

Florida Institute of Technology

Scholarship Repository @ Florida Tech

Theses and Dissertations

7-2021

Measuring the Relationship of Gender Misclassification and Automated Face Recognition Match Accuracy Relative to Skin Tone

Afi Edem-Edi Gbekevi

Follow this and additional works at: <https://repository.fit.edu/etd>



Part of the [Information Security Commons](#)

Measuring the Relationship of Gender Misclassification and Automated Face Recognition
Match Accuracy Relative to Skin Tone

by

Afi Edem-Edi Gbekevi

A thesis submitted to the College of Engineering and Science of
Florida Institute of Technology
in partial fulfillment of the requirements
for the degree of

Master of Science
in
Information Assurance and Cybersecurity

Melbourne, Florida
July 2021

We, the undersigned committee, hereby approve the attached thesis,
"Measuring the Relationship of Gender Misclassification and Automated Face
Recognition Match Accuracy Relative to Skin Tone"

by

Afi Edem-Edi Gbekevi

Michael King, Ph.D.
Associate Professor
Computer Engineering and Sciences
Major Advisor

Kevin Bowyer, Ph.D.
Professor
Computer Engineering and Sciences
University of Notre Dame
Committee Member

Vanessa Edkins, Ph.D.
Professor
School of Psychology
Committee Member

Philip Bernhard, Ph.D.
Associate Professor and Department Head
Computer Engineering and Sciences

Abstract

Title: Measuring the Relationship of Gender Misclassification on Automated Face Recognition Match Accuracy Relative to Skin Tone

Author: Afi Edem-Edi Gbekevi

Advisor: Michael King, Ph. D.

The gap of accuracy observed in some commercial face analytic systems based on race and gender raised questions about the equity and fairness of those systems. Since these systems are part of several applications today, some more critical than others, it urges designers to detect and mitigate any sources of bias. In this thesis, we begin by clarifying the confusion between face analytic, face recognition, and face processing systems. Then, we analyze gender classification accuracy using two datasets and three classifiers. The Pilot Parliaments Benchmark dataset is examined with an open-source algorithm to corroborate the gender shade. Secondly, the Morph dataset is employed to investigate the relationship between gender classification and face recognition as it is also suitable for face matching. Finally, we analyze the role of a person's skin in gender classification accuracy by correlating misclassified with false match pairs resulting from face match comparisons. We contribute to knowledge by providing evidence on the non-effect of gender classification on the face matching outcomes and providing the first investigation work on the skin tone-driven factor on the face processing results using an automated skin tone rating algorithm.

Keywords: gender classification, face recognition, face analytics, skin tone effect

Table of Contents

| | |
|---|-----|
| Abstract | iii |
| List of Figures | vii |
| List of tables | ix |
| Acknowledgments | xi |
| Dedication | xii |
| Chapter 1 Introduction..... | 1 |
| The edge of industrialization | 1 |
| Research questions | 2 |
| Chapter 2 Background..... | 4 |
| Biometric technologies | 4 |
| Face processing | 6 |
| Face Recognition | 8 |
| Face analytics | 10 |
| Chapter 3 Literature Review..... | 12 |
| Gender classification bias..... | 13 |
| Face Recognition bias | 15 |
| Relationship between face recognition and gender classification errors..... | 18 |
| Chapter 4 Experiment set-up | 19 |
| Gender Classification (GC) algorithms | 20 |
| Face Recognition (FR) matcher..... | 20 |
| Skin Tone (SK) classifier | 21 |
| PPB dataset..... | 21 |
| MORPH dataset..... | 22 |

| | |
|--|----|
| Chapter 5 Experiment results | 24 |
| PPB gender classification (GC) results | 24 |
| PPB GC with Skin tone classification | 26 |
| MORPH dataset GC results..... | 31 |
| Detailed analysis of the open-source GC results | 32 |
| Open-source GC result with Skin type classification | 33 |
| Comparison across the demographic cohort..... | 37 |
| Face matching (FM) results on the Morph dataset..... | 38 |
| Chapter 6 Gender classification errors and face matching analysis | 40 |
| GC errors placement across FM non -mated score distribution | 40 |
| African American Female (AAF): GC errors with the non-mated distribution .. | 44 |
| AAF false match errors analysis involving in the GC errors..... | 47 |
| African American Male (AAM): GC errors with the non-mated distribution | 48 |
| AAM False matching error analysis involved in the GC errors | 49 |
| Caucasian Female (CF) GC errors with the non-mated distribution..... | 50 |
| CF False matching error analysis involved in the GC errors..... | 52 |
| Caucasian Male (CM) GC errors with the non-mated distribution | 53 |
| CM False matching error analysis involved in the GC errors | 54 |
| Chapter 7 Skin tone (SK) factor in errors analysis | 56 |
| AAF SK assessment | 56 |
| AAF one image involved in GC errors | 56 |
| AAF SK assessment: two images involved in GC errors | 59 |
| AAF false match error with one image in GC error | 60 |
| AAF false match error with two images in GC error | 61 |

