered fair if the following premisse holds:

**Premisse 2.** $\mathrm{FNMR}^{d_i}(\tau) \geq \mathrm{FNMR}^{d_j}(\tau) \ \forall d_i, d_j \in D$.

Such premisse can be written with the following equation:

$$B(\tau) = \max(|\mathrm{FNMR}^{d_i}(\tau) - \mathrm{FNMR}^{d_j}(\tau)|) \leq \epsilon \ \forall d_i, d_j \in \mathcal{D}. \quad (5)$$

Since 4 and 5 are functions of $\tau$, both can be summarized in one figure of merit, that we refer as Fairness Discrepancy Rate (FDR) which is defined as:

$$FDR(\tau) = 1 - (\alpha A(\tau) + (1 - \alpha)B(\tau)), \quad (6)$$

where $\alpha$ is a hyper-parameter that defines the weight of $A(\tau)$ in the figure of merit (the importance of False Matches).

The values that FDR can take varies from 0 (maximum discrepancy between two demographic groups) to 1 (minimum discrepancy between two demographic groups). Worth noting that we have chosen to compute maximum differences in the computation of $A(\tau)$ and $B(\tau)$ instead of scaling them using some statistics (e.g., z-norm) or ratios between false matches or false non-matches of different demographic groups. We have chosen this path, so we can better explore the differentials for a large range of $\tau$, where $\mathrm{FNMR}^{d_i}(\tau) = 0$ is a possibility for some ranges and, the computation of such statistics in these cases leads to arithmetical problems.

In the computation of $A(\tau)$, only false matches from biometric-references and probe pairs from homogeneous groups are considered (e.g., comparison scores between Black-White, Female-Male, ... samples). Several works in the literature ( [15], [20], [33]) show that the number of false matches from non-homogenous groups is substantially lower than with homogenous groups. Hence, to enforce parity between homogenous and non-homogeneous groups seems counter-productive since non-homogeneous comparison pairs present naturally lower false-matches. As with equations 4 and 5 FDR can be possibly thresholded with a slack variable $\epsilon$ and an overall threshold defining what is fair and what is not can be defined as:

$$\begin{cases} \text{fair} & \text{if } FDR(\tau) \geq 1 - \epsilon \\ \text{unfair} & \text{otherwise} \end{cases} \quad (7)$$

The role of $\epsilon$ is discussed further in this section.

The following subsection presents one example of a desired fair biometric recognition system and one example of an undesired unfair biometric verification system that illustrates how FDR evaluates these two systems.

### 3.1 Fairness Discrepancy Rate using synthetic data

Figure 2 shows a canonical fictional example of a fair biometric recognition system. Each box plot shows the score distributions from both zeroth effort impostors (in red) and genuines (in blue) of three abstract demographics (labeled as 0, 1, and 2). It is possible to visually inspect that the score distribution from the three demographics is systematically aligned in all quartiles, which indicates that Premisses 1 and 2 can hold for both FMR and FNMR for any given $\tau$. In this experiment $\tau = \mathrm{FMR}_x(\tau)$ where $x$ varies from $10^{-3}$ to $10^{-6}$. Conversely, on the other side of the spectrum, an example of an unfair biometric verification system is presented in Figure 3. As it can be noticed, the score distributions from both zeroth effort impostors and genuines are not as aligned as in the previous example (see Figure 2). Intuitively, one can argue that having a single threshold $\tau$ that holds premisses 1 and 2 can be problematic.

Let us now test FDR using these two theoretical systems[11]. Table 1 presents $\mathrm{FNMR}(\tau)$, $\mathrm{FMR}(\tau)$ and $\mathrm{FDR}(\tau)$ for different values of $\tau$ of the fair synthetic biometric system presented in Figure 2. In this experiment $\tau = \mathrm{FMR}_x(\tau)$ where $x$ varies from $10^{-3}$ to $10^{-6}$. It is possible to observe that $\mathrm{FDR}(\tau)$ is stable and higher than 0.99 for all values of $x$, which indicates non-discrepant behavior concerning these abstract demographics. To analyse the other side of the spectrum, Table 2 presents $\mathrm{FNMR}(\tau)$, $\mathrm{FMR}(\tau)$ and $\mathrm{FDR}(\tau)$ for different values of $\tau$ of the unfair synthetic biometric system presented in Figure 3. The values of $\tau$ are set in the same way as in the previous experiment. It is possible to observe that $\mathrm{FDR}(\tau)$ is consistently higher for the fair biometric system than with the unfair one, which indicates consistency in this figure of merit in the evaluation of demographic differentials. On the other hand, Figure 4 shows the Detection Error Tradeoff (DET) curves of these two synthetic examples and the three demographics. It is possible to observe that all six DET curves present the same trends, overlapping each other. This example gives the false impression that the unfair synthetic verification system is fair, which is not as we could spot with the FDR.

