2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Sum of Group Error Differences: A Critical Examination of Bias Evaluation in Biometric Verification and a Dual-Metric Measure

Alaa Elobaid [1,2], Nathan Ramoly [3], Lara Younes [3], Symeon Papadopoulos [1], Eirini Ntoutsi [4], and Ioannis Kompatsiaris[1]
[1] Information Technologies Institute, CERTH, Thessaloniki, Greece
[2] Freie Universität Berlin, Berlin, Germany
[3] IDnow, Research CoE, Cesson-Sévigné, France
[4] Research Institute CODE, Bundeswehr University Munich, Munich, Germany
Email: alaa.elobaid@fu-berlin.de

***Abstract*— Biometric Verification (BV) systems often exhibit accuracy disparities across different demographic groups, leading to biases in BV applications. Assessing and quantifying these biases is essential for ensuring the fairness of BV systems. However, existing bias evaluation metrics in BV have limitations, such as focusing exclusively on match or non-match error rates, overlooking bias on demographic groups with performance levels falling between the best and worst performance levels, and neglecting the magnitude of the bias present.**

This paper presents an in-depth analysis of the limitations of current bias evaluation metrics in BV and, through experimental analysis, demonstrates their contextual suitability, merits, and limitations. Additionally, it introduces a novel general-purpose bias evaluation measure for BV, the "Sum of Group Error Differences (SED_G)". Our experimental results on controlled synthetic datasets demonstrate the effectiveness of demographic bias quantification when using existing metrics and our own proposed measure. We discuss the applicability of the bias evaluation metrics in a set of simulated demographic bias scenarios and provide scenario-based metric recommendations. Our code is publicly available under https://github.com/alaaobeid/SEDG.

I. Introduction

Biometric Verification (BV) systems suffer from demographic biases that manifest in the form of different accuracy levels influenced by demographic factors such as skin colour, age, and gender [7], [3], [1]. These biases are transferred to real-world BV applications, such as unconstrained face verification [13] in mobile phone access [11], biometric authentication [2] in identity-to-selfie matching [17], and identity verification [15] in border control [5]. However, there is no agreement over which metrics to use to quantify the amount of bias in a BV system [16]. Nevertheless, current bias evaluation methods and metrics suffer from at least one limitation at a time. Differential performance methods such as the fairness index measures [10], and the analysis of Receiver Operator Characteristics (ROC), False Match Rate (FMR), and False Non-Match Rate (FNMR) curves [12], [19] do not reflect real-world scenarios where a decision threshold is required. Methods that do not account for global performance and cross-demographic impostors such as the genuine and impostor distributions [19], the demographic group verification accuracies [6], [22], the Standard Deviation (STD) in Group Equal Error Rates (EER_G) [16], and the Gini Aggregation Rate for Biometric Equitability (GARBE) [8], lack a global reference point for measuring bias and neglect the scenario where the impostor does not share demographic attributes with the genuine identity. The absence of a reference point for bias measurement can make it impossible to know the magnitude of the bias. Metrics that quantify bias in the form of maximum and minimum error rates, such as the Skewed Error Rate (SER) [20], the Inequity Rate (IR) [8] and the Fairness Discrepancy Ratio (FDR) [4] disregard demographic groups that exhibit intermediate error rates. Hence, a system biased against one group is given the same score as a system biased against multiple groups as long as their minimum and maximum error values are identical. Additionally, most of these metrics require that an operational scenario is defined beforehand and may not be suitable for studying biases in systems in a manner that is independent of application.

This paper aims to highlight the limitations of current bias evaluation metrics in BV, evaluate their suitability and merits, and introduce a general-purpose bias evaluation measure that overcomes the identified limitations. The proposed bias evaluation measure captures deviations of demographic group error rates from global error rates, treats false matches and non-matches equally, and considers all demographic groups while calculating errors and when setting a suitable threshold. Our main contributions are as follows:

- We provide an analysis of bias evaluation metrics in the BV literature and identify their limitations.
- We evaluate the effectiveness and behaviour of the bias evaluation metrics through experimental analysis in a range of simulated scenarios with different levels and types of bias.
- We provide a detailed analysis of the strengths, weaknesses, and appropriateness of each metric in different bias scenarios.
- We propose a novel dual-metric bias evaluation measure that overcomes the limitations of existing bias evaluation metrics in BV.

II. Background and Related Work

We first introduce the basics of generic Biometric Verification (BV) before discussing related research on bias and

discrimination in BV.

A. Biometric Verification Problem Formulation

The generic BV problem is an application-independent formulation of various BV tasks that use an individual's biological traits to verify their identity. Therefore, it applies to various biometric data types, such as face images, fingerprints, speech, and iris data. BV is typically considered as a binary classification task, where the goal is to determine whether a pair of inputs is a **genuine pair**, meaning that the input pair genuinely belongs to the same person, or an **impostor pair** meaning that the pair belongs to different persons. In some cases, impostor pairs are selected to share the same demographic attributes. These are referred to as Within-Demographic Impostors (WDI) in contrast to Cross-Demographic Impostors (CDI), which do not necessarily share any demographic attributes.

More formally, the BV problem can be expressed as follows: Given a pair of biometric data samples s_1 and s_2, a biometric feature extraction model M is used to extract two embedding vectors e_1 and e_2, which represent these samples and the distance between the pair of embedding vectors in our case the Euclidean Distance (ED) is then calculated as follows:

$$e_1 = M(s_1)$$

$$e_2 = M(s_2)$$

$$\text{ED } (e_1, e_2) = \sqrt{\sum_{i=1}^{n} (e_{1i} - e_{2i})^2}$$

The goal of the generic BV problem is to determine whether the pair of samples s_1 and s_2, which are being represented by the embedding vectors (e_1, e_2), belong to the same individual or not. It can be expressed as follows:

$$D = \begin{cases} \text{True}, & \text{if ED } (e_1, e_2) \text{ is less than or equal to } T. \\ \text{False}, & \text{otherwise.} \end{cases}$$

where:

- D is the decision indicating whether the pair of sample embeddings (e_1, e_2) belong to the same individual or not.
- ED (e_1, e_2) is the Euclidian distance between the pair of biometric sample embedding vectors e1 and e2.
- T is the decision threshold used for verification.