| | |
|--|-----------|
| AAM skin tone assessment | 63 |
| AAM one image involved in classification errors | 63 |
| AAM two images involved in classification errors | 64 |
| AAM false match error with one image in GC error | 65 |
| AAM false match error with two images in GC errors | 66 |
| CF skin tone assessment | 69 |
| CF one image involved in GC errors | 69 |
| CF one image involved in GC errors | 70 |
| CF two images involved in GC errors | 71 |
| CM skin tone assessment | 72 |
| CM one image involved in GC errors | 72 |
| CM false match error with one image in GC error | 73 |
| Chapter 8 Conclusion | 75 |
| References | 77 |

List of Figures

| | |
|--|----|
| Figure 1. Face recognition processing flow, image from[9]..... | 10 |
| Figure 2.Face analytic processing flow, image from[23] | 11 |
| Figure 3.PPB GC results "with Morph used in training" and result "without Morph used in training" | 25 |
| Figure 4. Accuracy table from gender shade | 26 |
| Figure 5.PPB female comparative skin tone plot..... | 28 |
| Figure 6.PPB female errors per skin tone | 28 |
| Figure 7.PPB Male comparative skin tone plot. | 30 |
| Figure 8. AAF Comparative skin tone table and plot | 34 |
| Figure 9. CF comparative skin tone table and plot | 35 |
| Figure 10. AAM Comparative skin tone table and plot. | 36 |
| Figure 11.CM Comparative skin tone table and plot. | 37 |
| Figure 12.Face Matching distribution score..... | 39 |
| Figure 13. AAF Impostor distribution on the left the mean and std range, | 44 |
| Figure 14.AAF impostor distribution with GC errors placement | 45 |
| Figure 15. AAF open-source comparative plot between the impostors..... | 46 |
| Figure 16. AAF Amazon(left), Microsoft(right) comparative plot between the impostors | 46 |
| Figure 17. Impostor distribution showing on left the mean and std range,..... | 48 |
| Figure 18. AAM open-source comparative plot between the impostor and the pairs involved in GC errors..... | 49 |
| Figure 19. CF Impostor distribution showing on the left the mean and std,..... | 51 |
| Figure 20. CF open-source comparative plot between the impostor and the pairs involved..... | 52 |
| Figure 21.CM Impostor distribution showing on left the mean and std, | 53 |
| Figure 22. CM impostor distribution with GC errors placement | 54 |
| Figure 23. AAF SK distribution plot per std range..... | 57 |

| | |
|--|----|
| Figure 24. AAF Skin tone Amazon(left), Microsoft(right) SK distribution plot per std range | 58 |
| Figure 25. AAF Skin tone open-source comparative std plot per SK..... | 58 |
| Figure 26. AAF Skin tone Amazon(left) Microsoft(right) comparative std plot per SK..... | 59 |
| Figure 27. AAF two images involved in GC errors per classifiers..... | 59 |
| Figure 28. AAM open-source SK distribution plot per std range | 63 |
| Figure 29. AAM Amazon(left), Microsoft(right) SK distribution plot per std range | 63 |
| Figure 30. AAM Open-source comparative std plot per skin tone | 64 |
| Figure 31. AAM Amazon(left) Microsoft(right) comparative std plot per skin tone | 64 |
| Figure 32. AAM two images involved in GC errors per classifiers..... | 65 |
| Figure 33. CF Skin tone Open-source comparative std plot. | 69 |
| Figure 34. CF Amazon(left) Microsoft(right) comparative std plot per SK. | 70 |
| Figure 35. CM Open-source comparative std plot per SK..... | 73 |
| Figure 36. CM Amazon(left) Microsoft(right) comparative std plot per SK..... | 73 |