It is possible to plot FDR as a function of $x$ (or $\tau$) to compare two biometric systems. Figure 5 presents how the two biometric systems can be compared under this figure of merit. It is possible to observe that FDR is stable for all values of $x$ for the fair biometric system. FDR substantially decreases once $x$ decreases (when fewer false-matches are allowed) for the unfair biometric system. Another way to establish a comparison between two systems concerning demographic discrepancies is by analyzing the Area Under FDR. For a given range of $\tau$ (estimated by using $x$), the Area Under FDR can be calculated by merely integrating the $\mathrm{FDR}(\tau)$ over $x$. The value of $x$ is scaled from 0 to 1, so Area Under FDR is bounded to this range. However, by scaling it, the range of $x$ has to be reported. Hence, only Area Under FDR whose range of $x$ matches can be fairly compared. It is also possible to observe that, using this figure of merit, the

---

10. $\tau$ is set using an independent zero-effort impostor score distribution with scores from all demographics

11. This example is available in the following link: https://github.com/tiagofrepereira2012/fdr/
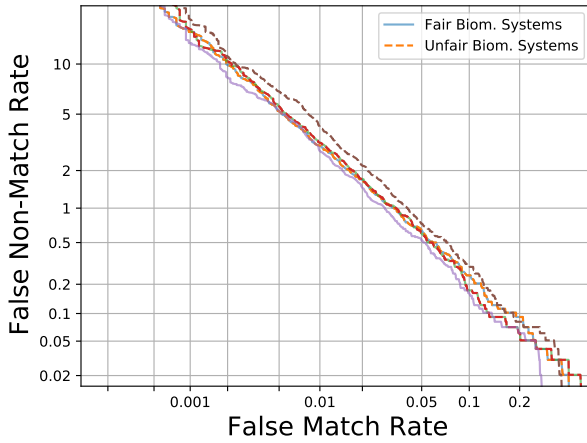
Fig. 4: DET curves for the canonical fair and unfair synthetic verification systems. It can be observed that analysing this curves gives a false impression that the unfair synthetic verification system is fair.

system that was intuitively considered as fair (see Figure 2) presents a Area Under FDR of $0.99$ and the one that was intuitively considered as unfair presents a Area Under FDR of $0.88$ (see Figure 3).
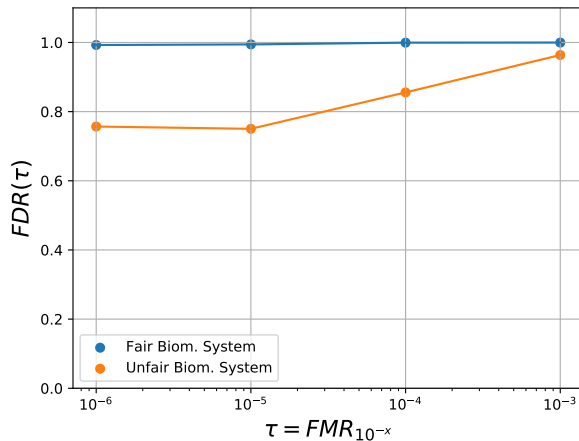


Fig. 5: FDR as a function of $\tau$ from two synthetic biometric systems from Figures 2 and 3.

## 3.2 The role of alpha

The hyper-parameter $\alpha \in [0, 1]$ in equation 6 has a crucial role in the computation of $\text{FDR}(\tau)$. As previously mentioned, it controls the weight of False Matches in the FDR computation. Such value is a business/application decision. For instance, a bank that deploys a biometric verification system in an ATM might prefer to favor parity towards False Non-Matches and, for this reason, $\alpha$ can assume low values. On the other hand, in a border control scenario, where false matches are critical, decision-makers might favor False-Matches' parity. Hence, $\alpha$ should be high.

Figure 6 shows the $\alpha$ trade-off between the two synthetic systems; the fair ones are represented by the solid lines and

the unfair by the dashed lines. It is possible to observe that the fair system presents a $\text{FDR}(\tau) \sim 0.98$ no matter the value of $\alpha$. For the unfair system, $\text{FDR}(\tau)$ presents a stepper decay once $\alpha$ decreases. This also can be seen via the Area Under FDR. As can be noticed in Table 3, for the unfair biometric system, the Area Under FDR substantially changes once $\alpha$ changes.
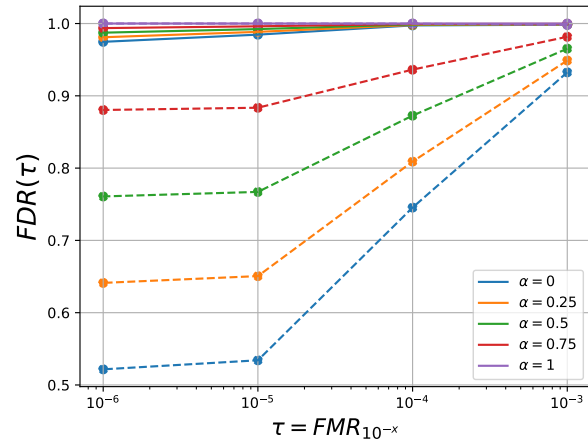


Fig. 6: $\text{FDR}(\tau)|\alpha$ for different values of $\alpha$ (see equation 6). Solid lines represent the $\text{FDR}(\tau)|\alpha$ for the synthetic fair verification system. Dashed lines represent the $\text{FDR}(\tau)|\alpha$ for the synthetic unfair verification system.

