**Appeared in 26th International Conference on Pattern Recognition (ICPR 2022), Fairness in Biometrics Workshop, Montreal, Quebec, August 2022.**

# Evaluating Proposed Fairness Models for Face Recognition Algorithms

John J. Howard*[1], Eli J. Laird*†[1], Rebecca E. Rubin[1], Yevgeniy B. Sirotin*[1], Jerry L. Tipton[1], and Arun R. Vemury[2]

[1] The Identity and Data Sciences Lab at The Maryland Test Facility, Maryland, USA
{jhoward, elaird, rrubin, ysirotin, jtipton}@idslabs.org
[2] The U.S. Department of Homeland Security, Science and Technology Directorate
arun.vemury@dhs.gov

**Abstract.** The accuracy of face recognition algorithms has progressed rapidly due to the onset of deep learning and the widespread availability of training data. Though tests of face recognition algorithm performance indicate yearly performance gains, error rates for many of these systems differ based on the demographic composition of the test set. These "demographic differentials" have raised concerns with regard to the "fairness" of these systems. However, no international standard for measuring fairness in biometric systems yet exists. This paper characterizes two proposed measures of face recognition algorithm fairness (fairness measures) from scientists in the U.S. and Europe, using face recognition error rates disaggregated across race and gender from 126 distinct face recognition algorithms. We find that both methods have mathematical characteristics that make them challenging to interpret when applied to these error rates. To address this, we propose a set of interpretability criteria, termed the Functional Fairness Measure Criteria (FFMC), that outlines a set of properties desirable in a face recognition algorithm fairness measure. We further develop a new fairness measure, the Gini Aggregation Rate for Biometric Equitability (GARBE), and show how, in conjunction with the Pareto optimization, this measure can be used to select among alternative algorithms based on the accuracy/fairness trade-space. Finally, to facilitate the development of fairness measures in the face recognition domain, we have open-sourced our dataset of machine-readable, demographically disaggregated error rates. We believe this is currently the largest open-source dataset of its kind.

**Keywords:** Face Recognition · Fairness · Socio-technical Policy

## 1 Introduction

Facial recognition is the process of identifying individuals using the physiological characteristics of their face [24]. Humans perform such tasks regularly, using dedicated neural pathways that are part of the larger human visual system [10]. In 2014 convolutional neural nets were first applied to the face recognition problem, allowing them to achieve near human performance for the first time [29].

* First authors contributed equally to this research. Authors listed alphabetically.
† Corresponding author.

Subsequently, public facing deployments of face recognition have been increasing steadily. However, there are also long standing reports of face recognition performance varying for people based on their demographic group membership [4, 14, 16, 21, 22, 27]. Of particular concern is the notion that false match rates in face recognition may run higher for certain groups of people, namely African Americans [16, 22].

In response, there has been considerable work around how to train and subsequently demonstrate a "fair" face recognition algorithm [2, 6, 12, 23, 31]. To address the latter, two definitions of "fairness" in face recognition applications were proposed by scientists seeking to quantify the equitability, or lack thereof, of various face recognition algorithms. The first, Fairness Discrepancy Rate (FDR), was proposed by scientists from the Idiap Research Institute, a Swiss artificial intelligence laboratory with a long history of contribution to the field of biometrics [25]. The second, called the Inequity Rate (IR), was proposed by scientists from the U.S. National Institute of Standards and Technology (NIST) [13], a leading scientific body with over 60 years of biometric test and evaluation experience.

However, to date, neither of these techniques has been extensively utilized in practice or audited using a large corpus of actual face recognition error rates. Further, there has been relatively little work to understand the utility of these measures for scoring the fairness of deployed algorithms or for selecting among alternative algorithms during procurement. To address these gaps, we apply these two fairness measures to error rates disaggregated across race and gender demographic groups from 126 commercial and open source face recognition algorithms. We assess their interpretability along three criteria, which we have termed the Functional Fairness Measure Criteria (Section 3.4). Finding no current measure meets all three of these criteria, we developed a new technique based on the Gini coefficient and coined the term the Gini Aggregation Rate for Biometric Equitability, or GARBE (Section 3.5) to describe it. We show how this measure can be used as part of a down-select protocol that also leverages Pareto optimization (Section 3.6). Finally, we discuss the lack of data currently available to developers of fairness measures so that audits of this kind can be executed. As a partial remedy for this, we have open-sourced our dataset of machine-readable, demographically disaggregated error rates. We believe this is currently the largest open-source dataset of its kind.

## 2 Background

### 2.1 Face Recognition

Face recognition algorithms operate by generating numerical representations of faces, referred to as templates. Two face templates can then be compared to produce a similarity score $s$ and if $s$ is greater than some discrimination threshold $\tau$ the corresponding faces are declared to be a "match" by the algorithm. This process can be used in both identification tasks, where an unknown probe face is matched to a gallery of faces, and face verification tasks, where a single face is matched to a claimed identity. The *false match rate* and the *false non-match rate* are two error rates used to measure the foundational accuracy of face

recognition algorithms. The false match rate (FMR) measures the proportion of face comparisons between different identities, or non-mated face pairs, that result in a match. The false non-match rate (FNMR) measures the proportion of face comparisons of the same identity, or mated face pairs, that do not result in a match. FMR and FNMR are specific to a given discrimination threshold $\tau$, which is almost universally set so that FMR $<<$ FNMR. In this paper, we discuss the notion of face recognition fairness with respect to the false match and false non-match rates.