In this problem, errors are usually quantified in terms of False Match Rates (FMR) and False Non-Match Rates (FNMR), which can be formulated as follows:

$$\text{FMR} = \frac{\text{Number of false matches}}{\text{Total number of impostor pairs}} \tag{1}$$

$$\text{FNMR} = \frac{\text{Number of false non-matches}}{\text{Total Number of genuine pairs}} \tag{2}$$

where:

- False matches are impostor pairs falsely flagged as genuine pairs.
- False non-matches are genuine pairs falsely flagged as impostor pairs.

Here, we define the True Match Rate (TMR) and the True Non-Match Rate (TNMR) as they are both used in our dataset generation process described in Section III.

$$\text{TMR} = \frac{\text{Number of true matches}}{\text{Total Number of genuine pairs}} \tag{3}$$

where:

- True matches are genuine pairs correctly flagged as genuine pairs.

TMR can also be expressed in terms of FNMR as:

$$\text{TMR} = 1 - \text{FNMR} \tag{4}$$

Similarly, the TNMR can be expressed in terms of FMR as:

$$\text{TNMR} = 1 - \text{FMR} \tag{5}$$

B. Bias and Discrimination in Biometric Verification

Bias in BV can be studied independently of any decision thresholds, referred to as Differential Performance, or by studying error rates at a decision threshold with a desired performance, referred to as Differential Outcomes. Both of those terms were introduced by [9]. Bias evaluation in terms of differential performance studies the differences in ROC curves or genuine-impostor distributions of the different demographic groups [12], [19], [10] independent of application. However, the practicality of current differential performance approaches is questionable because, in real-world operational scenarios, a threshold needs to be determined in advance. Taking this into consideration, differential outcome methods are, by design, more representative of real-world verification scenarios and, therefore, are the focus of our paper.

In the remaining part of this section, we discuss some of the shortcomings of current differential outcome methods.

Standard deviation-based measures: The use of the STD in demographic group-specific error rates computed on WDI such as the EER [16], TMR [12], [19], [6], [22], FMR [20], [21], [14], and FNMR [18] has been widely observed. However, by not accounting for CDIs, such approaches fail to reflect real-world scenarios where impostors may not always share demographic attributes with the genuine pairs. Additionally, quantifying biases in the form of the STD in group error rates obscures the magnitude of the bias. In other words, not using the global performance on WDIs and CDIs as a reference leads to loss of information about the significance of the bias. For instance, a model that performs three times worse than the global performance on a set of demographic groups would be treated the same as a model that performs five times worse than the global performance on the same set of groups. Depending on the fairness context, such behaviour may be undesirable.

Maximum disparity-based measures: In some metrics, bias is studied in terms of maximum discrepancy in error rates, such as the ratio of maximum and minimum values of FMRs in SER [20] or FMRs and FNMRs in IR [8] or the maximum absolute difference in FMRs and FNMRs in

Method	C1	C2	C3	L1	L2	L3
STD in FMR [20], [21], [14]	✓	✗	✗	✓	✓	✓
STD in FNMR [18]	✓	✗	✗	✓	✓	✓
STD in TMR [12], [19], [6], [22]	✓	✗	✗	✓	✓	✓
STD in EER_G [16]	✓	✗	✗	✗	✓	✗
IR [8]	✗	✓	✗	✗	✓	✓
FDR [4]	✗	✓	✗	✗	✓	✓
SER [20]	✗	✓	✗	✓	✓	✓
MAPE [20]	✗	✗	✓	✓	✗	✓
GARBE [8]	✗	✗	✓	✗	✓	✓
SED_G (Our method)	✓	✗	✓	✗	✗	✗

C1: Standard deviation-based measure
C2: Maximum disparity-based measure
C3: Summative aggregation-based measure
L1: Accounts for a single error type
L2: Exclusively uses WDI pairs, Lacks a global performance reference
L3: Requires a pre-defined policy FMR

TABLE I
SUMMARY OF LIMITATIONS OF EXISTING DIFFERENTIAL OUTCOME METHODS

the FDR [4]. Quantifying bias in the form of maximum and minimum error rates obscures a model's bias and performance on demographic groups with intermediate error rates. As a result, a model biased against multiple demographic groups may be assigned the same SER, IR, or FDR score as a model biased against a single demographic group. Since all three maximum-discrepancy-based metrics quantify bias in the form of WDI performance, in practice, they suffer from the same limitation of not capturing the magnitude of the bias as the STD-based metrics because both sets of metrics lack a global reference point for measuring bias.

Summative aggregation-based measures: Mean Average Percent Error (MAPE) [20] and GARBE [8] are two metrics that quantify bias as the average of the sum of the absolute difference in group performance from a reference performance. MAPE uses a global performance consisting of CDIs and WDIs as reference, while GARBE measures how much each group's performance differs on average from all other group performances using only WDIs. MAPE accounts exclusively for FMRs, while GARBE accounts for both FMRs and FNMRs. Therefore, each metric has its own strengths and can be useful in different scenarios.

Single error type measures: It is common for differential outcome methods to measure the differences in FMRs [20], [21], [14], FNMRs [18], or TMRs [12], [19], [6], [22] for the different demographic groups and treat them as bias. However, neglecting the other set of errors (respectively FNMRs, FMRs, or TNMRs) can lead to missing potential biases in those errors. As a result, bias evaluation metrics that account for FMRs only, such as SER [20] and MAPE [20], only partially capture the overall bias in a system. This means that, in theory, if a model suffers exclusively from demographic biases in the form of false non-match errors, no bias would be captured by those metrics.

We summarize the limitations of existing differential outcome methods in Table I. We point out that most methods, except for STD in EER_G and our proposed measure, require a policy FMR based on application or operational requirements. Hence, we see a need for a comprehensive application-independent metric that does not require any predefined operational threshold for bias quantification.