TABLE 3: Area Under the FDR for different values of $\alpha$ for $x$ varying from $10^{-3}$ to $10^{-6}$.

|  | Fair | Unfair |
|---|---|---|
| $\alpha = 0$ | 0.99 | 0.81 |
| $\alpha = 0.25$ | 0.99 | 0.86 |
| $\alpha = 0.5$ | 0.99 | 0.90 |
| $\alpha = 0.75$ | 0.99 | 0.95 |
| $\alpha = 1$ | 0.99 | 0.98 |

## 3.3 The role of epsilon

Another hyper-parameter in this figure of merit is the values that $\epsilon$ can assume. As mentioned in equation 7, this value is supposed to define what is fair and what is not with respect to $\text{FDR}(\tau)$. To the best of our knowledge, there is no reference value that we can rely upon. In some particular cases, there are some guidelines. For instance, as mentioned in [4, (sec 2, p.19)], the U.S Equal Employment Opportunity Commission[12] states that a disparate behavior between two groups occurs if the probability of selection between them differ from more than 20%. We can reasonably agree that having $\epsilon \sim 0.20$ in equation 7 is not realistic for a biometric verification system.

This work will not draw a line to define what is fair and what is not for biometric verification systems. As mentioned before, there is no legal or technical basis for such, and the ones that do exists are not suitable for biometrics. Instead, we will use both FDR and Area Under FDR to compare

12. https://www.eeoc.gov/laws/guidance/employment-tests-and-selection-procedures

different biometric verification systems in terms of demographic differentials.

## 4 FACE VERIFICATION USE CASE

This section presents a case study of the Fairness Discrepancy Rate using three FR systems based on DCNNs and one Commercial-Off-The-Shelf System (COTS). The **first** DCNN system is based on the Inception-Resnet v2 architecture [2]. This model was trained with the MS-Celeb-1M dataset using a joint loss function combining the cross-entropy loss and center loss. Check in [34, p.147] more details on how this DCNN was trained. The **second** DCNN is the ArcFace model [35] from InsightFace (ArcFace-InsightFace)[13]. This model is based on Resnet-101 architecture [36], and trained using the ArcFace loss. The **third** is Resnet-50 based architecture [36] trained using the ArcFace loss. Such a model was trained using the VGG2Face dataset [37]. More details on this DCNN was trained can be found in [14]. For these three biometric systems, comparisons between samples are made with the embeddings of each DCNN using the cosine distance metric. Given the embeddings $e$ and $p$ for enrollment and probing, respectively, the distance $s(e, p)$ is given by:

$$s(e, p) = 1 - \frac{e \cdot p}{\|e\| \cdot \|p\|} \tag{8}$$

Finally, the **fourth** face verification system a COTS developed by RankONE [15].

### 4.1 Dataset setup

Several datasets are publicly available in the literature with privacy-sensible attributes where FR systems assessments can be carried out. The most recent ones available are based on images from celebrities scraped from the web, such as Racial Faces in the Wild (RFW) [5], Balanced Faces in the Wild [29] and, IARPA Janus Benchmark C (IJB-C) [38]. Although all the aforementioned datasets contain meta-information that can be used to assess demographic differentials, they were captured in the so-called "in a wild" conditions. Several factors of variation, such as pose, illumination, occlusion, and image quality, can play a role in the final recognition rates, which might interfere with this initial FDR assessment. Since this is the first work with this figure of merit, we have focused on three datasets where capture conditions are "less in the wild" and whose demographic attributes are available.

NIST developed the MEDS II database to support and assists their biometrics evaluation program. Five hundred eighteen identities compose it from both men/women (labeled as M and F) and five different race annotations. They are Asian, Black, American Indian, Unknown, and White (labeled as A, B, I, U, and W). Unfortunately, the distribution of gender and race is hugely unbalanced. Furthermore, only 256 subjects have more than one image sample (obviously, it is impossible to do a biometric evaluation with one sample per subject). For this reason, we have performed our evaluation in a subset of this dataset, which is composed

13. https://github.com/deepinsight/insightface
14. https://gitlab.idiap.ch/bob/bob.bio.face
15. https://www.rankone.io version 1.22.1

only of 223 subjects composed of White and Black men only (where we have 109 white subjects and 114 back subjects). More details about the organization of this evaluation setup can be found on its website[13].

Although dating from 2008, the MORPH dataset is getting some traction recently ( [17], [19], [28]) mostly because of its richness concerning sensitive attributes. It is composed of 55,000 samples from 13,000 subjects from men and women and five race clusters (called ancestry), and they are the following: African, European, Asian, Hispanic, and Others. More details about the organization of this evaluation setup can be found on its website[13].