### 2.2 Fairness in Face Recognition

The fairness of software applications in general has garnered much attention in recent years from organizations across a wide swath of disciplines, including computer science, sociology, policy, and others [3, 5, 9, 17, 34]. This focus on algorithmic fairness has been spurred by cases of disparate outcomes for members of different demographic groups in AI-driven software applications. However, until recently there has been relatively little activity on measuring the fairness of face recognition software specifically. One particular challenge in the face recognition domain is that there are numerous ways in which a system can fail, each with different impacts to different users. In the law enforcement use case in particular, a false positive identification has the resulting harm of possible false arrest and imprisonment for a member of the community. A false negative identification, whereby a known suspect in a database is missed, carries the harm of a suspect continuing to be at large in a given community. The favourable outcome in police use of face recognition is therefore a combination of the probability of two distinct error cases, weighted by some social cost of each error case. The fair outcome is that this favourable outcome occurs equally often across demographic groups.

In the absence of other domain specific guidance on fairness, scientists from NIST and the Swiss Idiap Research Institute have proposed two independent measures of fairness with respect to differential error rates. These two methods are known as the Inequity Rate and Fairness Discrepancy Rate, respectively and are discussed in detail in the following sections.

## 3 Methods

### 3.1 Fairness Discrepancy Rate

Fairness Discrepancy Rate (FDR) was proposed by scientists at the Idiap Research Institute, a Swiss artificial intelligence laboratory, in November of 2020 [26]. It was subsequently published in a leading IEEE biometrics journal in August, 2021 [25] as the ".. first figure of merit in this field" and highlights that it "consider[s] the FMR and FNMR trade-off in the demographic differential assessment..". Essentially, this metric advocates for calculating the max difference in false match rate (FMR) and false non-match rate (FNMR) performance between any two demographic groups $d_i$ and $d_j$ and a given discrimination threshold $\tau$. Those differences are then weighed by parameters $\alpha$ and $\beta = 1 - \alpha$, which represent the level of concern applied to differences in FMR and FNMR respectively.

The resulting FDR metric is on a scale of 0 to 1, with 1 being "fair" and 0 being "unfair" [26]. The exact equations for calculating FDR are shown in Equations 3.

$$A(\tau) = max(|FMR_{d_i}(\tau) - FMR_{d_j}(\tau)|) \;\; \forall d_i, d_j \in D \tag{1}$$

$$B(\tau) = max(|FNMR_{d_i}(\tau) - FNMR_{d_j}(\tau)|) \;\; \forall d_i, d_j \in D \tag{2}$$

$$FDR(\tau) = 1 - (\alpha A(\tau) + (1-\alpha)B(\tau)) \tag{3}$$

### 3.2 Inequity Rate

The Inequity Rate (IR) was proposed by scientists at NIST in March of 2021 [13]. Unlike FDR, the IR metric takes ratio differences between min, max FMR and FNMR rates per demographic groups $d_i$ and $d_j$. It then raises these differences to weighing factors $\alpha$ and $(1-\alpha)$ and multiplies the results as shown in Equation 6.

$$A(\tau) = \frac{\max_{d_i} FMR_{d_i}(\tau)}{\min_{d_j} FMR_{d_j}(\tau)} \;\; \forall d_i, d_j \in D \tag{4}$$

$$B(\tau) = \frac{\max_{d_i} FNMR_{d_i}(\tau)}{\min_{d_j} FNMR_{d_j}(\tau)} \;\; \forall d_i, d_j \in D \tag{5}$$

$$IR = A(\tau)^{\alpha} B(\tau)^{1-\alpha} \tag{6}$$

### 3.3 Data

Evaluating the properties of summative measures of face recognition fairness requires data. In the case of the FDR and the IR, the data required must have false match, and non-match rates across demographic groups at a single threshold. We note this is a non-trivial dataset to develop. Most users and developers of face recognition only have access to a small number of algorithms. There are also a limited number of large datasets with ground truth demographic data. The only source (to our knowledge) of this data in a single, consolidated report is the NIST Face Recognition Vendor Test (FRVT) Part 3. The FRVT evaluation is open to face recognition companies and researchers from around the world. Applicants submit their face recognition algorithm packaged in a NIST defined API. NIST then runs these algorithms over several large corpora of face images where the identity of the individuals in the photo is known (VISA photos, MUGSHOT photos, WILD photos, etc.). From these face comparisons, various metrics are produced such as false match and non-match rates at various thresholds.

Part 3 of the FRVT report was released in 2019 and specifically focused on demographic effects [14]. Specifically, Annex 15 of this report contains demographically disaggregated error rates for eight demographic groups, across 126 face recognition algorithms. The demographic groups included in the report consist of two gender groups (Male and Female) paired with four race groups

**Table 1.** Data criteria for summative face recognition fairness metric evaluation

| Criteria | Description |
|---|---|
| C.1 | False match rates |
| C.2 | False non-match rates |
| C.3 | Criteria C.1 and C.2 at a single threshold per algorithm |
| C.4 | Criteria C.1 and C.2 dis-aggregated by demographic group |
| C.5 | Criteria C.1 - C.4 across a representative number of face recognition algorithms |

(American Indian, Asian, Black, and White), resulting in eight gender-race pairs. The discrimination threshold $\tau$ used to calculate error rates in Annex 15 was set to the value that produced a false match rate of $1e^{-4}$. The face pairs used to generate these metrics are derived from a subset of a dataset known as "Mugshots", which contains images of individuals involved in routine U.S. law enforcement booking procedures. Demographic labels are assigned by law enforcement officers and encoded in a record known as the Electronic Biometric Transmission Specification, or EBTS.