In conclusion, existing bias evaluation metrics in BV can only partially capture bias due to the limitations mentioned before. This emphasizes the need for a more comprehensive application-independent bias evaluation metric that simultaneously accounts for match and non-match errors, global CDI and WDI performance, and intermediate error rates.

III. Demographic Bias Simulation and the Sum of Group Error Differences

In this section, we describe our demographic bias simulation, synthetic dataset generation process, and our proposed measure, the Sum of Group Error Differences (SED_G).

A. Demographic Bias Simulation

To evaluate the demographic biases in a BV system, we typically need a dataset that consists of biometric data, demographic data, and unique person identifiers. For example, a typical bias evaluation dataset in face verification consists of face images, race labels, and unique person identifiers. To evaluate a BV system, genuine and impostor biometric sample pairs are generated randomly and sometimes selected following specific criteria, e.g. selecting only difficult pairs [22] or alternatively using all available pairs [16]. Distances between the pairs' respective embedding vectors are then calculated to quantify different error rates, such as FMRs and FNMRs, which can be disaggregated based on the demographic membership of the individuals being represented by the biometric samples to quantify the system's demographic bias. Therefore, the minimum data required to evaluate a biometric system's bias are the distances between pairs of embedding vectors, labels indicating whether the pairs belong to the same person or not, and information about the demographic membership of the individuals being represented by the sample pairs. A biased system is then expected to give higher false matches and non-matches for some demographic groups compared to others.

We simulate two demographic bias scenarios, one where a single group is disadvantaged at various levels in terms of model performance and a second scenario where more than one group suffer from different degrees of disadvantage in model performance. To achieve this, we synthetically generate model output distances and ground truth labels with specific FMRs at a TMR of 0.95, denoted as $\text{FMR}(\text{TMR}_{95})$, and simulate different disadvantage levels against different demographic groups using a disadvantage increase factor denoted by x. We describe this process more closely in the remainder of this section.

1) Single disadvantaged group: In the first scenario, there is a single disadvantaged group (g_{dis}); meanwhile, the remaining groups have the same performance levels. This scenario aims to test the bias evaluation metrics' ability to quantify disadvantage against a single demographic group. x represents the disadvantage increase factor, which takes

values 1, 2, 3, 5, 10, 20, and 50 depending on the level of bias being simulated. To simulate this scenario, we need to generate two synthetic datasets that satisfy the condition that the FMR at a TMR of 0.95 for the disadvantaged demographic group is x times the FMR at a TMR of 0.95 for the remaining groups.

$$\mathrm{FMR}_{dis}(\mathrm{TMR}_{95}) = x \cdot \mathrm{FMR}_{oth}(\mathrm{TMR}_{95}), \tag{6}$$

where:

- FMR_{dis} is the FMR at a TMR of 0.95 for the disadvantaged group g_{dis}.
- FMR_{oth} is the FMR at a TMR of 0.95 for the remaining groups.
- x is the simulated disadvantage increase factor and takes the values 1, 2, 3, 5, 10, 20, and 50.

Using (6), we simulate different levels of demographic bias by varying the FMR at a TMR of 0.95 value of a single demographic group using the variable x. Using this equation, we obtain two sets of synthetic model outputs, one for the disadvantaged group denoted as D_{dis} and another for all the remaining groups denoted as D_{oth}.

$$\mathrm{D}_{dis} = \{d \in \mathrm{D} | \mathrm{FMR}(\mathrm{TMR}_{95}, d) = \mathrm{FMR}_{dis}(\mathrm{TMR}_{95}, d)\}$$

$$\mathrm{D}_{oth} = \{d \in \mathrm{D} | \mathrm{FMR}(\mathrm{TMR}_{95}, d) = \mathrm{FMR}_{oth}(\mathrm{TMR}_{95}, d)\}$$

These two synthetic model outputs can be combined to simulate a model's output on n demographic groups as follows:

$$\mathrm{D}_{single} = \mathrm{D}_{dis} \underbrace{\cup \mathrm{D}_{oth}}_{n \text{ times}}$$

2) Multiple disadvantaged groups: In this scenario, multiple demographic groups suffer from different levels of disadvantage. This aims to test a metric's ability to capture intermediate-level biases, i.e., biases against groups other than the most disadvantaged. Multiple synthetic model outputs with varying levels of disadvantage need to be combined to simulate this scenario. For simplicity, we use the same values of x as in the previous scenario and set the minimum (or best) FMR at a TMR of 0.95 denoted by FMR_{best} as a reference for generating the $\mathrm{FMR}(\mathrm{TMR}_{95})$ values for the demographic groups. This process is described in (7).

$$\mathrm{FMR}_{i}(\mathrm{TMR}_{95}) = x \cdot \mathrm{FMR}_{best}(\mathrm{TMR}_{95}) \tag{7}$$

where:

- $\mathrm{FMR}_{i}(\mathrm{TMR}_{95})$ is the FMR at a TMR of 0.95 of a demographic group i.
- FMR_{best} is the FMR at a TMR of 0.95 of the group with the lowest FMR at a TMR of 0.95.
- x takes 1, 2, 3, 5, 10, 20, and 50 depending on the level of disadvantage simulated for demographic group i.

Using this equation, we obtain one synthetic model output dataset per each demographic group i denoted by D_i.

$$\mathrm{D}_i = \{d \in \mathrm{D} | \mathrm{FMR}(\mathrm{TMR}_{95}, d) = \mathrm{FMR}_i(\mathrm{TMR}_{95}, d)\}$$

These outputs can then be combined to simulate the outputs of a model that is biased against more than a single group as follows:

$$\mathrm{D}_{full} = \mathrm{D}_1 \cup \mathrm{D}_2 \cup \mathrm{D}_3 \cup ... \cup \mathrm{D}_n$$

where:

- n corresponds to the number of demographic groups in the simulation.

3) Dataset generation: To simulate the scenarios described in III-A.1 and III-A.2, we must generate model outputs with specific FMRs at a TMR of 0.95. To achieve this, we rely on hill climbing as in Algorithm 1.

