A Study of Bias Estimation in Biometric Systems

Kaïra Neily Sanon, Joël Di Manno, Tanguy Gernot, Christophe Charrier, Christophe Rosenberger

A Study of Bias Estimation in Biometric Systems

Kaïra Neily SANON[1,2], Joël DI MANNO[1], Tanguy GERNOT[2], Christophe CHARRIER[2], and Christophe ROSENBERGER[2]

[1]*FIME EMEA, 14000 Caen, France, Email:* {*neily.sanon, joel.dimanno*}*@fime.com*
[2]*Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France, Email:* {*tanguy.gernot, christophe.charrier, christophe.rosenberger*}*@unicaen.fr*

Abstract

In the current context of biometric system certification, it is essential to address inherent biases to ensure both fairness and accuracy. Our research introduces a new method for estimating biases in these systems, particularly under grey box conditions, which are commonly encountered in certification settings. We aim to quantify biases in gender and ethnicity using advanced metrics applied to a selected database that combines the datasets VGGFace, VGGFace2, and CWD. Our methodology implies the variation of decision thresholds to observe changes in metrics, thereby uncovering biases. The goal of this research is to compare metrics on their biases evaluation way. The outcomes are intended to aid in establishing a congruous protocol for bias estimation in biometric systems, thereby enhancing the fairness and dependability of biometric authentication methods.

Keywords: Biometric system, Fairness, Evaluation, Bias

1 Introduction

Biometric systems, which identify individuals through unique physical or behavioral characteristics, have become increasingly widespread across various industries. However, the effectiveness and fairness of these systems are often undermined by some biases, posing significant challenges, particularly in environments where certification is crucial. Such biases, if not properly addressed, can result in unequal performance across diverse demographic groups. In light of the limited access to the internal mechanisms of these systems, especially in scenarios known as grey box situations, this paper presents a refined approach to standardize the estimation of biases in biometric systems. Our study tackles the challenge of assessing bias with limited system transparency, focusing on how decision thresholds can affect performance metrics in the facial recognition (FR) context. This research would be useful not only for enhancing the fairness and dependability of biometric technologies but also for setting new benchmarks for future assessments in environments with similar access restrictions. Thus, in the second section, we will explore biometric systems and the definition of biases. The third section will concentrate on related research in this field. The fourth section will outline the protocol details, followed by the fifth section which will cover experimental aspects. Finally, we will conclude with a discussion in the last section.

2 Background

We present in this section some information as background on the proposed study.

2.1 Biometric system

A biometric system works by utilizing unique physical or behavioral traits to verify identity or identification. Typical biometrics includes fingerprints, facial recognition, iris or retina patterns, voice, signature, and even an individual's manner of walking, as noted in [Jain et al., 1999]. To fulfill its objectives, a biometric system functions through a series of stages:

1. **Capture:** Initially, the system captures raw biometric data, like a facial image or fingerprint. This data is then converted into a biometric template, a digital representation of the individual's unique characteristics.

2. **Extraction:** The system begins by extracting key features from raw data. In the case of facial recognition, it often employs convolutional neural networks (CNNs), which are adept at processing images. Examples include Inception v4 [Szegedy et al., 2016] and Resnet50 [He et al., 2016]. Inception v4 is known for its deep architecture that enhances image recognition accuracy, while Resnet50, a variant of

CNN, is notable for its ability to train extremely deep networks effectively. Alongside these modern methods, traditional techniques like Principal Components Analysis (PCA) are also used. PCA works by reducing the dimensionality of data, simplifying the dataset while retaining its essential characteristics. Additionally, autoencoders, as mentioned in [Wang et al., 2016], are utilized. Autoencoders are a type of neural network that learns a compressed representation of the input data, which is particularly useful in dimensionality reduction tasks.

3. **Comparison:** After extraction, features are compared against a database for identity verification or identification, using metrics like cosine or Manhattan distance for similarity assessment. The reality is that captures of the same identity are not consistently identical, yet they tend to be similar.

4. **Decision:** The system determines identity verification or denial based on the outcomes of comparisons. This decision-making process is influenced by the system's level of transparency, whether it's a grey box, black box, or white box system, and also hinges on the predefined decision threshold. In scenarios where cosine distance serves as the metric, we can create a scoring function s which is dependent on e (enrollment) and p (probability). For a given threshold, can effectively dictate the decision.