For this work, the values contained in NIST FRVT Part 3, Annex 15 were hand transcribed into a machine readable comma separated value file (CSV). This CSV contains 126 columns (one per algorithm) and 17 rows (algorithm name, 8 false match rates, 8 false non-match rates, one per demographic group). We believe this is currently the largest, machine readable collection of disaggregated face recognition error rates. We have made this dataset available at our organizations GitHub page for the benefit of the ML fairness community (see Acknowledgements Section).

### 3.4 Functional Fairness Measure Criteria

One primary objective of any proposed fairness measure is to rank classification algorithms by that measure and select the top or "most fair". We argue this objective is aided when the fairness measure has three properties that make the measure intuitive and more easily reasoned about. These properties are listed below. We collectively refer to these three conditions as the Functional Fairness Measure Criteria, or FFMC.

- FFMC.1 - The net contributions of FMR and FNMR differentials to the overall fairness measure should be intuitive when using a normal range of risk parameter weights and operationally relevant error rates.
- FFMC.2 - There should be recognizable points of reference in the domain of the fairness measure. The easiest way to achieve this objective is to have a bounded fairness measure, with a minimum and maximum possible value.

– FFMC.3 - The fairness measure should be calculable when no errors are observed for a demographic group. Particularly in the context of face recognition, as an increasing number of intersectional demographic groups are considered, the likelihood of experiencing a group with a FNMR of zero also increases. Furthermore, in face recognition if cross group FMR numbers are considered, the likelihood of experiencing a group pair with FMR of zero also rises. The fairness measure should be able to be computed in the presence of either one of these conditions.

### 3.5 The Gini Aggregation Rate for Biometric Equitability

Sections 4.1 and 4.2 examine the properties of the FDR and IR metrics using real, disaggregated face recognition error rates against the FFMC criteria. We find each metric does not fully satisfy the criteria. We thus propose a third fairness aggregation, called the Gini Aggregation Rate for Biometric Equitability (GARBE), inspired by the mathematics of the Gini coefficient. The Gini coefficient is a long-standing measure of statistical dispersion of a set of numbers [11] that is often applied to measure wealth disparity [8]. The formula for the generic Gini coefficient, given $n$ observations of a discrete variable $x$ is shown in Equation 7. For our purposes, we use a variant that normalizes the upper bound of the sample by $\frac{n}{n-1}$. This corrects for downward bias in Gini coefficient calculations when the number of samples is small, as demonstrated in [7].

$$G_x = \left(\frac{n}{n-1}\right)\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \mid x_i - x_j \mid}{2n^2\bar{x}}\right) \forall d_i, d_j \in D \tag{7}$$

Given this definition, a simple extension of the Gini coefficient to the face recognition, or general biometric, use case, taking account risk parameters for weighting the impact of a false match versus false non-match error is shown in Equation 9. We coin the term Gini Aggregation Rate for Biometric Equitability (GARBE) to describe this measure.

$$A(\tau) = G_{FMR_\tau};\ B(\tau) = G_{FNMR_\tau} \tag{8}$$

$$GARBE(\tau) = \alpha A(\tau) + (1-\alpha)B(\tau) \tag{9}$$

One potential drawback to the approach proposed by Equations 7 - 9 is that various studies have documented grouping effects in Gini calculations that can result in underestimation of numeric dispersion [32]. For example, consider calculating the Gini coefficient as shown in Equation 7 on error counts as experienced across three groups, A, B, and C. For the data $x = \{5, 5, 10\}$, the corresponding $G_x = 0.25$. However, were we to combine the error counts for groups A and B such that $x = \{10, 10\}$ the corresponding would $G_x$ would be 0. It therefore becomes possible to "cheat the system" by grouping the data in such a way that minimizes the Gini coefficient, giving an impression of a "more fair" system that would not exist had data been grouped otherwise. To discourage the intentional

use of grouping to bias comparisons involving Gini coefficients, we recommend the specification of grouping variables and group sizes when reporting calculations of the Gini coefficient and derivatives of the metric, such as GARBE.

### 3.6 The Pareto Curve Optimization with Overall Effectiveness

As others have noted, fairness is often part of a trade space with another optimization criteria, accuracy [33, 35]. For example, one way to achieve "fairness" in a face recognition system is to simply declare every face pair as non-matching. Each demographic group would therefore have precisely equal FNMRs (100%) and precisely equal FMRs (0%). While fair, this solution is less than desirable when one also considers the overall performance of the system.

One common technique for optimization around multiple performance criteria in economics and engineering is Pareto efficiency. One can say a pair of performance measures for a solution is Pareto efficient if it satisfies the following condition. Given a set of performance measures $p_1 = \{p_{1,1}, p_{1,2}...p_{1,m}\}$ and $p_2 = \{p_{2,1}, p_{2,2}...p_{2,m}\}$ for $m$ solutions, a pair $\{p_{1,n}, p_{2,n}\}$ is Pareto efficient if both of the following conditions is met:

$$\begin{aligned} p_{1,n} &< p_{1,x} \forall x \in \{1, .., m\} \mid x \neq n \\ p_{2,n} &< p_{2,x} \forall x \in \{1, .., m\} \mid x \neq n \end{aligned} \tag{10}$$

Similarly, if a pair of performance measures satisfies one condition but not the other then we can say this pair is weakly Pareto efficient.