Algorithm 1 Hill Climbing Algorithm for BV data synthesis

S_TMR $\leftarrow a$ ▷ Desired TMR
S_FMR $\leftarrow b$ ▷ Desired FMR
GT $\leftarrow$ array of i True and j False ground truth labels ▷ Binary labels for genuine and impostor pairs
$\mathrm{dist}_{gen} \leftarrow$ array of i random values [0.0, 0.5] ▷ Distance values for the genuine pairs
$\mathrm{dist}_{imp} \leftarrow$ array of j random values [0.5, 1.0] ▷ Distance values for the impostor pairs
dist $\leftarrow$ concatenate dist_{gen} and dist_{imp}
n $\leftarrow$ number of iterations
$\mathrm{fitness}_{best} = |\mathrm{FMR}(\mathrm{GT}, \mathrm{dist}) - \mathrm{S\_FMR}| + |\mathrm{TMR}(\mathrm{GT}, \mathrm{dist}) - \mathrm{S\_TMR}|$
for $i \leftarrow 1$ to n **do** ▷ Start the hill climbing algorithm
    $\mathrm{dist}_{new} \leftarrow$ dist + small random change
    $\mathrm{fitness}_{new} = |\mathrm{FMR}(\mathrm{GT}, \mathrm{dist}_{new}) - \mathrm{S\_FMR}| + |\mathrm{TMR}(\mathrm{GT}, \mathrm{dist}_{new}) - \mathrm{S\_TMR}|$
    if $\mathrm{fitness}_{new} < \mathrm{fitness}_{best}$ **then**
        $\mathrm{dist} \leftarrow \mathrm{dist}_{new}$
        $\mathrm{fitness}_{best} \leftarrow \mathrm{fitness}_{new}$
    end if
end for
return dist ▷ Return the distance values closest to the desired values

This hill climbing algorithm performs a local search that iteratively improves a randomly generated solution, specifically fit for BV data synthesis. To this end, we use a simple fitness function that calculates the absolute difference between the desired FMR and TMR values and the actual FMR and TMR values for the input distances and ground truth labels. Using this algorithm, it is possible to generate different model output distances and ground truth labels that satisfy desired FMR at a TMR of 0.95 or FNMR at a TNMR of 0.95 values. In our experiment, 1000 iterations were enough to reach the desired values for our demographic bias simulation.

B. Sum of Group Error Differences

SED$_G$ is our proposed measure for simultaneously addressing the limitations of the previous metrics highlighted in Section II-B. It relies on the average of the thresholds needed to achieve EER (T_{EER_g}) for each demographic group (g) in a set of demographic groups (G) as a reference point for measuring the deviations of individual demographic group performances from a global performance. T_{EER_g} is considered as the point (threshold) where the FMR is equal to the FNMR. Hence, our proposed measure accounts for both FMRs and FNMRs and quantifies demographic bias in reference to a global performance that includes WDIs and CDIs. First, to quantify the performance (error rate) deviations, we adapt the relative difference formula as follows:

$$\delta\mathrm{FMR}_g = |1 - \frac{\mathrm{FMR}_g(\bar{T}_{\mathrm{EER}})}{\mathrm{FMR}_{global}(\bar{T}_{\mathrm{EER}})}| \tag{8}$$

$$\delta\mathrm{FNMR}_g = |1 - \frac{\mathrm{FNMR}_g(\bar{T}_{\mathrm{EER}})}{\mathrm{FNMR}_{global}(\bar{T}_{\mathrm{EER}})}| \tag{9}$$

where:

- $\delta\mathrm{FMR}_g$ (respectively $\delta\mathrm{FNMR}_g$) represents the relative difference between the FMR (respectively FNMR) value for demographic group g at the average EER threshold $\bar{T}_{EER}$ and the global FMR (respectively FNMR) values at this same threshold. Note: We consider absolute values because we also want to treat better performance than the global reference performance as a form of bias.
- FMR_g (respectively FNMR_g) denotes the FMR (respectively FNMR) value for demographic group g.
- FMR_{global} (respectively FNMR_{global}) denotes the FMR (respectively FNMR) value using the full dataset.

$$\mathrm{SED}_g = \delta\mathrm{FMR}_g + \delta\mathrm{FNMR}_g \tag{10}$$

The values of $\delta\mathrm{FMR}_g$ and $\delta\mathrm{FMNR}_g$ are summed into a single value named the Sum of Group Error Differences (SED_g) for simplicity, which represents the over- and under-performance of each demographic group as shown in (10).

$$\mathrm{SED}_G = \{\mathrm{SED}_g \mid g \in G\} \tag{11}$$

The SEDs of all demographic groups are combined in a single set SED_G in (11). The average of the set SED_G represents the amount of deviation of demographic group performance from global performance. Meanwhile, the standard deviation of the set represents the variability of performance across the different demographic groups, and in that way, combines some of the strengths of aggregation- and STD-based measures.

IV. Experimental setup

A. Synthetic Dataset Description

To test the bias evaluation metrics in the scenarios with different types and levels of biases described in Sections III-A.1 and III-A.2, model output distances and ground truth labels are created to simulate the within-demographic performance of a model with different FMR(TMR$_{95}$) values on seven hypothetical demographic groups as described in Table II. The generated distances and ground truth labels can then be combined to simulate the outputs of models with varying types and magnitudes of demographic bias.

Increase factor	FMR$_{95}$ for g_dis	Genuine:Impostor
1	0.001	3000:3000
2	0.002	3000:3000
3	0.003	3000:3000
5	0.005	3000:3000
10	0.01	3000:3000
20	0.02	3000:3000
50	0.05	3000:3000

TABLE II
WITHIN-DEMOGRAPHIC ONLY SYNTHETIC DATASET CHARACTERISTICS

Increase factor	FMR$_{95}$	Genuine:Impostor
0.1	0.0001	3000:24000

TABLE III
WITHIN AND CROSS-DEMOGRAPHIC SYNTHETIC DATASET CHARACTERISTICS

Each set of distance and ground truth pairs represents a demographic group with 3,000 identities and two samples per identity. This allows for generating 3,000 genuine and 35,988,000 impostor pairs. However, we select an equal number of pairs (3,000 impostor and genuine pairs) to treat both types of errors with the same significance. We choose to have four demographic groups similar to some datasets in the demographic bias evaluation literature [22].