TABLE 4: MEDS II - ArcFace-InsightFace: FNMR($\tau$), FMR($\tau$), and FDR($\tau$) per demographic (Demog.) in the test set. These figures of merit are fragmented by the race of the samples used for enrollment and the race of the samples used for probe ("(e-p)" in the table.)[*].

| $\tau = FMR_{10^x}$ | $10^{-1}$ | $10^{-2}$ | $10^{-3}$ | $10^{-4}$ | $10^{-5}$ |
|---|---|---|---|---|---|
| Demog (e-p) | FMR$_x(\tau)$ | | | | |
| White - White | 0.106 | 0.008 | 0.001 | 0 | 0 |
| White - Black | 0.058 | 0.002 | 0 | 0 | 0 |
| Black - White | 0.068 | 0.001 | 0 | 0 | 0 |
| Black - Black | 0.162 | 0.026 | 0.003 | 0 | 0 |
| | FNMR$_x(\tau)$ | | | | |
| White - White | 0 | 0 | 0 | 0 | 0.016 |
| Black - Black | 0 | 0 | 0 | 0 | 0 |
| $FDR(\tau)$ | 0.972 | 0.991 | 0.995 | 0.998 | 0.995 |

[*]In this example $\tau = \mathrm{FMR}_x^{\mathrm{dev}}$ where **dev** is the development-set.
FMR($\tau$), FNMR($\tau$) and $FDR(\tau)$ are reported using the test-set

The MOBIO dataset is a video database containing bimodal data (face/speaker). One hundred fifty-two people compose it (split in the two genders **male** and **female**), mostly Europeans, split into five sessions (few weeks time lapse between sessions). The database was recorded using two types of mobile devices: mobile phones (NOKIA N93i) and laptop computers (standard 2008 MacBook). In this paper, we only use mobile phone data. As with other datasets, its evaluation protocol is also published as a python package[16].

### 4.2 Experiments

This section discusses the demographic differentials of the four FR systems using the Fairness Discrepancy Rate. Each one of the following subsections discusses each database in isolation. Both False Matches and False Non-Matches have the same weight; therefore, $\alpha$ is equal to $0.5$. As aforementioned, we will not set a value for $\epsilon$; instead, we will use FDR and Area Under FDR to compare different biometric verification systems concerning demographic discrepancies.

### 4.3 MEDS II database (Demographic differentials concerning race)

Table 4 presents the FMR($\tau$), FNMR($\tau$) and FDR($\tau$) in the test set for the ArcFace-InsightFace system. For the sake of brevity, we present only this system in this comprehensive manner. Please, check the supplementary material to have information about the other systems. In this experiment,

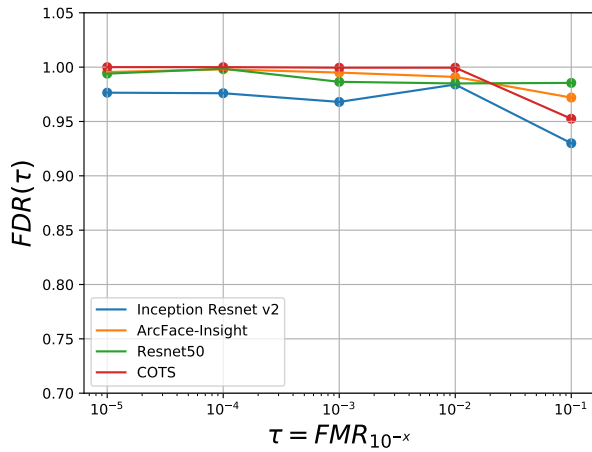16. https://gitlab.idiap.ch/bob/bob.db.mobio

Fig. 7: MEDS II: Fairness Discrepancy Rate of different face verification systems for different decision thresholds

$\tau$ was set at different operational points in the impostor score distribution from an independent set (development set in this case). It is worth noting that such impostor score distribution contains samples from all races, which is the closest scenario to operational conditions, where one single threshold has to be fair to all demographic groups. Both $\text{FMR}_x(\tau)$ and $\text{FNMR}_x(\tau)$ tables are fragmented by demographics (race in this case). Hence, in Table 4, "White - White" means biometric references from White subjects compared with probe samples from White subjects, and so on.

In terms of $\text{FMR}_x(\tau)$ it is possible to observe that for $x = 10^{-1}$ and $x = 10^{-2}$ ($\tau = \text{FMR}_{10^{-1}}$ or $\tau = \text{FMR}_{10^{-2}}$) the face verification system tends to have more false alarms for comparison between biometric references and probes from Black subjects. We can also observe that the number of false alarms using the comparison scores from different races (e.g. "White-Black" and "Black-White") is substantially lower than the number of false alarms using the pair gallery-probe from the same race. This behavior was observed and reported in several publications [15], [20], [33]. In terms of $\text{FNMR}_x(\tau)$, it is possible to notice that such a system tends to reject more White subjects than Black for $x > 10^{-4}$.

Figure 7 presents the FDR plot of the four different biometric systems covering the same decision thresholds showed in Table 4. We can observe that all the four FR systems tend to have more discrepancies between $x = 10^{-1}$ and $x = 10^{-2}$. This reflects the amount of the false alarms raised in these operational points (see Table 4). The system based on Inception-Resnet-v2 presents more discrepancies than the other ones for all operational points. This possibly indicates that the recent state-of-the-art approaches are naturally fairer than the past approaches, even if fairness constraints are not considered in the current approaches (see [35]).

Table 5 (a) presents the Area Under FDR (varying from $10^{-1}$ to $10^{-5}$) of every biometric verification system. With this summarized figure of merit, we can spot some of the trends spotted with Figure 4. Furthermore, we can observe that the current state-of-the-art FR approaches are fairer

than the evaluated COTS.