In Figure 1, we can see the workflow of a voice biometric system, indicating varied identification accuracies for different demographic groups, the process of feature extraction from a voice sample, and the classification steps leading to speaker identification.

2.2 Biases definition

In this study, we adopt [Danks and London, 2017] interpretation of bias, viewing it as a deviation from a standard rather than inherently negative. Bias manifests in various contexts: statistical bias refers to numerical deviations from expected values, moral bias concerns deviations from ethical norms, and other forms encompass legal, social, and psychological aspects. An action or policy might be biased according to one criterion but unbiased by another. For instance, a tech company's recruitment strategy focusing on top engineering schools may be unbiased in terms of educational merit but biased against socioeconomic diversity. This concept of 'statistical bias' highlights differences from a broader population.
Moving forward, we will define bias as performance disparities among different groups. A biometric system is viewed as fair if it consistently delivers equivalent results across all considered people categories. Thus, for our study on authentication involving gender and ethnicity, fairness implies equal recognition accuracy across these groups.

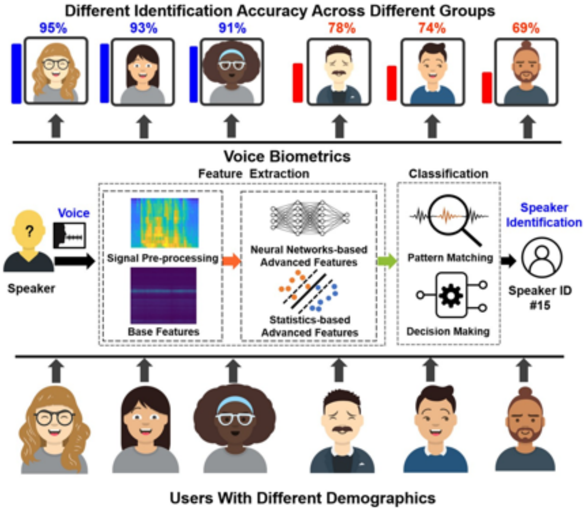

Figure 1: Example of biometric system [Chen et al., 2022]

3 Related works

In biometric authentication systems, assessing performance involves examining two primary types of errors: False Match Rate(FMR), where impostors are mistakenly accepted as genuine users, and False Non-Matches Rate(FNMR), where genuine users are wrongly denied. The occurrence of false matches can often be attributed to biometric features that lack uniqueness, whereas false non-matches may be caused by factors like noise during sample collection, poor-quality biometric templates, or changes in the user's biometric data over time. We call ERR for Error Relative Rate where FNMR and FMNR are same.

The evaluation of biometric systems lead sometimes to the use of some metrics in other to quantify biases. [Grother, 2021] found a way to assess demographic disparities by calculating the ratio with pondering of differences between the minimum and maximum values of the FMR and FNMR for various demographic groups. Even if the ratio implies that we are limited when we are a none FNMR, IR is a way to quantify biases between 0 and 1. Furthermore, another metric cited by [Grother, 2021], [DeAlcala et al., 2023], [Schuckers et al., 2022], [Howard et al., 2019], contributing to the understanding and measurement of fairness in these sys-

tems. In the context of assessing system vulnerability to identity concealment attacks, we have for example the metric Balanced Fairness (ABF) metric proposed by [Fang et al., 2024] which discuss about how to combine performance and fairness.
In our study, we will focus on two main metrics:

- **Fairness Discrepancy Rate (FDR) [De Freitas Pereira and Marcel, 2020]:** This metric evaluates a biometric system's fairness by measuring performance discrepancies across different demographic groups, focusing on gender and ethnicity. It consider that a Biometric Verificatin (BV) system is considered fair if different demographic groups share the same FMR and FNMR for a given decision threshold. FDR is defined by the following formula:

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

  where

  $$A(\tau) = \max\left(\left|FMR^{d_i}(\tau) - FMR^{d_j}(\tau)\right|\right),$$
  $$B(\tau) = \max\left(\left|FNMR^{d_i}(\tau) - FNMR^{d_j}(\tau)\right|\right).$$

  FDR ranges from 0 (indicating unfairness) to 1 (representing fairness).