For simplicity, we refer to FMR(TMR$_{95}$) as FMR$_{95}$ in the remainder of the paper.

An additional set of simulated model output distances and ground truth labels is also created to simulate a model's global performance as described in Table III. This performance is usually better than within-demographic performance because it is dominated by cross-demographic impostor pairs, which are the easiest for a typical BV system to distinguish. Therefore, the number of genuine pairs becomes 12,000, and while it is possible to generate 575,952,000 impostor pairs, 600,000 impostor pairs are enough to simulate a global performance with an FMR$_{95}$ of 0.0001.

B. Compared Bias Evaluation Metrics

We test our proposed dual metrics and other established metrics, such as IR and GARBE [20], FDR [4], and STD in EER$_G$ [16]. To allow an equitable comparison, we only use metrics that account for both FMRs and FNMRs to quantify the bias in our scenarios with different levels and types of bias. The evaluated metrics are described in detail in the following.

IR is a maximum-disparity-based metric that combines the ratio between the highest and lowest FMR and FNMR values among a set of demographic groups. It calculates the maximum disparity in FMRs and FNMRs across different demographic groups at a desired or policy FMR threshold (T_{FMR_p}).

$$\mathrm{A}(\mathrm{T}_{\mathrm{FMR_p}}) = \frac{\max\left\{\mathrm{FMR}_g(T_{\mathrm{FMR}_p}), \forall g \in G\right\}}{\min\left\{\mathrm{FMR}_g(T_{\mathrm{FMR}_p}), \forall g \in G\right\}} \quad (12)$$

$$\mathrm{B}(\mathrm{T}_{\mathrm{FMR_p}}) = \frac{\max\left\{\mathrm{FNMR}_g(T_{\mathrm{FNMR}_p}), \forall g \in G\right\}}{\min\left\{\mathrm{FNMR}_g(T_{\mathrm{FNMR}_p}), \forall g \in G\right\}} \quad (13)$$

$$\mathrm{IR} = A(T)^{\alpha} B(T)^{1-\alpha} \quad (14)$$

where:

- α is the hyperparameter that determines the weight assigned to FMRs relative to FNMRs during the calculation of FDR and can be adjusted according to operational requirements.

We set the alpha (α) hyperparameter for all of the metrics to 0.5 to treat FMRs and FNMRs with the same level of significance so that they are comparable.

FDR is also a maximum-disparity-based metric that quantifies the rate of false match and non-match errors. Similar to IR, FDR takes into account the maximum disparity (in the form of a maximum difference rather than a ratio) in FMRs and, additionally, FNMRs at T_{FMR_p}.

$$\begin{aligned}\mathrm{FDR} = 1 - \big[&\alpha(\max\{|\mathrm{FMR}_{g1}(T_{\mathrm{FMR}_p}) \\ &- \mathrm{FMR}_{g2}(T_{\mathrm{FMR}_p})|, \forall g1, g2 \in G\}) \\ &+ (1-\alpha)(\max\{|\mathrm{FNMR}_{g1}(T_{\mathrm{FMR}_p}) \\ &- |\mathrm{FNMR}_{g2}(T_{\mathrm{FMR}_p})|, \forall g1, g2 \in G\})\big]\end{aligned} \quad (15)$$

GARBE is a metric inspired by a measure of statistical dispersion called the Gini coefficient, used to quantify the average difference between the FMRs and FNMRs of available demographic groups at a predefined policy FMR threshold.

$$\mathrm{Gini}_x(T_{\mathrm{FMR}_p}) = \big(\frac{|G|}{|G|-1}\big)\big(\frac{|x_{g1}(T_{\mathrm{FMR}_p}) - x_{g2}(T_{\mathrm{FMR}_p})|}{2n^2\bar{x}}\big), \quad \forall g1, g2 \in G \quad (16)$$

$$\mathrm{A} = \mathrm{Gini}_{\mathrm{FMR}}(T_{\mathrm{FMR}_p}); \mathrm{B} = \mathrm{Gini}_{\mathrm{FNMR}}(T_{\mathrm{FMR}_p}) \quad (17)$$

$$GARBE = \alpha A + (1-\alpha)B \quad (18)$$

STD in EER$_G$ (σ_{EER_G}) is a simplistic measure of bias that quantifies the standard deviation in demographic group-specific EERs computed on within-demographic pairs (see (21)), with EER being the value at which the FMR is equal to the FNMR (see (19)).

$$\mathrm{EER}_g = \frac{\mathrm{FMR}_g + \mathrm{FNMR}_g}{2} \mid \mathrm{FMR}_g = \mathrm{FNMR}_g \quad (19)$$

$$\mathrm{EER}_G = \{\mathrm{EER_g} \mid g \in G\} \quad (20)$$

EER_G represents the set of all demographic group-specific EER values, with each EER_g corresponding to a specific demographic group g in the set of all demographic groups G. Note: In [16], the EER is expressed in percentage form.

$$\sigma_{\mathrm{EER}_G} = \sqrt{\frac{1}{|\mathrm{EER}_G|}\sum_{\mathrm{EER}_g \in \mathrm{EER}_G}(\mathrm{EER}_g - \mu_{\mathrm{EER}_G})^2} \quad (21)$$

STD in SED$_G$ and Average of SED$_G$ are the proposed metrics in our dual-metric measure, designed to simultaneously address the limitations of the previous metrics highlighted in Section II-B. Set SED$_G$ is described in detail in Section III-B. The average of set SED$_G$ values ($\overline{\mathrm{SED}}_G$) measures the magnitude of the deviation of demographic group performance from global performance, while the STD in set SED$_G$ (σ_{SED_G}) measures the degree of variation in demographic group performance.