TABLE 5: Area Under the Fairness Discrepancy Rate for (a) MEDS II, (b) MORPH and (c) MOBIO

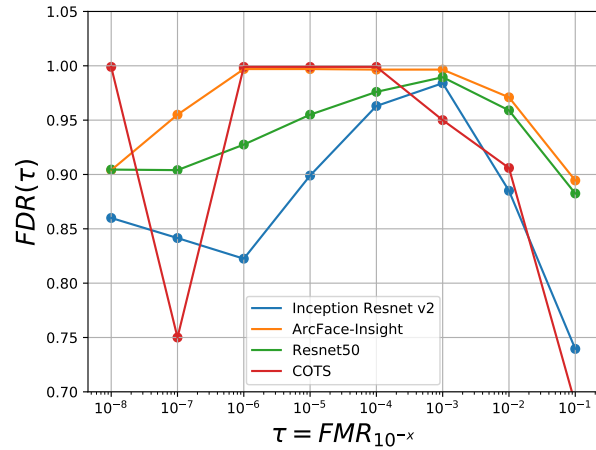| | FDR AUC | | |
|---|---|---|---|
| | (a) MEDS II | (b) MORPH | (c) MOBIO |
| Inception-Resnet-v2 | 0.959 | 0.82 | 0.93 |
| ArcFace-InsightFace | **0.991** | **0.938** | 0.96 |
| Resnet50 | 0.985 | 0.926 | 0.92 |
| COTS | 0.97 | 0.81 | **0.97** |



Fig. 8: MORPH: Fairness Discrepancy Rate of different face verification systems for different decision thresholds

## 4.4 MORPH database (Demographic differentials concerning race)

Table 6 presents the $\text{FMR}(\tau)$, $\text{FNMR}(\tau)$ and $\text{FDR}(\tau)$ in the test set (Male subjects only) for the ArcFace-InsightFace verification system. As with the last section, only this system will be presented in this extensive manner for brevity. In this experiment, $\tau$ was set at different operational points in the impostor score distribution from the development set. Both $\text{FMR}_x(\tau)$ and $\text{FNMR}_x(\tau)$ tables are fragmented by demographics (race in this case) in the same manner as in the previous experiment. However, in this one, we have four demographic groups, and they are the following: Asian, Black, Hispanic, and White (samples labeled as "Others" were left aside).

In terms of $\text{FMR}_x(\tau)$ it is possible to observe that from $x = 10^{-1}$ to $x = 10^{-2}$ (from $\tau = \text{FMR}_{10^{-1}}$ to $\tau = \text{FMR}_{10^{-2}}$) the face verification system tends to have more false alarms for comparisons between biometric references and probes from Hispanic, Asian and Black subjects. Interesting to observe that for $x = 10^{-1}$, a significant amount of false alarms are observed between Asian biometric references with Hispanic Probes and vice-versa. In terms of $\text{FNMR}_x(\tau)$, it is possible to notice that such a system tends to reject more Hispanic subjects from $x \geq 10^{-4}$ and more White subjects from $x \geq 10^{-5}$.

Figure 8 presents the FDR plot of the four different biometric verification systems covering the same decision thresholds showed in Table 6. We can observe that the four

TABLE 6: MORPH - ArcFace-InsightFace: FNMR$(\tau)$, FMR$(\tau)$, and FDR$(\tau)$ per demographic (Demog.) in the test set. These figures of merit are fragmented by the race of the samples used for enrollment and the race of the samples used for probe ("(e-p)" in the table.)[*]

| $\tau = FMR_{10^x}$ | $10^{-1}$ | $10^{-2}$ | $10^{-3}$ | $10^{-4}$ | $10^{-5}$ | $10^{-6}$ | $10^{-7}$ | $10^{-8}$ |
|---|---|---|---|---|---|---|---|---|
| Demog. (e-p) | | | | FMR$_x(\tau)$ | | | | |
| Asian - Asian | 0.314 | 0.066 | 0 | 0 | 0 | 0 | 0 | 0 |
| Asian - Black | 0.076 | 0.003 | 0 | 0 | 0 | 0 | 0 | 0 |
| Asian - Hisp. | 0.108 | 0.009 | 0 | 0 | 0 | 0 | 0 | 0 |
| Asian - White | 0.063 | 0.002 | 0 | 0 | 0 | 0 | 0 | 0 |
| Black - Asian | 0.068 | 0.002 | 0 | 0 | 0 | 0 | 0 | 0 |
| Black - Black | 0.148 | 0.019 | 0 | 0 | 0 | 0 | 0 | 0 |
| Black - Hisp. | 0.067 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Black - White | 0.053 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Hisp. - Asian | 0.102 | 0.006 | 0 | 0 | 0 | 0 | 0 | 0 |
| Hisp. - Black | 0.066 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Hisp. - Hisp. | 0.217 | 0.005 | 0.001 | 0 | 0 | 0 | 0 | 0 |
| Hisp. - White | 0.063 | 0.001 | 0 | 0 | 0 | 0 | 0 | 0 |
| White - Asian | 0.064 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| White - Black | 0.054 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| White - Hisp. | 0.063 | 0. | 0 | 0 | 0 | 0 | 0 | 0 |
| White - White | 0.103 | 0.008 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | | | FNMR$_x(\tau)$ | | | | |
| Asian - Asian | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.182 |
| Black - Black | 0 | 0 | 0 | 0 | 0 | 0.001 | 0.002 | 0.092 |
| Hisp. - Hisp. | 0 | 0 | 0 | 0.006 | 0.006 | 0.006 | 0.006 | 0.116 |
| White - White | 0 | 0 | 0 | 0 | 0.002 | 0.005 | 0.01 | 0.284 |
| $FDR(\tau)$ | 0.8945 | 0.971 | 0.996 | 0.996 | 0.997 | 0.997 | 0.995 | 0.904 |