- **Gini Aggregation Rate:** Developed in response to the issue of interpretability, [Howard et al., 2023] introduces the Functional Fairness Measure Criteria (FFMC). This set of criteria proposes an alternative metric, the Gini Aggregation Rate for Biometric Equitability (GARBE), which employs the Gini coefficient to:

  $$G_x = \left(\frac{n}{n-1}\right)\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n^2\bar{x}}\right), \quad (2)$$

  leading to:

  $$\text{GARBE}(\tau) = \alpha A(\tau) + (1 - \alpha) B(\tau), \quad (3)$$

  where $A(\tau)$ and $B(\tau)$ are the Gini coefficients for FMR and FNMR, respectively.

In this work, we intend to study the behavior of these metrics for the estimation of biaiases in biometrics systems.

4 Experimental protocol

To achieve the objective of comparing metrics variations and thereby quantify potential biases, it is imperative to have a reliable biometric system in place. For this purpose, we utilize a combination of Multi-task Cascaded Convolutional Networks (MTCNN) [Zhang et al., 2016] and Inception ResNet V1 [Li et al., 2022] for facial detection and feature extraction, respectively. These are complemented by the use of cosine similarity measures for the final comparison and classification stages. Additionally, a specialized database, referred to as DemogPairs [Hupont Torres and Fernández, 2019] database, has been curated to facilitate the assessment of biases across different demographics.

4.1 Algorithm Implementation

In the implementation phase of our research, we employed state-of-the-art feature extraction systems to enhance the precision of our biometric system certification. For the critical task of facial detection, we utilized the MTCNN, known for its efficacy in detecting faces across a wide range of scales and orientations with high accuracy. This algorithm operates in 3 phases: In Phase 1, the algorithm resizes the image multiple times to detect faces of varying sizes using the P-network (Proposal), intentionally introducing false positives. Phase 2 involves the R-network (Refine), refining initial detection, and "filtering" false positives to achieve precise bounding boxes. The final refinement is carried out in Phase 3 by the O-network (Output), ensuring accurate face detection and bounding box localization. Figures 2 and 1 illustrate these steps.

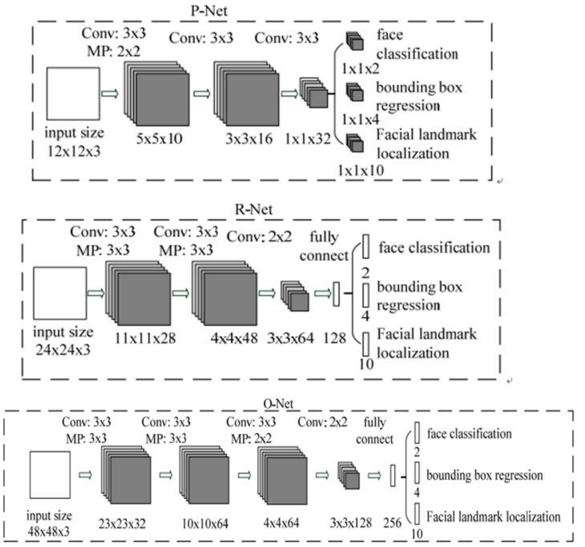

Figure 2: MTCNN architecture

Following the detection, we use the power of the Inception ResNetV1 architecture for feature extraction. This hybrid model combines the Inception architecture's [Szegedy et al., 2014] efficiency in handling different scales within an image with

Algorithm 1 Multi-task Cascaded Convolutional Networks (MTCNN)

```
Input: original test image
Output: image with face bounding boxes and facial landmarks
**Stage 1: Proposal Network (P-Net)**
Resize image to different scales to form an image pyramid
for each scaled image do
    Perform forward pass of P-Net
    Generate bounding box proposals
    Apply non-maximum suppression (NMS) to reduce overlap
end for
**Stage 2: Refine Network (R-Net)**
for each proposal from P-Net do
    Crop and resize the proposal region
    Perform forward pass of R-Net
    Refine bounding boxes and Apply NMS
end for
**Stage 3: Output Network (O-Net)**
for each refined proposal from R-Net do
    Crop and resize the proposal region
    Perform forward pass of O-Net
    Output final bounding boxes
    Output facial landmarks then apply NMS
end for
```

the ResNet model's residual learning capabilities to avoid vanishing gradient issues, thereby ensuring robust feature extraction that contributes significantly to the bias estimation process in our study. Then, to complete the items of a biometric system we use cosine similarities for comparison.
We acknowledge that MTCNN which we use is pretrained on the VGGFace dataset, which consists of 59.3% male subjects.