List of tables

| | |
|---|-----------|
| Table 1. PPB dataset split by gender and country..... | 22 |
| Table 2. GC and FR result samples..... | 23 |
| Table 3. Gender prediction result from the 1st and 2nd experiment..... | 24 |
| Table 4. PPB female classification raw results | 27 |
| Table 5. PPB female relative frequency results | 27 |
| Table 6. PPB Male Classification raw results..... | 29 |
| Table 7. PPB Male Classification relative frequency results..... | 29 |
| Table 8. Morph dataset GC result with three classifiers. | 32 |
| Table 9. Morph dataset GC result with the Open-source classifier. | 32 |
| Table 10. AAF Open-source GC result per skin tone | 33 |
| Table 11. CF Open-source GC result per skin tone | 34 |
| Table 12. AAM Open-source GC result per skin tone..... | 36 |
| Table 13. CM Open-source GC result per skin tone..... | 37 |
| Table 14. Open-source GC cross skin tone comparaisn..... | 38 |
| Table 15. One impostor image involved in GC errors split by std range..... | 41 |
| <i>Table 16. Two impostor images involved in GC errors split by std range</i> | <i>43</i> |
| Table 17. AAF False match errors with one image involved | 47 |
| Table 18. AAF False match errors with two images involved..... | 48 |
| Table 19. AAM FM errors with one and two images involved in GC errors | 50 |
| Table 20. CF FM errors with one and two images involved in GC errors..... | 52 |
| Table 21. CM FM errors with one and two images involved in the GC errors | 54 |
| Table 22. AAF SK relative frequency for one image involved in open-source GC errors | 56 |
| Table 23. AAF Total impostor involving one image in GC errors per SK..... | 60 |
| Table 24. AAF FMR involving one image in GC errors per SK. | 61 |
| Table 25. AAF FMR involving one image in GC errors per SK. | 61 |
| <i>Table 26. AAF FMR involving two images in GC errors per SK.</i> | <i>62</i> |

| | |
|--|----|
| Table 27. AAF FMR with same SK for the pairs involved | 62 |
| Table 28. AAM Total impostor involving one image in GC errors per SK. | 65 |
| Table 29. AAM FM involving one image in GC errors per SK. | 66 |
| Table 30. AAM variation of FMR involving one image in GC errors per SK. | 66 |
| Table 31. AAM total Imp (left), False match (right) involving two images in GC errors per SK. | 67 |
| Table 32. AAM FMR involving two images in GC errors per SK. | 67 |
| Table 33. AAM FMR with same skin tone for the pairs involved..... | 68 |
| Table 34. CF SK relative frequency for one image involved in open-source GC errors | 69 |
| Table 35. CF Total impostor involving one image in GC errors per SK. | 70 |
| Table 36. CF FM involving one image in GC errors per SK. | 70 |
| Table 37. CF FMR involving one image in GC errors per SK. | 71 |
| Table 38. CF total Imp (left), False match (right) involving two images in GC | 71 |
| Table 39. CF of Open-source FMR involving two images in GC errors per SK..... | 71 |
| Table 40. CF FMR with same SK for the pairs | 72 |
| Table 41. CM Total impostor involving one image in GC errors per SK..... | 73 |
| Table 42. CM FM involving one image in GC errors per SK..... | 74 |
| Table 43. CM FMR involving one image in GC errors per SK..... | 74 |

Acknowledgments

Above all, I want to thank God for his grace and strength upon me over these years.

I am grateful to the Fulbright program for the opportunity and scholarship that help me pursuing this program.

I want to thank Dr. William Allen, my academic advisor, for his unconditional support during this master's program.

I want to thank and express my gratitude to my supervisor Dr. Michael King for introducing me to the topic and the support on the way, and for all the valuable comments, remarks, and engagement through the learning process of this master thesis.

Furthermore, I want to thank my coworkers, Dr. Krishnapriya k. S. and Kushal Vangara (Ph.D. Student), who willingly shared their precious time by helping me throughout the process.

Special thanks to my English tutor Lucille Serody for helping all along the writing process.

To all my loved ones, thanks for your prayers, support, and for keeping me harmonious all the time. I will be grateful forever for your love.