[*]In this example $\tau = \text{FMR}_x^{\text{dev}}$ where **dev** is the development-set. FMR$(\tau)$, FNMR$(\tau)$ and $FDR(\tau)$ are reported using the test-set

face verification systems present demographic differentials from $x = 10^{-1}$ to $x = 10^{-3}$ mostly due to false matches. From $x = 10^{-3}$ to $x = 10^{-6}$ the ArcFace-InsightFace and the COTS presents very little demographic differentials. Both operate at $FDR(\tau) \geq 0.99$. Demographic differentials from Resnet50 and Inception-Resnet-v2 substantially increase from $x > 10^{-3}$, mostly due to the differences in false non-matches.

Table 5 (b) presents a full picture of the above observations with the Area Under FDR ($x$ varying from $10^{-1}$ to $10^{-8}$)[17] of every biometric verification system. We can observe, in this experiment, that the ArcFace-InsightFace is the fairest FR system, followed by the Resnet50, both based on the ArcFace loss. COTS and the Inception-Resnet-v2 systems present similar $FDR$ ($\approx 0.82$).

### 4.5 MOBIO Database (Demographic differentials concerning gender)

Table 7 presents the FMR$(\tau)$, FNMR$(\tau)$ and FDR$(\tau)$ in the test set for the ArcFace-InsightFace system. MOBIO dataset is composed basically of Caucasians, and for that reason, this experiment focuses on gender differentials only. Hence, FMR$_x(\tau)$ and FNMR$_x(\tau)$ tables are fragmented by gender in the same manner as in the previous experiments. In this setup, $\tau$ is set at different operational points in an independent zero-effort impostor score distribution (from the development set).

In terms of FMR$_x(\tau)$ it is possible to notice that from $x = 10^{-1}$ to $x = 10^{-2}$ (from FMR$_{10^{-1}}(\tau)$ or FMR$_{10^{-2}}(\tau)$) the face verification system tends to have more false alarms for comparison between biometric references and probes from female subjects. We can also observe that the number

TABLE 7: MOBIO - ArcFace-InsightFace: FNMR$(\tau)$, FMR$(\tau)$, and FDR$(\tau)$ per gender in the test set. These figures of merit are fragmented by the gender of the samples used for enrollment and the race of the samples used for probe ("(e-p)" in the table.)[*].

| $\tau = FMR_{10^x}$ | $10^{-1}$ | $10^{-2}$ | $10^{-3}$ | $10^{-4}$ | $10^{-5}$ |
|---|---|---|---|---|---|
| Demog.(e-p) | | | FMR$_x(\tau)$ | | |
| Male-Male | 0.077 | 0.006 | 0 | 0 | 0 |
| Male-Female | 0.052 | 0.001 | 0 | 0 | 0 |
| Female-Male | 0.043 | 0.001 | 0 | 0 | 0 |
| Female-Female | 0.235 | 0.027 | 0.004 | 0 | 0 |
| | | | FNMR$_x(\tau)$ | | |
| Male-Male | 0.001 | 0.001 | 0.001 | 0.001 | 0.002 |
| Female-Female | 0 | 0 | 0 | 0 | 0 |
| $FDR(\tau)$ | 0.9205 | 0.989 | 0.997 | 0.999 | 0.999 |

[*]In this example $\tau = \text{FMR}^{\text{dev}}x$ where **dev** is the development-set. FMR$(\tau)$, FNMR$(\tau)$ and $FDR(\tau)$ are reported using the test-set

of false alarms using the comparison scores from different genders (e.g. "Male-Female" and "Female-Male") is substantially lower than the number of false alarms from comparison scores from the same gender. These are the same trends observed before with race and, as already mentioned, the same trends observed in the literature. In terms of FNMR$_x(\tau)$ such a system tends to slightly reject more Male subjects at $x = 10^{-5}$.