## 4 Results

### 4.1 Properties of Fairness Discrepancy Rate in Practice

When we apply the data described in Section 3.3 to the FDR measure (Section 3.1) we see the distribution of FDR measures as shown in Figure 1. We notice that, despite having a theoretical range of 0 to 1, the practical range of the FDR measure, with the alpha and beta set to 0.5, is closer to 0.9 to 1, with over 95% of FDR values falling in that range. This is a straightforward mathematical extension of the fact that, while the act of aggregating error rates makes sense in principle, for the face recognition problem in particular these error rates almost always exist on vastly different scales. For example, using sample data from NIST FRVT part 3 we see FNMRs ranging from 1.29% to 6.54%. Conversely, the false non-match rates are orders of magnitude smaller, ranging from 0.001% ($1e^{-5}$) to approximately 0.05% ($10^{3.3} = 0.000501$). This is generally true of all face recognition error rates found in our dataset (see Figure 1B, note the log scale of the y axis). This has the effect of limiting the FDR measure, for all practical purposes, to 1 minus the difference in FNMR *only*, hence the practical range from 0.9 to 1.0 (FNMR differences typically vary by <1% to 10%).

Furthermore, this aggregation of error rates that exist on significantly different scales has the extended effect of making the risk parameter $\alpha$ a challenge to configure correctly. Recall from Section 3.1 that alpha is the "weight" of the false

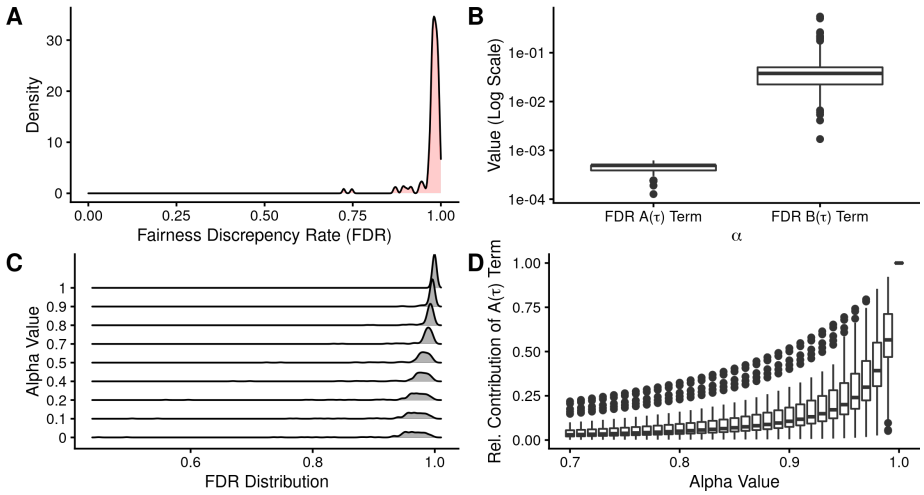


**Fig. 1.** Fairness Discrepancy Rate values using NIST FRVT Part 3 face recognition error rates.**A.** Overall distribution of FDR values (alpha = 0.5). **B** Magnitude of the alpha and beta terms in Equation 3. **C.** Minimum and maxiumum values for FDR given an alpha setting. Note the convergence of the range as alpha increases. **D.** Relative contribution of the alpha term to the overall FDR value. Note the truncated x scale (0.7 - 1.0) and that the median contribution of the alpha term does not surpass 50% until alpha is set to 0.99. Error rates used in FDR calculation are across the eight demographic groups described in Section 3.3.

match discrepancy in the overall FDR calculation. However, because of the small magnitude of FMR differences, these differences only begin to impact the FDR calculation on an equal scale as FNMR differences when alpha is set to greater than 0.99. Indeed, from Figure 1D we see that the median relative contribution of the FMR difference to the FDR only surpasses 50% when alpha is 0.99 and higher.

### 4.2 Properties of Inequity Rate in Practice

Because of the ratio rather than aggregation based summative nature of the Inequity Rate (IR) metric, the issues discussed in Section 4.1 are largely absent. The distribution of IR values at the default alpha of 0.5 spans a range from 2.4 to 26.38 with lower values representing is more "fair" algorithms in this metric system (Figure 2A). The $A(\tau)$ and $B(\tau)$ terms are on more similar scales, with $A(\tau)$ typically having a value in the 40 - 50 range and the $B(\tau)$ term typically ranging from 4 to 9. This more congruous relationship between the $A(\tau)$ and $B(\tau)$ terms means the IR reacts to changes in false match rate weight ($\alpha$) with IR distributions continuing to span representative portions of the metric space at all values of $\alpha$ (Figure 2C) and the $A(\tau)$ term having more of an impact as alpha rises (Figure 2D).

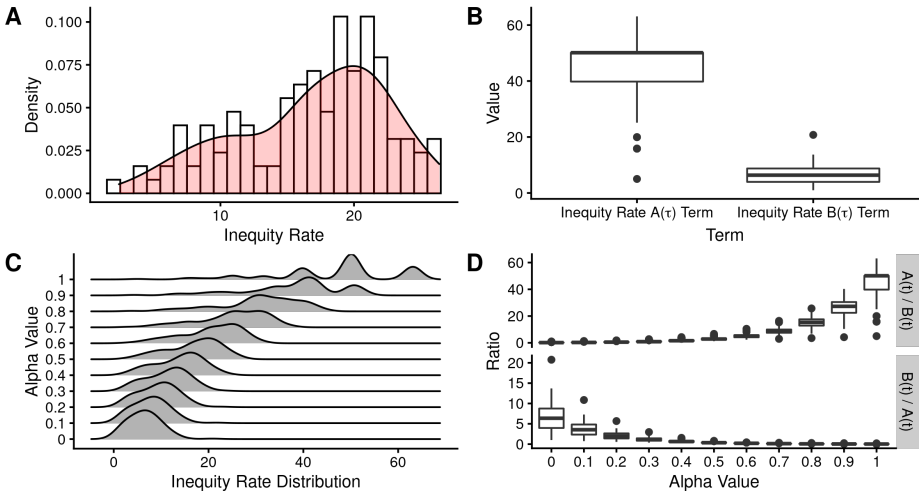


**Fig. 2.** Inequity Rate (IR) values using NIST FRVT Part 3 face recognition error rates. **A.** Overall distribution of IR values (alpha = 0.5). **B** Magnitude of the alpha and beta terms in Equation 6. **C.** Distribution of IR values given an alpha setting. **D.** Relative contribution of the alpha term to the overall IR value. Error rates used in IR calculation are across the eight demographic groups described in Section 3.3.