4.2 Database

For our experiment, we use DemogPairs [Hupont Torres and Fernández, 2019] which comprises 10.8K images across six demographic groups: Asian females and males, Black females and males, and White females and males, each with 100 subjects and 18 images per subject. Sourced primarily from CWF, VGGFace, and VGGFace2, it ensures minimal impact on training processes due to its low overlap with these datasets. This dataset presents 50% of males and females and 33.3% for each considered race.

DemogPairs enables the creation of 58.3 million evaluation pairs, including 29.1M cross-gender and 38.7M cross-ethnicity pairs. The dataset's design allows for analyzing challenging pairs within the same demographic group, which share similar features, and more distinct cross-demographic pairs. This diversity is instrumental in assessing the performance and bias of biometric systems.

4.3 Evaluation methodology

As previously discussed, the biometric system we use is inherently biased towards gender due to the imbalance in the training dataset, which comprises 59.3% male subjects. To address this, we suggest a methodology to compare metrics, as established in the current state of the art, specifically for the gender category. A similar approach will be applied to the ethnicity category.

1. **Feature and Label Extraction:** For our dataset, we extract features and labels referring to each category chosen (ethnicity/gender). This step was crucial to ensure the accuracy and relevance of the scores calculated later.

2. **Calculation of Match Scores:** We then determined genuine and imposter scores. A genuine score corresponds to a situation where the reference photo and the matching attempt come from the same person, while an imposter score is attributed when they come from different individuals.

3. **Error Rate Analysis (FAR and FRR):** The analysis involved calculating the false acceptance and rejection rates (FAR and FRR) over different thresholds, based on cosine similarity scores ranging from 0 (identical) to 1 (completely different). This step will permit us to view the performance of our system.

4. **Applying Measures to Various Performance Indicators:** Finally, we applied FAR and FRR measures to various indicators to assess the robustness of our method. We used detailed visualizations to show the impact of demographic variations on biases and accuracy in the biometric system.

5 Experimental results

In this experiment, we employed Python on a Windows laptop featuring an Intel® Core™ i5-9600K CPU running at 3.70 GHz, coupled with 16 GB of RAM. The extraction of features required approximately 5 minutes for every 3600 images, while the score calculation process took about 50 minutes for the same number of images. The remaining procedures in the experiment each took less than 1 minute to complete.

For this section, we present the results obtained in terms of the performance of our system and also in terms of fairness with selected metrics.

5.1 Performance evaluation

We evaluate the performance of the studied face biometric system by considering gender (see Figure 3) then ethnicity (see Figure 4). We observe that the system provides relatively better performance for males compared to females (difficult to say if this difference is significant). This outcome aligns with the expectation of bias arising from the system training on a gender-imbalanced dataset. Moreover, there are not significant differences with ethnicity.

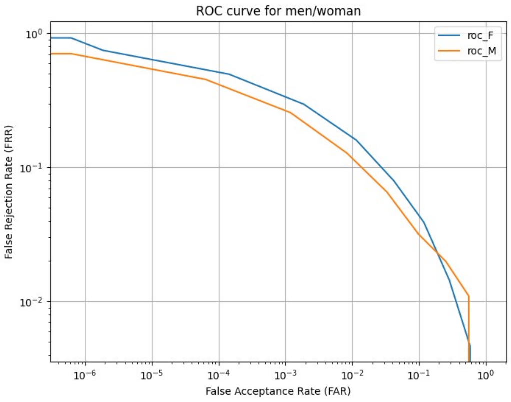

Figure 3: ROC curve concerning gender with log scale

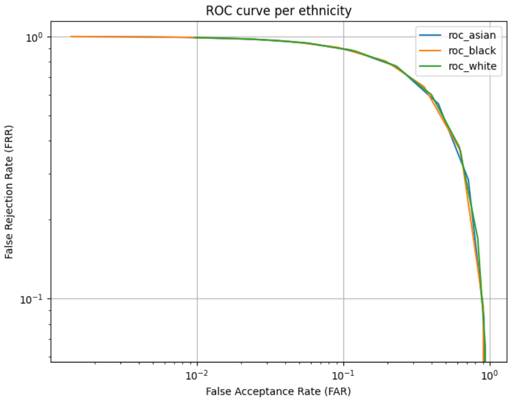

Figure 4: ROC curve concerning ethnicity with log scale

In the next part, we analyze the behavior of bias metrics from the literature.