Figure 9 presents the overall picture about demographic differentials using FDR. We can observe that the four face verification systems present demographic differentials from $x = 10^{-1}$ to $x = 10^{-2}$ mostly due to false matches. As in the previous experiment, the ArcFace-InsightFace and the COTS presents very little demographic differentials from $x = 10^{-3}$ to $x = 10^{-5}$; with both operating at $FDR(\tau) \geq 0.99$. The FDR for the systems based on Inception-Resnet-v2 and Resnet50 decreases from $x \geq 10^{-3}$, mostly due to false non-

17. In this experiment, we have enough scores to place a $\tau$ at $10^{-6}$

matches. Finally, Table 5 (c) presents the Area Under FDR. We can observe that, in this experiment, the COTS is the fairest FR system, followed by the ArcFace-InsightFace by a short margin.
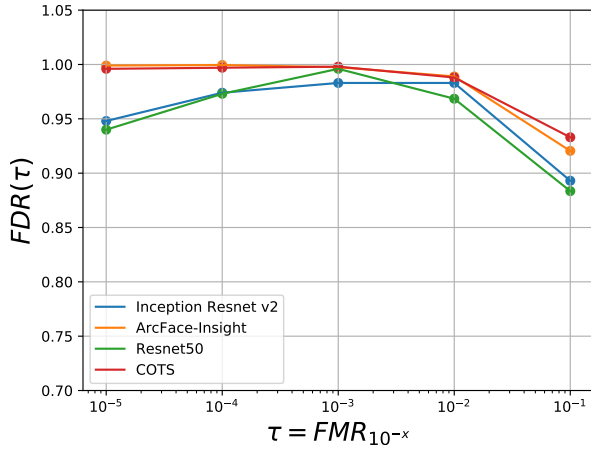


Fig. 9: MOBIO: Fairness Discrepancy Rate of different face verification systems for different decision thresholds

## 4.6 Discussion

This section presented a case study using the proposed Fairness Discrepancy Rate to assess demographic differentials. Experiments with three open-source FR baselines based on DCNNs and one COTS system were used along with three databases where gender and racial differentials were studied. We could notice that FDR could summarize and compare the demographic differentials concerning FMR and FNMR between several FR systems. Furthermore, the Area Under FDR gives a single scalar estimate of such differentials where further rankings can be made.

With this assessment, we could notice that the current state-of-the-art Face Recognition systems based on ArcFace loss are fairer than the state-of-the-art from a few years ago and with the evaluated COTS system. Worth noting that the DCNNs based on ArcFace do not have any fairness constraints. We can hypothesize that the margins imposed in the ArcFace loss play a role in minimizing within-class variability and maximizing between-class separation (embeddings from the same identity tends to be more compact than DCNNs training "vanilla" cross-entropy loss). This possibly impacted the scoring behavior allowing us to "safely" use fair decision thresholds.

We could observe in Table 6 with the MORPH dataset that false alarms are more frequent with Asian, Hispanic, and Black subjects (using operation points around $10^{-2}$). The same trends were observed with female subjects using the Mobio dataset. We can also observe the number of false alarms using the comparison scores between non-homogeneous samples (gallery and probe samples from different demographics) is substantially lower than homogeneous samples (gallery and probe samples from the same demographic groups). This corroborates with the findings of the literature [15], [20], [33].

FDR can be trivially extended to other biometric recognition tasks, such as closed-set or opened-set identification. For instance, for **closed-set** identification, it is possible to compute the differentials between the rank-n [39, chap.21] of different demographic groups as a figure of merit of fairness. For **opened-set**, it is possible to stablish an aggregation figure of merit between Detection and Identification Rate ($\text{DIR}(\tau)$) and False Alarm Rate ($\text{FAR}(\tau)$) for different demographic groups. Further assessment of those extensions for identification will be carried out as future work.

## 5 CONCLUSIONS

This work introduced the Fairness Discrepancy Rate (FDR) to assess demographic differentials in biometric verification systems. FDR tackles a threshold problem, which is the main issue of how the biometric community addresses such differentials. A substantial amount of works in the biometrics community assess demographic differentials in verification systems by comparing DET curves or ROC curves of different demographic groups separately. This type of comparison assumes that decision thresholds are demographic-specific, which is not feasible or ethical in operational conditions. FDR addresses that by assessing demographic differentials assuming single decision thresholds. In this work, we consider that fair biometric recognition systems are fair if a decision threshold $\tau$ is "fair" for all demographic groups concerning $\text{FMR}(\tau)$ and $\text{FNMR}(\tau)$ and FDR proxies this behavior. Furthermore, the $\text{FMR}(\tau)$ and $\text{FNMR}(\tau)$ trade-off with respect to fair behavior can be set by addressing the value $\alpha$ in Equation 6. Finally, the Area Under FDR provides a general overview of demographic differentials under a range of decision thresholds and allows a quick comparison between different biometric verification systems.

Two groups of experiments were carried out to evaluate this new figure of merit. In the first one, a case study using synthetic data was presented, and it was demonstrated how FDR behaves in extreme cases of fair and unfair scenarios. In the second, a case study using four different face verification systems and three databases was carried out. We could observe via the FDR plots that all evaluated face verification systems presents gender and racial biases to some degree. Furthermore, it was possible to quickly compare different face recognition systems concerning their demographic discrepancies using the Area Under FDR. Worth noting that neither FDR nor Area Under FDR is direct proxies for how "accurate" a biometric verification system is. Possible error rates have to be analyzed in parallel to picture the trade-off between accuracy vs. fairness fully.

We also briefly presented a possible trivial extension of such a figure of merit to closed and opened-set identification problems. Further work in this direction will be carried out.

For reproducibility purposes of the work, all the source code, trained models, and recognition scores are made publicly available.