Dedication

“I dedicate this thesis to my husband, who encouraged me to pursue my dreams, accepted the distance, and supported me all the way through.”

Thanks for the love and the patience.

Chapter 1

Introduction

The introduction section browses technologies revolution to date and states the investigation research questions.

The edge of industrialization

The world is continually evolving, and today we are at the edge of the fourth industrial revolution[1]. Automation invades many sectors, from entertainment to sensitive fields such as health care or criminal services. Industry 4.0 is the new era of convergence of Information Technology (IT) and Operational Technology (OT). The former deals with digital information flow, while the latter manages the machinery and physical processes used to carry the information out. This association of hardware and software resulted in the rise of the Internet of things(IoT)[2]. It is a transformative world where automation, advanced robotics, big data, intelligent factories, machine learning, and artificial intelligence (AI) become a part of it entirely.

The application of machines and algorithms to perform tasks once performed by humans has revolutionized many areas, increased productivity, and sped up many processes. However, when it comes to the reliability, effectiveness, and accuracy of those automated systems, one can realize that perfection is not achievable in the real world; we cannot design 100% accurate systems. At some point, these innovations shrink the space for flexibility and intuition and, more importantly, cause the attempt to perform tasks not capable of being performed by machines. Moreover, the increased dependency on technology raises questions about security, privacy, and human rights.

The years 2018-2019 and 2020 recorded most of the discussion around racial bias in face recognition systems. We can read the bold headline in the New York Times

[3], "Face recognition is accurate if you are a white guy," and the ACLU news titled "How is Face recognition, surveillance racist?" [4]; on Harvard University's blog page [5], we read "Racial discrimination in face recognition technology," and the list goes on. However, in 2010, in the Multiple-Biometric Evaluation (MBE), the investigation¹⁹ stated that the link between race and accuracy showed the "race-effect." In their studies, black subjects were more straightforward to be recognized than the white cohort for five out of six algorithms, and American and Asia were easier to identify for three algorithms. Additionally, the investigation across sex concluded that males generated fewer non-match errors than females[6]. The bias topic became louder when Buolamwini et al. published the project Gender Shades. Following the observation that the university biometric system misclassified her gender, she investigated three commercial classification systems that revive the topic.

Research questions

Regarding the controversy on face recognition characterized as racist following the gender shades project [7] that audited three commercial gender classifiers where darker subjects recorded the more significant error rate, two questions that have yet to be widely explored enough stick in our minds:

- Do errors generated in gender detection carry over into face recognition?
- To what extent does skin tone influence the result observed in question 1?

These are the two essential questions that will guide the progression of this document. First, we replicate the experiment using the PPB dataset from the gender shade project to corroborate the previous results and extend the work with the Morph dataset, which is suitable and used later for face recognition. Second, we correlated the errors resulting from morph dataset gender detection and false match errors from face recognition. Finally, we classify those common errors on the ITA skin tone scale to measure skin color's impact.

The rest of this document is divided as follows: we provide the Background and the Literature Review. Following, we explain the Experiment set-up and present the Experiment results. We conducted analysis regarding: Gender classification errors and face matching analysis and Skin tone (SK) factor in errors analysis, and closed the work with the Conclusion.

Chapter 2

Background

In the background section, we provide the literature on biometric areas and clarify different terms related to face processing.

Biometric technologies

The cognitive computing wave is spreading, and AI is becoming increasingly prevalent. Because of risks and flaws embedded in technologies, the industry has turned to cybersecurity to help limit the impact of their deficiencies. Among the methods of protecting systems or individual information, we have used a knowledge-based approach in the form of a password or pin, a physical or digital token such as a passport, credit card, or digital security keys. However, those former methods suffer from disadvantages such as theft, loss, forgetting, and the inability to prove the identity. Hence, the automation of such a process for better controls became evident. This automation of the authentication method, referred to as Biometrics[8], uses human biology attributes for identification purposes.

Biometric technologies use several distinct physiological or behavioral parts of a human to validate or reject the claimant's identity. For a human attribute to become a biometric trait, it has to meet seven requirements. [9]Universality: all humans possess the quality; Distinctiveness: the predicate must be discriminative among the population; Invariance: it remains the same over time; Collectivity: features are extractable and processible; Performance provides high accuracy; Acceptability: population will submit the attribute willingly; Circumvention: not prone to attack or mimicry.