The only challenge to interpreting IR values that arises from this analysis is the unbounded nature of the metric. Because of its multiplicative nature and the exponential risk weights, there is no theoretical upper bound on the IR measure. Although the practical upper limit in this study was 63.1, different face recognition algorithms could, in theory, give IR results that approach infinity. Similarly, this ratio property also has the drawback of making IR incalculable when the min FNMR or FMR for any group is 0.

### 4.3 Properties of the Gini Aggregation for Biometric Equitability in Practice

The Gini Aggregation for Biometric Equitability (GARBE) measure combines the positive characteristics of the FDR and IR measures. Namely, it's a summative aggregation, meaning the bound can be reasonably controlled but it does not add or subtract error rate values that, in practice, exist on markedly different scales. Instead the GARBE calculates the Gini coeffecient as an approximation to the "spread" or dispersion of these error rates and leverages the fact that the resulting coefficient is already scaled from 0 to 1. This coefficient can then be weighed using the same basic, multiplicative weighing technique utilized in the FDR metric. We see that using a default $\alpha$ of 0.5, GARBE metrics for algorithms in [14] span about half of the theoretically usable range (0.165 - 0.618, Figure 3A). This range continues to span representative portions of the metric space as false match error weight ($\alpha$) is modulated (Figure 3C). We also note that the $A(\tau)$ and $B(\tau)$ terms are the only terms in any of the summative fairness

measures presented here that are scaled to the same order of magnitude, with the median $A(\tau)$ value found at 0.74 and the median $B(\tau)$ at 0.33 (Figure 3B). Finally, because of the consistent scaling of the Gini coefficient calculation, the relative contribution of the $A(\tau)$ term to the overall GARBE metric increases approximately linearly as alpha increases (Figure 2D), with the mean contribution of $A(\tau)$ surpassing the contribution of $B(\tau)$ when $\alpha = 0.4$. Contrast this with Figure 1D where the mean contribution of $A(\tau)$ did not surpass 0.5 until $\alpha = 0.99$.

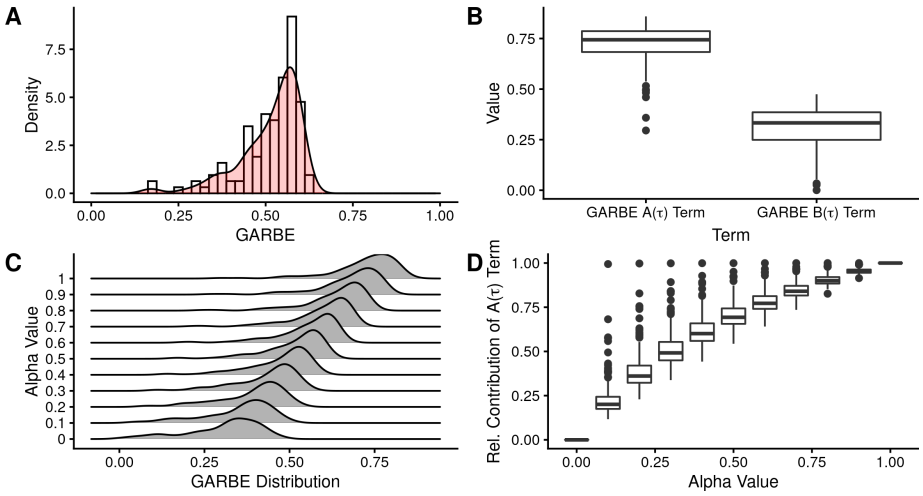


**Fig. 3.** GARBE values using NIST FRVT Part 3 face recognition error rates. **A.** Overall distribution of GARBE values (alpha = 0.5). **B** Magnitude of the alpha and beta terms in Equation 9. **C.** Distribution of GARBE values given an alpha setting. **D.** Relative contribution of the alpha term to the overall GARBE value. Error rates used in GARBE calculation are across the eight demographic groups described in Section 3.3.

### 4.4 In Summary of Summative Fairness Measures

Because the FDR metric is bounded, we find it is possible to create reference points in its domain. For example, a perfectly fair algorithm (no differences in group based FNMR or FMR) has a FDR of 1 and an perfectly unfair algorithm (all FNMR or FMR occurring for one group) has a FDR of 0. FDR is also calculable in the presence of zero percent FNMR or FMR. However, the FDR measure's differential terms exist at vastly different scales when using a normal range of risk parameters and operationally relevant error rates (Figure 1B). In face recognition deployments where false match rate differences across group are of concern, the FDR *alpha* term should be set on the scale from $(0.99, 1]$ in order to allow the contributions from the $A(\tau)$ term to contribute to the overall FDR measure. This is not documented anywhere outside this audit but is an

important point should the FDR measure be used to select fair face recognition algorithms in practice.