We hope that these tools are useful for the biometrics community to assess demographic differentials, and we advocate for some standardization towards ISO/IEC 19795-10[18]. Hence, the demographic differential can be easily assessed as any other figure of merit, such as FMR or FNMR.

18. https://www.iso.org/standard/81223.html

# REFERENCES

[1] C. O'neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.

[2] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *arXiv preprint arXiv:1602.07261*, 2016.

[3] M. et al., "Bi-modal person recognition on a mobile phone: using mobile phone data," in *2012 IEEE International Conference on Multimedia and Expo Workshops*. IEEE, 2012, pp. 635–640.

[4] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. fairmlbook.org, 2019, http://www.fairmlbook.org.

[5] 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 *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 692–702.

[6] M. Chen and M. Wu, "Towards threshold invariant fair classification," in *Conference on Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 560–569.

[7] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, "The variational fair autoencoder," *arXiv preprint arXiv:1511.00830*, 2015.

[8] F. Locatello *et al.*, "On the fairness of disentangled representations," in *Advances in Neural Information Processing Systems*, 2019, pp. 14 611–14 624.

[9] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International Conference on Machine Learning*, 2013, pp. 325–333.

[10] M. Feldman and othes, "Certifying and removing disparate impact," in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 259–268.

[11] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, 2016, pp. 3315–3323.

[12] Z. et al., "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th international conference on WWW*, 2017, pp. 1171–1180.

[13] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[14] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," *arXiv preprint arXiv:1709.02012*, 2017.

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

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

[17] J. G. Cavazos, P. J. Phillips, C. D. Castillo, and A. J. O'Toole, "Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?" *arXiv preprint arXiv:1912.07398*, 2019.

[18] R. S. Malpass and J. Kravitz, "Recognition for faces of own and other race." *Journal of personality and social psychology*, vol. 13, no. 4, p. 330, 1969.

[19] K. Vangara *et al.*, "Characterizing the variability in face recognition accuracy relative to race," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[20] D. Michalski *et al.*, "The impact of age and threshold variation on facial recognition algorithm performance using images of children," in *2018 International Conference on Biometrics (ICB)*. IEEE, 2018, pp. 217–224.

[21] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 87–102.

[22] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter, "Analyzing and reducing the damage of dataset bias to face recognition with synthetic data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[23] A. Amini, A. P. Soleimany, W. Schwarting, S. N. Bhatia, and D. Rus, "Uncovering and mitigating algorithmic bias through learned latent structure," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 289–295.

[24] E. M. Rudd, M. Günther, and T. E. Boult, "Moon: A mixed objective optimization network for the recognition of facial attributes," in *European Conference on Computer Vision*. Springer, 2016, pp. 19–35.

[25] P. e. a. Terhörst, "Post-comparison mitigation of demographic bias in face recognition using fair score normalization," *Pattern Recognition Letters*, vol. 140, pp. 332–338, 2020.

[26] T. de Freitas Pereira *et al.*, "Periocular biometrics in mobile environment," in *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2015, pp. 1–7.

[27] S. Gong, X. Liu, and A. K. Jain, "Debface: De-biasing face recognition," *arXiv preprint arXiv:1911.08080*, 2019.

[28] 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 Transactions on Technology and Society*, vol. 1, no. 1, pp. 8–20, 2020.

[29] J. P. Robinson, G. Livitz, Y. Henon, C. Qin, Y. Fu, and S. Timoner, "Face recognition: too bias, or not too bias?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 0–1.

[30] M. Wang *et al.*, "Mitigating bias in face recognition using skewness-aware reinforcement learning," in *P. of the IEEE/CVF C. on Computer Vision and Pattern Recognition*, 2020, pp. 9322–9331.

[31] C. M. Cook *et al.*, "Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 1, pp. 32–41, 2019.

[32] N. Poh *et al.*, "Group-specific score normalization for biometric systems," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010, pp. 38–45.

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

[34] T. De Freitas Pereira, "Learning how to recognize faces in heterogeneous environments," EPFL, Tech. Rep., 2019.

[35] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[37] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[38] B. e. a. Maze, "Iarpa janus benchmark-c: Face dataset and protocol," in *2018 International Conference on Biometrics (ICB)*. IEEE, 2018, pp. 158–165.

[39] A. K. Jain and S. Z. Li, *Handbook of face recognition*. Springer, 2011, vol. 1.

**Tiago de Freitas Pereira** received the Ph.D. degree in Electrical Engineering from Ecole Polytechnique Federale de Lausanne ( http://www.epfl.ch) in Switzerland (2019) whose main contribution was the development of the Domain Specific Units (DSU) framework for Heterogeneous Face Recognition. Currently he is a postdoctoral researcher at the Idiap Research Institute in the Biometrics Security and Privacy group and conducts research on face recognition.

**Sébastien Marcel** is the head of the Biometrics group and senior research scientist at the Idiap Research Institute (CH), where he leads a research team and conducts research on biometrics and more particularly face recognition, speaker recognition, vein recognition and spoofing/anti-spoofing. He received the Ph.D. degree in signal processing from Université de Rennes I in France (2000) at CNET, the research center of France Telecom (now Orange Labs). He is currently interested in pattern recognition and machine learning with a focus on multimodal biometric person recognition.