Based on the criteria above, different categories of human traits participate in biometric systems:[10]

- Hand Region with fingerprint, palm print, hand geometry, hand vein pattern, or finger knuckle print. These features are the oldest in the biometric technologies used for identification at a crime scene; for example, the government uses them to establish unique identity cards, and, with the advantage of low-cost imaging sensors and the small size of templates needed for fingerprinting technologies, they are widely adopted by many applications and devices for authentication purposes[8]
- Facial regions represent the most natural attribute for recognizing humans[11], hence has gained more interest from researchers. However, the nonlinear structure of the human face makes facial technology a complex pattern recognition problem. The new trend is to use 3D representation to rectify some challenges associated with 2D facial recognition, such as sensitiveness to illumination conditions, pose variations, aging, and other occlusions[12]. Apart from the face itself, other attributes from face regions used in biometrics are ear shape, teeth, and tongue.
- The ocular region possesses more accurate and highly reliable, stable, and almost impossible to forge biometric signatures. It includes the retina, iris, sclera, and vasculature. The image acquisition could be quite invasive, especially for the retina, and the development of synthetic iris images from stored iris code has opened a debate on iris template protection.
- Medico-Chemical systems are identified by body odor, DNA, heart sound, or electrocardiogram, requiring medical or chemical sensors for acquisition. DNA is the most well-established for identification, whereas heart sound and ECB for identification are still undergoing studies.
- Behavioral systems focus on the way humans perform some activities such as type styles (keystroke dynamics), vocal characteristics(voice), signature dynamics, and walk(gait). Even though the research work in the speech and gait arena is underway, it has been noticed that human behavior is linked to emotions and external factors and could be easily mimicked by an impostor.

- Soft biometrics have gained attention due to the imaging errors from hard biometric traits (face, iris). One can list gender, ethnicity, height, scars, marks, or tattoos as a part of the soft attributes. These latter lack distinctiveness and permanence because they are the most common among humans. Nonetheless, they can be used for categorization and to limit research space [13]. Nevertheless, several studies have claimed to achieve significant improvement by combining soft and hard biometric attributes[10], [13].

Biometric technologies' prospects go from government to private stakeholders and are part of many applications beyond individual security. It impacts the life of the masses. One makes decisions based on those technologies' output. Hence, it is crucial to invest more in research to better those algorithms and automation processes to protect people and ensure the right choices.

In the following section, we focus on the study related to facial attributes and soft biometrics. The human face study has appeared in various areas, including cognitive neuroscience, psychology, personality, and mental illness on the one hand. On the other hand, we have biometric face recognition and biometric face analytics, where the intersection of those areas is confusing and can lead to wrong interpretation. For this reason, we have dedicated the following section to clarifying each domain and to drawing the intersection line between them.

Face processing

Humans are known to be experts at recognizing faces; we are a social species with the natural ability to classify, identify, and memorize the known face for a long time. This capability observed since childhood raises curiosity and yields several studies. The terminology Face processing refers to the ability of humans to categorize and recognize faces. [14]The researchers observed that this capability starts from the first days of birth, contrary to speech or walking capability. For

example, newborns in their early days show preferences for human faces versus non-human faces. They will prefer familiar-looking visages versus non-familiar faces and attend to attractive faces when paired with an unattractive look. Gradually as time passes, a 3-month-old baby will have a visual preference for faces that match the gender of his primary caregiver and the face from its ethnicity. The explicit recognition with subtle differences in morphology will follow adulthood with more exposure and experience[15].

Researchers investigated the brain region responsible for human face processing to clarify questions regarding humans' perceptions and behavior, nature versus nurture, such as [9]; are we predisposed to attend to what we focus on and interpret cues in the face? Is this learned through experience? Second, can we diagnose and treat some neural diseases related to recognition and know why it is difficult to recognize people from other races? This study extends into psychology, neuroscience, and human development science.