TABLE IV

EVALUATION METRICS AT DIFFERENT FMR$_{95}$ VALUES FOR A SINGLE DISADVANTAGED GROUP

Ratios	IR	GARBE	FDR	σ_{EER_G}	σ_{SED_G}	$\overline{\mathrm{SED}}_G$
1:1:1:1	1.0	0.0000	1.00	0.00	0.00	0.24
1:1:1:2	1.33	0.0090	0.9990	8.66e-4	0.17	0.32
1:1:1:3	3.63	0.0397	0.9958	1.29e-3	0.28	0.85
1:1:1:5	13.62	0.0758	0.9923	3.96e-3	0.75	1.26
1:1:1:10	22.40	0.0891	0.9890	4.69e-3	0.91	1.30
1:1:1:20	87.54	0.1170	0.9821	6.92e-3	1.32	1.56
1:1:1:50	368.7	0.1427	0.9693	1.54e-2	2.99	2.55

V. EXPERIMENTAL RESULTS AND DISCUSSION

The experiments aim to test the efficacy of the bias evaluation metrics in quantifying the biases of BV systems with different types and magnitudes of bias using the scenarios described in Sections III-A.1 and III-A.2.

A. Single Disadvantaged Group

In the case of a single disadvantaged group g_{dis} (4th group in the ratios in Table IV), maximum disparity metrics such as IR and FDR rank all systems accurately due to the absence of intermediate-level biases. All metrics consistently report higher scores for systems with a greater FMR$_{95}$ for the disadvantaged group, making them equally usable in this scenario. It is worth mentioning that for system 1:1:1:1, four metrics, IR, GARBE, FDR, STD in EER$_G$, and STD in SED$_G$, report no biases in their score. This is because such metrics measure the bias based on the maximum and minimum error difference or the variation in demographic group errors. The Average of SED$_G$ is the only metric that indicates a bias score equal to 0.24, as it measures bias by comparing the errors for each demographic group against the global errors. Depending on the fairness context, any of the metrics can be suitable for this scenario.

B. Multiple Disadvantaged Groups

1) Two-disadvantaged groups (Table V):

- IR is a maximum-disparity-based metric, meaning it relies on maximum and minimum FMR values. Therefore, it fails to capture intermediate-level FMR$_{95}$ value

TABLE V

EVALUATION METRICS AT DIFFERENT FMRS AT A TMR OF 0.95 RATIOS FOR TWO DISADVANTAGED GROUPS

Ratios	IR	GARBE	FDR	σ_{EER_G}	σ_{SED_G}	$\overline{\text{SED}_G}$
1:1:2:2	1.33	0.0120	0.9990	9.99e-4	0.20	0.37
1:1:2:3	3.63	0.0428	0.9958	1.78e-3	0.35	0.81
1:1:2:5	13.62	0.0793	0.9923	4.33e-3	0.82	1.15
1:1:3:3	3.63	0.0530	0.9958	1.50e-3	0.38	1.05
1:1:3:5	13.62	0.0905	0.9923	3.74e-3	0.70	1.32
1:1:5:5	13.62	0.1011	0.9923	4.58e-3	0.86	1.68

TABLE VI

EVALUATION METRICS AT DIFFERENT FMR_{95} RATIOS FOR THREE DISADVANTAGED GROUPS

Ratios	IR	GARBE	FDR	σ_{EER_G}	σ_{SED_G}	$\overline{\text{SED}_G}$
1:2:2:2	1.33	0.0090	0.9990	8.66e-4	0.16	0.46
1:2:2:3	3.63	0.0399	0.9958	0.0399	0.39	0.69
1:2:2:5	13.62	0.0767	0.9920	0.0767	0.85	0.98
1:3:3:2	3.63	0.0502	0.9958	2.12e-3	0.43	0.94
1:3:3:3	3.63	0.0397	0.9958	1.29e-3	0.37	1.23
1:3:3:5	13.62	0.0786	0.9923	3.33e-3	0.58	1.41
1:5:5:2	13.62	0.0990	0.9923	5.13e-3	0.95	1.50
1:5:5:3	13.62	0.0906	0.9923	3.97e-3	0.73	1.70
1:5:5:5	13.62	0.0758	0.9923	3.96e-3	0.73	2.02

changes. Hence it assigns identical scores to 1:1:2:3 and 1:1:3:3 although system 1:1:3:3 is simulated to have a higher disadvantage for the third group. A similar case is observed for 1:1:2:5, 1:1:3:5, and 1:1:5:5. Similarly, FDR is a maximum disparity-based metric that relies on minimum and maximum FMR and FNMR values. Therefore, it also fails to capture intermediate-level biases for those same examples.

- STD in EER_G and STD in SED_G give a lower bias score to 1:1:3:5 in comparison to 1:1:2:5 although the former is simulated to have a higher disadvantage for the third group. This is because they are both STD-based measures and in that sense, system 1:1:3:5 has a lower degree of variation in group errors. The only metrics capable of correctly quantifying the magnitude of bias, in this case, are GARBE and Average of SED_G.
- The only metrics with no failure cases are GARBE and the Average of SED_G followed by STD in SED_G and STD in EER_G, which both share the same failure cases due to their shared STD-based characteristic.

2) Three-disadvantaged groups (Table VI):

- The highlighted cases in red help demonstrate the functionalities of GARBE, STD in EER_G, and STD in SED_G. All three metrics quantify the difference in performance between demographic groups rather than the bias in each group separately against a reference point. Depending on the fairness context and definition, it is likely that system 1:5:5:2 is considered less biased than systems 1:5:5:3 and 1:5:5:5, which suffer from a higher level of disadvantage for the fourth group. However, the three metrics fail to capture this relation and, therefore, might not be suitable for this scenario.
- Similar to the previous scenario, the scores of IR and FDR are dictated by the highest FMR_{95}, causing them to neglect intermediate biases. This is also observable when comparing elements from Tables V and VI. Systems 1:1:2:2 and 1:2:2:2 are given the same IR and FDR scores, 1.33 and 0.9990, respectively. Similarly, systems 1:1:3:2, 1:1:3:3, 1:3:3:2, and 1:3:3:3 all have the same IR and FDR scores, 3.63 and 0.9958 respectively. A distinction between such systems might be necessary depending on the fairness definition.
- The dual SED_G metrics provide a good understanding of the type and magnitude of bias in each system. The Average of SED_G helps us understand the overall bias present in the system, while the STD in SED_G helps us understand the difference in performance across demographic groups. When used together, these metrics provide the most accurate results in this scenario.