5.2 Metrics comparison

Observing the results illustrated in Figures 5, 6, 7, and 8, it becomes evident that the values of FDR and GARBE vary with the selected threshold. That means the fairness of the biometric system depends on its decision threshold. This is an important information to consider.

Upon further examination of the impact of the α parameter, which is employed to balance the False Match Rate (FMR) and False Non-Match Rate (FNMR) in these metrics, a notable correlation with threshold levels emerges. Indeed, for lower thresholds, opting for a smaller value of α is advisable to reach fairness, while for higher thresholds, a larger α is more appropriate. This observation is consistent with the role of α in moderating the balance between FMR and FNMR, with a higher threshold implying an increased FNMR. The exception is made for FDR applied to ethnicity where the values of α are not impacted by the thresholds.

Moreover, except the figure 7, the other figures present a way where all the curves of FDR/GARBE for a given alpha meet.

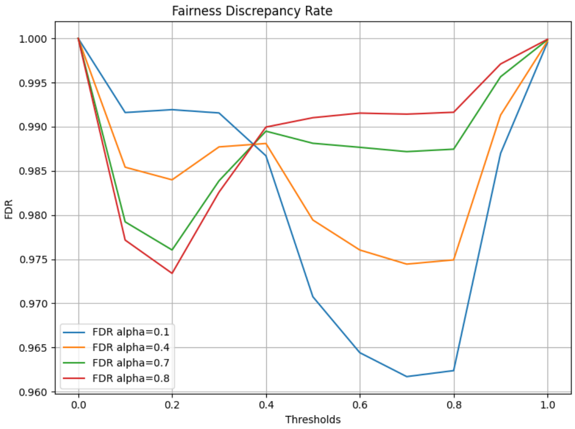

Figure 5: Gender evaluation with FDR

6 Conclusion and perspectives

The results of this work reveal that the Gini Aggregation Rate (GARBE) exhibits a higher sensitivity to biases compared to the Fairness Discrepancy Rate (FDR). This aligns with existing knowledge, as outlined in the GARBE paper, which highlights the limitations of FDR, such as the increased significance of the alpha parameter at higher values. Nevertheless, despite its limitations, FDR appears to more consistently reflect the actual performance

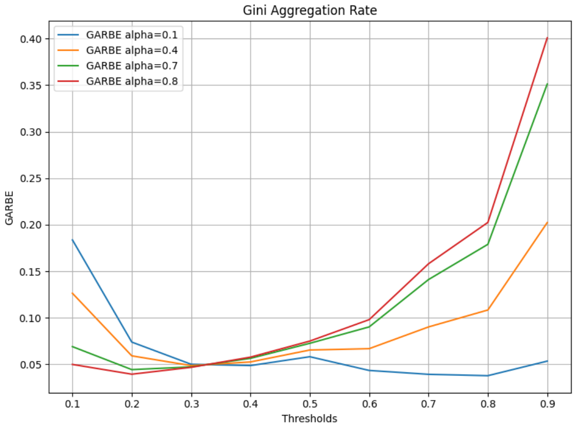

Figure 6: Gender evaluation with GARBE

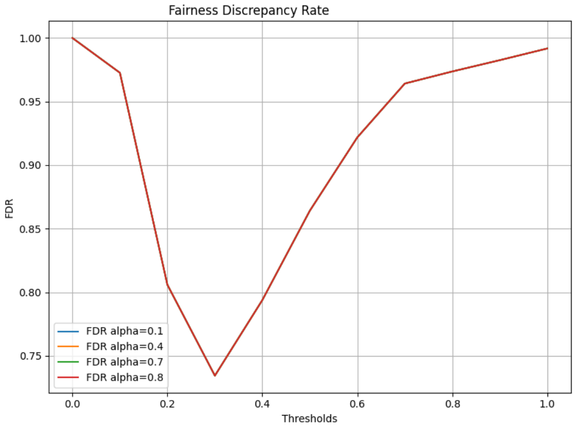

Figure 7: Ethnicity evaluation with FDR

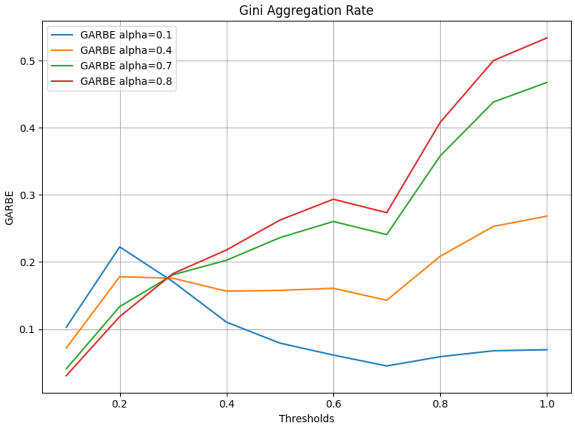

Figure 8: Ethnicity evaluation with GARBE

of biometric systems, which remains relatively stable. The inclusion of confidence intervals in the performance calculation would be beneficial to examine this question in detail.