The IR fairness measure largely rectifies the scaling issues encountered with FDR measure by taking a ratio as opposed to minmax aggregation of FMR and FNMR numbers. This results in the IR measure having a dynamic range that spans from a supposed minimum of 1 ("fair" algorithm) to a practical maximum of 63.1, in this study. Furthermore, the contribution from $A(\tau)$ and $B(\tau)$ are on relatively similar scales when alpha values are set to normal ranges. However, also because of this ratio aggregation, the IR measure can approach $\infty$ as $\min_{d_j} \mathrm{FNMR}_{d_j}(\tau)$ or $\min_{d_j} \mathrm{FMR}_{d_j}(\tau)$ approaches 0 and is indeed incalculable should one of these rates reach 0. Its also challenging to interpret and compare IR values both within and across studies. Because of the unbounded nature of the measure, the most direct approach to establishing a "fair" algorithm is to partition the IR space and select algorithms in the Nth quartile. However, this quartile can shift from study to study, depending on the minimum FNMR and FMR's per group encountered. This makes comparing IR values a challenge should the IR measure be used to select fair face recognition algorithm in practice.

Finally, the GARBE fairness measure, proposed in this study, builds on the strength of the FDR and IR measures. Instead of aggregating minmax FNMR and FMR differences, the GARBE measure weighs and aggregates measures of dispersion of these error rates, namely the Gini coefficient (Equation 7). This has several advantages. One, this measure is calculable in the presence of error rates being 0. Second, this measure first converts two sets of numbers that exist on markedly different scales to a single common metric space before weighing and aggregating. In this fashion we can both avoid the poor relationship between risk ratios and relative contribution of $A(\tau)$ and $B(\tau)$ terms (Figure 1C-D and 3C-D) and retain a bounded domain (Figure 2A &C and 3A & C). Because of these properties, the GARBE measure is able to satisfy all the FFMC criteria. Table 2 summarizes the three fairness measures with respect to the FFMC criteria.

**Table 2.** Summary of Summative Fairness Measures

| **FFMC Criteria** | **FDR** | **IR** | **GARBE** |
|---|---|---|---|
| FFMC.1 | | ✓ | ✓ |
| FFMC.2 | ✓ | | ✓ |
| FFMC.3 | ✓ | | ✓ |

### 4.5 Pareto Curve Optimization with the Gini Aggregation Rate for Biometric Equitability

Finally, this study advocates for evaluating face recognition algorithms along multiple axes of performance, namely overall effectiveness and fairness, using the Pareto curve method (Section 3.6). This technique requires computing both

overall effectiveness and a fairness measure. We select the GARBE fairness measure for the reasons outlined in Section 4.4. As a measure of overall performance we select total FNMR across all demographic groups. This value is a weighted average of the error rates as reported in Annex 15 (Section 3.3) and the mated comparison counts, also provided in the introductory material to Annex 15 of [14].

This result is shown in Figure 4, with overall performance (FNMR) plotted on the x axis and the GARBE fairness measure plotted on the y axis. Each point represents an algorithm, while the Pareto efficient algorithms are connected with a red line and have their names printed. We note the Pareto frontier provides a perceptive means of down-selecting which algorithms should be considered in this optimization space. Any algorithm not on the Pareto frontier can be discarded, as there exists another selection that is either mathematically more fair or better performing. This effectively reduces the search space for the "optimal" algorithm from the 126 algorithms tested in [14] to the 9 on the Pareto frontier, a savings of over 90%. Additionally, if we further refine our search to algorithms that had very good performance overall, we only have to consider the six algorithms in the inset of Figure 4. Algorithm didiglobalface-001 is the highest performing in this space, having achieved the lowest overall FNMR of $\sim$0.0022. However, it is also the least fair of the Pareto efficient set, having achieved the highest GARBE measure of $\sim$0.54. Conversely, algorithm intellifusion-001 was the least performative of this set, with a total FNMR of $\sim$0.0038, but it also had a somewhat improved GARBE fairness measure at $\sim$0.37. Whether this trade-off of a 0.0016 increase in total performance is worth a decline in fairness of 0.17 is a question that can be posed to system designers. However, the Pareto curve, frontier, and process we have outlined here allow this trade-space to be explored effectively.

## 5 Discussion

In this study we have executed the first audit of two proposed face recognition fairness measures using demographically disaggregated false match and false non-match error rates from 126 commercial and open source algorithms. We've found that both proposed models have benefits and drawbacks when it comes to interpreting their outcomes on face recognition error rates commonly found in practice. We've attempted to consolidate the benefits of each approach into a set of interpretability criteria, called the FFMC, and hope these can serve as a guide for future development of fairness measures, particularly in the face recognition domain. We've also proposed an alternative fairness measure, the Gini Aggregation Rate for Biometric Equality or GARBE that satisfies all of these criteria and demonstrated a protocol using Pareto efficiency that can rapidly identify optimal algorithms in both the overall performance and fairness domains. The main takeaways and areas of future work are delineated below.

### 5.1 Audit the Audit

As discussed in Section 2.2, there are currently a plethora of definitions for both bias sources and fairness measures propagating throughout the ML fairness