TABLE VII

EVALUATION METRICS AT DIFFERENT FMR_{95} RATIOS FOR FOUR DISADVANTAGED GROUPS

Ratios	IR	GARBE	FDR	σ_{EER_G}	σ_{SED_G}	$\overline{\text{SED}_G}$
2:2:2:2	1.00	0.0000	1.00	0.00	0.00	0.49
2:2:2:3	2.72	0.0310	0.9968	2.16e-3	0.44	0.70
2:2:2:5	10.20	0.0682	0.9933	4.83e-3	0.92	0.93
2:2:3:3	2.72	0.0413	0.9968	2.50e-3	0.46	0.97
2:2:3:5	10.20	0.0795	0.9933	4.59e-3	0.83	1.24
2:2:5:5	10.20	0.0909	0.9933	5.58e-3	1.07	1.46
2:3:3:3	2.72	0.0310	0.9968	2.16e-3	0.48	1.29
2:3:3:5	10.20	0.0702	0.9933	3.95e-3	0.65	1.43
2:3:5:5	10.20	0.0826	0.9933	4.68e-3	0.82	1.69
2:5:5:5	10.20	0.0682	0.9933	4.83e-3	0.92	2.00
3:3:3:3	1.00	0.0000	1.00	0.00	0.00	1.77
3:3:3:5	3.74	0.0402	0.9965	2.67e-3	0.35	1.89
3:3:5:5	3.74	0.0536	0.9965	3.08e-3	0.45	1.91
3:5:5:5	3.74	0.0402	0.9965	2.67e-3	0.41	2.19
5:5:5:5	1.00	0.0000	1.00	0.00	0.00	2.53

3) Four-disadvantaged groups (Table VII): In this scenario, all four groups suffer from some level of disadvantage in reference to global performance.

- Only the Average of SED_G metric is capable of differentiating between systems 2:2:2:2, 3:3:3:3, and 5:5:5:5. This makes it useful in all four scenarios discussed so far, having the advantage of providing additional information about the type of bias present when used alongside the STD in SED_G. This example shows the benefit of using a metric that quantifies bias in terms of a difference from a global performance because it can capture the magnitude of the disadvantage for all the groups, as opposed to metrics that consider deviations and maximum differences in group error rates as bias.
- When all four groups have the same level of disadvantage as in systems 2:2:2:2, 3:3:3:3, 5:5:5:5, WDI-based metrics that lack a global reference such as IR, GARBE, FDR, STD in EER_G, and STD in SED_G cannot quantify the magnitude of this disadvantage and assigns them the same score. As previously discussed, this behaviour, in some cases, may be undesirable.
- Similar to the scenarios discussed in Sections V-B.1 and V-B.2, IR and FDR fail to capture intermediate-level

TABLE VIII
EVALUATION METRICS AT DIFFERENT $FNMR_{95}$ VALUES FOR A SINGLE DISADVANTAGED GROUP

Ratios	IR	GARBE	FDR	σ_{EER_G}	σ_{SED_G}	$\overline{SED}_G$
1:1:1:1	1	0.0000	1.00	0.00	0.00	2.14
1:1:1:2	13	0.0745	0.9963	2.02e-3	1.63	3.07
1:1:1:3	64	0.1166	0.9913	3.60e-3	2.76	4.04
1:1:1:5	134	0.1263	0.9890	5.19e-3	3.42	4.55
1:1:1:10	502	0.1572	0.9821	7.93e-3	6.47	6.18
1:1:1:20	1136	0.1667	0.977	1.04e-2	6.96	7.48
1:1:1:50	9396	0.1981	0.9568	1.94e-2	14.99	10.69

biases for all systems. Additionally, for systems 2:2:2:2, 3:3:3:3, and 5:5:5:5 which do not suffer intermediate biases but exhibit the same level of disadvantage for all four groups, both metrics assign them identical scores. As a result, IR and FDR are completely unusable when all demographic groups suffer from some level of disadvantage i.e. this scenario.

- As previously stated, when the Average of SED_G is used alongside the STD in SED_G, it is possible to measure the magnitude of bias present and also get an idea about the type of bias present. For instance, STD in SED_G assigns the same score (a value of zero) to systems 2:2:2:2, 3:3:3:3, and 5:5:5:5 that have the same type of bias. This enables us to know that the level of bias is consistent among all the groups, and by examining the Average of SED_G, it is possible to know the magnitude of this bias.

C. Bias in False Non-match Errors

We conduct an additional experiment to evaluate the effectiveness of the bias evaluation metrics in capturing bias in the form of false non-match errors. In this experiment, we fixed the FMR value to 0.05, which is equivalent to a TNMR of 0.95. We then introduce the bias by varying the values of FNMR at a TNMR of 0.95 for a single disadvantaged group. We only simulate a single disadvantaged group scenario as the purpose of this experiment is to test the metrics' ability to detect bias in the form of false non-match errors. We tested and demonstrated the limitations of the metrics in the multiple disadvantaged group scenarios in the previous experiments. We use the notation $FNMR_{95}$ to represent the FNMR at a TNMR of 0.95, similar to how FMR_{95} represents the FMR at a TMR of 0.95.

The behaviour of the metrics in this scenario (Table VIII) is identical to their behaviour in the scenario where bias is introduced in the form of different FMR_{95} values for a single disadvantaged group in Section V-A. All metrics are capable of accurately ranking the systems, which is because all the tested metrics rely on false match and non-match errors for bias quantification. The Average of SED_G is the only metric scoring bias for system 1:1:1:1, which might be useful depending on the fairness definition. IR and FDR rank all systems accurately, mainly due to the absence of intermediate-level biases.