To reinforce these conclusions, further research should expand to include a more diverse biometric array of systems and datasets. In our opinion, mixing biases in the FDR and GARBE metrics with α value is not suitable. As for performance evaluation, we should consider errors or biases separately for false rejection or acceptance. We intend in the future to contribute for the proposal of significant and useful fairness metrics for biometric systems.

References

[Chen et al., 2022] Chen, X., Li, Z., Setlur, S., and Xu, W. (2022). Exploring racial and gender disparities in voice biometrics. *Scientific Reports*, 12(1):3723. 2

[Danks and London, 2017] Danks, D. and London, A. J. (2017). Algorithmic Bias in Autonomous Systems. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 4691–4697, Melbourne, Australia. International Joint Conferences on Artificial Intelligence Organization. 2

[De Freitas Pereira and Marcel, 2020] De Freitas Pereira, T. and Marcel, S. (2020). Fairness in Biometrics: A Figure of Merit to Assess Biometric Verification Systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(1):19–29. 3

[DeAlcala et al., 2023] DeAlcala, D., Serna, I., Morales, A., Fierrez, J., and Ortega-Garcia, J. (2023). Measuring Bias in AI Models: An Statistical Approach Introducing N-Sigma. arXiv:2304.13680 [cs]. 2

[Fang et al., 2024] Fang, M., Yang, W., Kuijper, A., Struc, V., and Damer, N. (2024). Fairness in face presentation attack detection. *Pattern Recognition*, 147:110002. 3

[Grother, 2021] Grother, P. (2021). Demographic differentials in face recognition algorithms. EAB Virtual Event Series - Demographic Fairness in Biometric Systems. 2

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA. IEEE. 1

[Howard et al., 2023] Howard, J. J., Laird, E. J., Rubin, R. E., Sirotin, Y. B., Tipton, J. L., and Vemury, A. R. (2023). Evaluating Proposed Fairness Models for Face Recognition Algorithms. In

Rousseau, J.-J. and Kapralos, B., editors, *Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges*, Lecture Notes in Computer Science, pages 431–447, Cham. Springer Nature Switzerland. 3

[Howard et al., 2019] Howard, J. J., Sirotin, Y. B., and Vemury, A. R. (2019). The Effect of Broad and Specific Demographic Homogeneity on the Imposter Distributions and False Match Rates in Face Recognition Algorithm Performance. In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8, Tampa, FL, USA. IEEE. 2

[Hupont Torres and Fernández, 2019] Hupont Torres, I. and Fernández, C. (2019). pages 1–7. 3, 4

[Jain et al., 1999] Jain, A. K., Bolle, R., and Pankanti, S., editors (1999). *Biometrics: Personal Identification in Networked Society.* Springer US, Boston, MA. 1

[Li et al., 2022] Li, Z., Chen, Z., Che, X., Wu, Y., Huang, D., Ma, H., and Dong, Y. (2022). A classification method for multi-class skin damage images combining quantum computing and Inception-ResNet-V1. *Frontiers in Physics*, 10. 3

[Schuckers et al., 2022] Schuckers, M., Purnapatra, S., Fatima, K., Hou, D., and Schuckers, S. (2022). Statistical Methods for Assessing Differences in False Non-Match Rates Across Demographic Groups. arXiv:2208.10948 [stat] version: 1. 2

[Szegedy et al., 2016] Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. (2016). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. arXiv:1602.07261 [cs] version: 2. 1

[Szegedy et al., 2014] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2014). Going Deeper with Convolutions. arXiv:1409.4842 [cs]. 3

[Wang et al., 2016] Wang, Y., Yao, H., and Zhao, S. (2016). Auto-encoder based dimensionality reduction. *Neurocomputing.* 2

[Zhang et al., 2016] Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, 23(10):1499–1503. 3