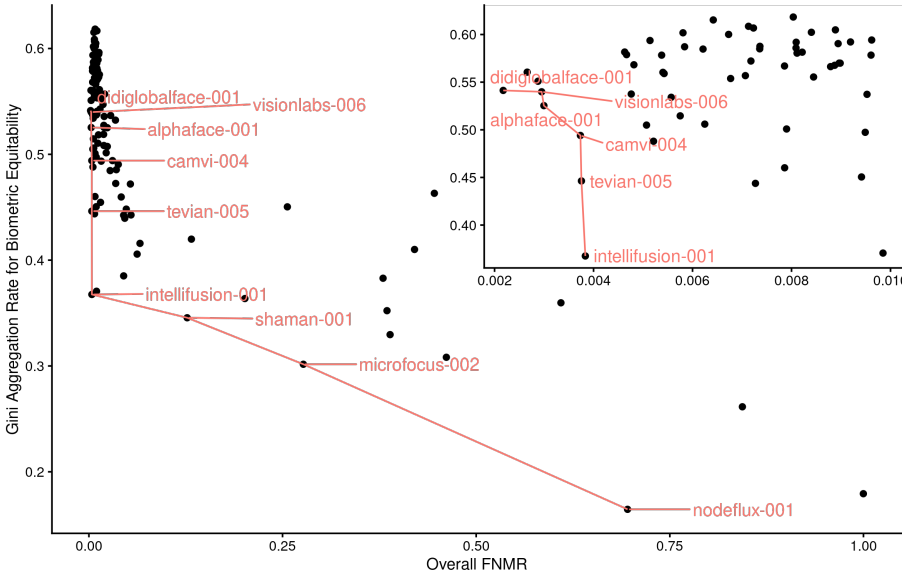


**Fig. 4.** Pareto curve of Gini Aggregation Rate for Biometric Equitability (GARBE) values plotted against overall performance (Total FNMR) using NIST FRVT Part 3 face recognition error rates. Red line connects algorithms that are Pareto efficient. No algorithms are weakly Pareto efficient. Inset shows zoom1ed area where total FNMR performance is less than 1%

space. This increased attention is a positive development. However, in such an environment, it is critical to evaluate the merits of different approaches when applying a given technique to a specific use case. Often times, when analysing the positive outcome in a specific application of a ML decision system, there is not one specific failure case that can cause harm but a set, which requires the aggregation of different error probabilities into a new metric, as we have shown here. As new metrics are developed and proposed, purveyors and evaluators of ML algorithms should strive to ensure the statistical properties of their proposed methods are well documented via the kind of audit we have performed here. In this way they can be of maximum utility to the broader ML fairness community.

### 5.2 On the Need for Additional Fairness Data

One obvious yet often illusive requirement for auditing fairness measures in any domain is data. This study documented a set of criteria necessary for the evaluation of face recognition fairness measures in particular (Table 1). Access to data of this nature is a necessary for auditing fairness measures, yet datasets of this nature are limited at best. We have attempted to provide one such dataset by open sourcing the error rates used in this research. However, even this dataset has certain drawbacks. For example, our dataset only shows error rates at a single population-wide FMR threshold. FNMR and FMR measures across a range

of representative thresholds would allow for a more complete investigation. Additionally, our dataset only includes intra-demographic false match error rates (Male-to-Male and Female-to-Female, for example). However, equatability in the face recognition domain may depend on *inter*-demographic false match rates (Male-to-Female, etc.) [15]. To promote future work in this area, evaluators of face recognition algorithms should consider making more robust datasets available to the community in a readily parsable format.

### 5.3 Limitations of Mathematical Formulations of Fairness

Finally, we conclude with a discussion of the general term "fairness" in relation to the kind of mathematical audits we have performed here. As others have noted, fairness is a broad concept without a concise definition [1, 30]. Additionally, as observed by individuals, fairness is not primarily a mathematical construct but a social and perceptual one [18, 19]. We've used the term "fairness measure" as have others in the sense that these metrics relate to the *topic* of fairness, as they are used to reason about differential error rates. However, one area that is currently under-researched in the ML fairness community is how mathematical notions of fairness translate to perceptual notions of fairness. Human perception is often nonlinear and we have accounted for these non-linearities in measurements of physical intensity (e.g. light and sound [28]) and in economic models [20]. Furthermore, if a system has precisely equal odds that a privileged and unprivileged group will receive a positive outcome in a given fairness space (e.g. a disparate impact of 1), does a human observing this system operate perceive it to be fair? There very well may be entire classes of AI systems, face recognition included, that regardless of their performance may be perceived as unfair in some applications. Should this be the case, then, despite current consensus in the literature, the term "fairness" may not be appropriate for describing the class of metrics that deal more narrowly with differential performance of the system rather than the perceptual fairness of a particular application of the system. We think studies to understand human perception of fairness will help bridge current gaps between notions of mathematical fairness based on accuracy and social/perceptual fairness in the ML fairness community.

## Acknowlegments

This research was funded by the Department of Homeland Security, Science and Technology Directorate (DHS S&T) on contract number W911NF-13-D-0006-0003. The views presented do not represent those of DHS, the U.S. Government, or the author's employers.

The dataset used in this report is available on the Maryland Test Facility's github: https://github.com/TheMdTF/mdtf-public/tree/master/datasets/nist-frvt-annex15.

Paper contributions: All authors conceived the work; Eli J. Laird and John J. Howard performed the statistical analysis and wrote the paper; Yevgeniy B. Sirotin advised on statistical analysis and edited the paper. Jerry L. Tipton (IDSLabs) and Arun R. Vemury (DHS S&T) also conceived the work.