VI. CONCLUSION

In conclusion, we revisit the bias evaluation metrics and discuss some of their advantages, disadvantages, and nuances.

IR and FDR: Since IR and FDR are both based on maximum disparity, it is sensible to discuss them together. Additionally, they rank all systems similarly. They are most suitable when there is only a single disadvantaged group. Being based on maximum disparity, both metrics study biases in terms of maximum and minimum error rates, hence not capturing intermediate biases. They are unusable in the presence of intermediate biases or when all demographic groups exhibit some level of disadvantage.

GARBE: GARBE quantifies bias in terms of the difference in false match and non-match errors among the demographic groups. This makes it suffer from similar issues and failure cases as the STD-based metrics, STD in EER_G and STD in SED_G. Since it relies exclusively on WDIs, it does not capture the magnitude of bias making it unusable in scenarios where all demographic groups suffer from the same level of disadvantage.

STD in EER_G: STD in EER_G is a WDI-based metric that considers both false match and non-match errors and measures the variation in demographic group error rates. While this means that it captures intermediate biases, it also means that it does not provide information about the magnitude of the bias present, which is an advantage that Average of SED_G has when used alongside STD in SED_G.

STD in SED_G and Average of SED_G: the value of the SED by design takes into account the differences in false match and false non-match errors in reference to the global values of those errors (see (10)). By accounting for global CDI and WDI performance, SED_G metrics hold a clear advantage over STD in EER_G. In our simulation, STD in EER_G and STD in SED_G ranked systems similarly. However, only the Average of SED_G metric ranked all systems correctly. In combination, the SED_G metrics provide a clearer understanding of the magnitude and type of bias present in a system making them usable in all scenarios.

In summary, we present our proposed measure of demographic bias in BV as a dual-metric measure that overcomes the identified limitations of previous metrics. In combination, the dual metrics in our measure offer an understanding about the type and magnitude of bias present in a BV system. Therefore, we encourage the research community around BV systems to include them in their evaluation.

VII. ACKNOWLEDGMENTS

This work has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie Actions (grant agreement number 860630) for the project "NoBIAS - Artificial Intelligence without Bias". This work reflects only the authors' views and the European Research Executive Agency (REA) is not responsible for any use that may be made of the information it contains.

REFERENCES

[1] V. Albiero, K. W. Bowyer, and M. C. King. Face regions impact recognition accuracy differently across demographics. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9, 2022.

[2] D. Bhattacharyya, R. Ranjan, F. Alisherov, and M. Choi. Biometric authentication: A review. *International Journal of u-and e-Service, Science and Technology*, 2(3):13–28, 2009.

[3] C. M. Cook, J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury. Demographic Effects in Facial Recognition and Their Dependence on Image Acquisition: An Evaluation of Eleven Commercial Systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(1):32–41, 2019.

[4] T. de Freitas Pereira and S. Marcel. Fairness in biometrics: A figure of merit to assess biometric verification systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(1):19–29, 2022.

[5] R. V. Garcia, L. Wandzik, L. Grabner, and J. Krueger. The harms of demographic bias in deep face recognition research. In *2019 International Conference on Biometrics (ICB)*, pages 1–6, 2019.

[6] S. Gong, X. Liu, and A. K. Jain. Jointly de-biasing face recognition and demographic attribute estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 330–347. Springer, 2020.

[7] P. Grother, M. Ngan, and K. Hanaoka. Face recognition vendor test part 3: Demographic effects. 2019-12-19 2019.

[8] J. J. Howard, E. J. Laird, R. E. Rubin, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury. Evaluating Proposed Fairness Models for Face Recognition Algorithms. In J.-J. Rousseau and B. Kapralos, editors, *Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges*, pages 431–447, Cham, 2023. Springer Nature Switzerland.

[9] 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)*, pages 1–8. IEEE, 2019.

[10] K. Kotwal and S. Marcel. Fairness index measures to evaluate bias in biometric recognition. In *International Conference on Pattern Recognition*, pages 479–493. Springer, 2022.

[11] E. López-López, X. M. Pardo, C. V. Regueiro, R. Iglesias, and F. E. Casado. Dataset bias exposed in face verification. *IET Biometrics*, 8(4):249–258, 2019.

[12] B. Lu, J.-C. Chen, C. D. Castillo, and R. Chellappa. An Experimental Evaluation of Covariates Effects on Unconstrained Face Verification. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(1):42–55, Jan. 2019.

[13] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother. IARPA Janus Benchmark - C: Face Dataset and Protocol. In *2018 International Conference on Biometrics (ICB)*, pages 158–165, Feb. 2018.

[14] J. P. Robinson, G. Livitz, Y. Henon, C. Qin, Y. Fu, and S. Timoner. Face Recognition: Too Bias, or Not Too Bias? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–10, 2020.

[15] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[16] I. Serna, A. Morales, J. Fierrez, and N. Obradovich. Sensitive loss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. *Artificial Intelligence*, 305:103682, Apr. 2022.

[17] Y. Shi and A. K. Jain. Docface: Matching id document photos to selfies. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8, 2018.

[18] P. Terhörst, J. N. Kolf, M. Huber, F. Kirchbuchner, N. Damer, A. M. Moreno, J. Fierrez, and A. Kuijper. A comprehensive study on face recognition biases beyond demographics. *IEEE Transactions on Technology and Society*, 3(1):16–30, 2021.

[19] K. Vangara, M. C. King, V. Albiero, and K. Bowyer. Characterizing the variability in face recognition accuracy relative to race. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 2278–2285, 2019.

[20] E. Villalobos, D. Mery, and K. Bowyer. Fair face verification by using non-sensitive soft-biometric attributes. *IEEE Access*, 10:30168–30179, 2022.

[21] M. Wang and W. Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9319–9328, 2020.

[22] 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 *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 692–702, 2019.