## References


1. Barocas, S., Hardt, M., Narayanan, A.: Fairness and Machine Learning. fairmlbook.org (2019), http://www.fairmlbook.org
2. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. pp. 77–91 (2018)
3. Cavazos, J.G., Phillips, P.J., Castillo, C.D., O'Toole, A.J.: Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? IEEE Transactions on Biometrics, Behavior, and Identity Science **3**(1), 101–111 (2021). https://doi.org/10.1109/TBIOM.2020.3027269
4. Cook, C.M., Howard, J.J., Sirotin, Y.B., Tipton, J.L., Vemury, A.R.: Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems. Transactions on Biometrics, Behavior, and Identity Science **1**(1) (2019)
5. Danks, D., London, A.J.: Algorithmic bias in autonomous systems. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. p. 4691–4697. IJCAI'17, AAAI Press (2017)
6. Das, A., Dantcheva, A., Brémond, F.: Mitigating bias in gender, age and ethnicity classification: A multi-task convolution neural network approach. In: ECCV Workshops (2018)
7. Deltas, G.: The small-sample bias of the gini coefficient: Results and implications for empirical research. The Review of Economics and Statistics **85**(1), 226–234 (2003), http://www.jstor.org/stable/3211637
8. Department of Economic and Social Affairs: World economic sitatuon and prospects, monthly briefing. United Nations (2018), https://www.un.org/development/desa/dpad/tag/gini-coefficient/
9. Drozdowski, P., Rathgeb, C., Dantcheva, A., Damer, N., Busch, C.: Demographic bias in biometrics: A survey on an emerging challenge. vol. 1 (03 2020). https://doi.org/10.1109/TTS.2020.2992344
10. Freiwald, W., Duchaine, B., Yovel, G.: Face processing systems: from neurons to real-world social perception. Annual Review of Neuroscience **39**, 325–346 (2016)
11. Gini, C.: Variabilità e mutabilità. Reprinted in Memorie di metodologica statistica (Ed. Pizetti E (1912)
12. Gong, S., Liu, X., Jain, A.K.: Jointly De-Biasing Face Recognition and Demographic Attribute Estimation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **12374 LNCS**, 330–347 (2020). https://doi.org/10.1007/978-3-030-58526-6_20, iSBN: 9783030585259
13. Grother, P.: Demographic differentials in face recognition algorithms. EAB Virtual Event Series - Demographic Fairness in Biometric Systems (2021)
14. Grother, P., Ngan, M., Hanaoka, K.: Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. Tech. rep., United States National Institute of Standards and Technology (2019)
15. Howard, J.J., Sirotin, Y.B., Tipton, J.L., Vemury, A.R.: Quantifying the extent to which race and gender features determine identity in commercial face recognition algorithms. Tech. rep., United States Department of Homeland Security, Science and Technology Directorate, Technical Paper Series (2021)
16. Howard, J.J., Sirotin, Y.B., Vemury, A.R.: The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face

recognition algorithm performance. In: 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS). pp. 1–8. IEEE (2019)
17. Hutchinson, B., Mitchell, M.: 50 years of test (un)fairness: Lessons for machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. p. 49–58. FAT* '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3287560.3287600, https://doi.org/10.1145/3287560.3287600
18. Jacobs, A.Z., Wallach, H.: Measurement and fairness. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. pp. 375–385 (2021)
19. Kahneman, D., Knetsch, J.L., Thaler, R.: Fairness as a constraint on profit seeking: Entitlements in the market. The American economic review pp. 728–741 (1986)
20. Kahneman, D., Tversky, A.: Prospect theory: An analysis of decision under risk. In: Handbook of the fundamentals of financial decision making: Part I, pp. 99–127. World Scientific (2013)
21. Klare, B.F., Burge, M.J., Klontz, J.C., Bruegge, R.W.V., Jain, A.K.: Face recognition performance: Role of demographic information. IEEE Transactions on Information Forensics and Security **7**(6), 1789–1801 (2012)
22. Krishnapriya, K., Albiero, V., Vangara, K., King, M.C., Bowyer, K.W.: Issues related to face recognition accuracy varying based on race and skin tone. IEEE Transactions on Technology and Society **1**(1), 8–20 (2020)
23. Liu, B., Deng, W., Zhong, Y., Wang, M., Hu, J., Tao, X., Huang, Y.: Fair loss: Margin-aware reinforcement learning for deep face recognition. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10051–10060 (2019). https://doi.org/10.1109/ICCV.2019.01015
24. Organization, I.S., Commission, I.E.: ISO/IEC 2382-37:2015: Information technology - vocabulary - part 37: Biometrics. ISO/IEC, Editor (2015)
25. Pereira, T.d.F., Marcel, S.: Fairness in biometrics: a figure of merit to assess biometric verification systems. IEEE Transactions on Biometrics, Behavior, and Identity Science pp. 1–1 (2021). https://doi.org/10.1109/TBIOM.2021.3102862
26. Pereira, T.d.F., Marcel, S.: Fairness in biometrics: a figure of merit to assess biometric verification systems. arXiv preprint arXiv:2011.02395 (2020)
27. Phillips, P.J., Jiang, F., Narvekar, A., Ayyad, J., O'Toole, A.J.: An other-race effect for face recognition algorithms. ACM Transactions on Applied Perception (TAP) **8**(2), 1–11 (2011)
28. Raub, M.: Bots, bias and big data: artificial intelligence, algorithmic bias and disparate impact liability in hiring practices. Ark. L. Rev. **71**, 529 (2018)
29. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1701–1708 (2014)
30. Verma, S., Rubin, J.: Fairness definitions explained. In: 2018 ieee/acm international workshop on software fairness (fairware). pp. 1–7. IEEE (2018)
31. Wang, T., Zhao, J., Yatskar, M., Chang, K.W., Ordonez, V.: Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
32. Warrens, M.: On the negative bias of the gini coefficient due to grouping. Journal of Classification **35** (10 2018). https://doi.org/10.1007/s00357-018-9267-9
33. Wei, S., Niethammer, M.: The fairness-accuracy pareto front. arXiv preprint arXiv:2008.10797 (2020)

34. Wilson, C., Ghosh, A., Jiang, S., Mislove, A., Baker, L., Szary, J., Trindel, K., Polli, F.: Building and auditing fair algorithms: A case study in candidate screening. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. p. 666–677. FAccT '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3442188.3445928, https://doi.org/10.1145/3442188.3445928
35. Zafar, M.B., Valera, I., Rogriguez, M.G., Gummadi, K.P.: Fairness constraints: Mechanisms for fair classification. In: Artificial Intelligence and Statistics. pp. 962–970. PMLR (2017)