In their article [16], Pascalis and D. J. Kelly reviewed models and evidence from development, evolutionary and comparative psychology and concluded that a dedicated and complex neural system leads to this capability. For example, in developmental psychology, evidence provided by models such as CONSPEC and Gestational Proprioceptive Feedback (GPF) points out this capability starts at the gestation stage. At the same time, the fetus acquires abilities such as arm and leg movement, listening, eyes blinking around 24 weeks of gestation, and learning and developing preferences from the mother's behavior in the last six weeks of pregnancy. Once born, newborns will recognize and prefer their mother when paired with others [11],[12]. Also, the Perceptual narrowing model showed that human face recognition becomes tuned during the 6-to-9-month age range. However, the process is not human-specific and has been found with non-human primates such as monkeys in comparative psychology. Evolutionary psychology focused on the hormonal and physiological factors involved in the adult choice of mate. The newest methods, such as Functional Magnetic Resonance Imaging

(fMRI) and Transcranial Magnetic Stimulation(TMS)[15], [17] discovered that there are specialized cells in the brain responsible for face processing, and this process goes through hierarchies that transform visual information through multiple levels of processing. Three central parts in the cortical areas activate when a human is presented with faces: the Inferior Occipital Gyrus ('OFA'), the Middle Fusiform Gyrus ('FFA'), and the Superior Temporal Sulcus (STS)[15].

Human face processing expertise uses two types of information for recognition [15]. The featural attributes isolate internal features (eyes, nose, and mouth) and external ones (hairstyle and jawline). The second refers to the spatial relationship between elements, i.e., the distance between eyes, nose, and mouth—the combination of the dual information yields the holistic representation of the face. While the process started at birth, several investigations confirm that it becomes performant with age and experience—some factors such as environment and exposure impact the preference and the ability to recognize non-familiar faces.

Overall, studies agree that face processing is a complex and arduous task that requires multiple processing levels, whether started at gestation or through experience.[17] "The face processing systems are modular and distributed and appear to proceed in parallel through hierarchies." All information collected from the way the brain represents and processes faces helped create computers and algorithms that mimic this ability and automated the face recognition process with landmarks, discriminate features, holistic representation, and neural algorithm training.

Face Recognition

When it comes to using human recognition expertise for control and security, eyewitness evidence, or mass surveillance, this capability seems limited to perform this task to a large extent. For example, we have - the notion of the "own-race advantage" or the "other-race effect" [15], where humans have difficulty uniquely

identifying people from other races than theirs. Adapting the capability to fields other than human socializing needs leads to the creation of machines able to perform in an automated way. Hence, a biometric face recognition system is a computerized process to verify and identify a subject.

The pioneers [18] Woody Bledsoe, Helen Chan Wolf, and Charles Bisson, in the 1960s, manually computed the landmarks and used the RAND Tablet to perform face recognition before the development of the first automated system in 1973 by a Ph.D. student Takeo Kande. Then in the early 1990s, Sirovich and Kirby's work on the low dimensional representation of a face using the Eigen method and PCA analysis provided a significant breakthrough[9].

Today, thanks to the improvements in camera technologies, feature mapping, machine learning, and processing speeds, and due to the noninvasive method to acquire the data, face recognition is widely spread and meets general acceptance. It is present on mobile phones, airports, social media, forensic investigation, retail business, etc.

How does the system work? Face Recognition is performed in five stages[19].

Referring to Figure 1 below, the first stage, preprocessing: detects the location of the face and its size, rescaling and realigning the image if needed. Secondly, there is the extraction of features or landmarks to create a graph representation. Thirdly, the system compares the graph to all the previous photos present in the database. A high similarity score from the matching confirms the individual's existence in the database and leads to a verification or identification operation.

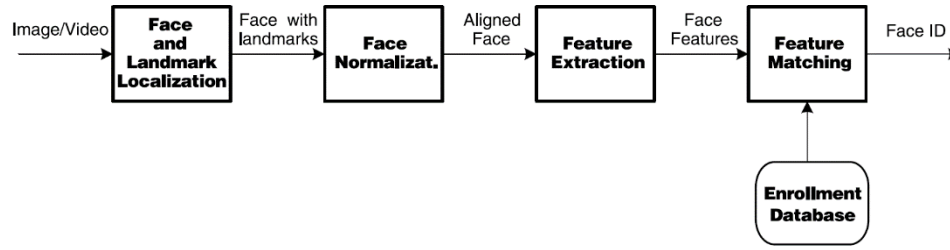


Figure 1. Face recognition processing flow, image from[9]

For both operations, the system requires a threshold that will be used to make the decision. The point is defined according to the application domain and whether security or convenience is more important.

During the matching process, the systems generate the genuine score (matching score between a pair of the same subject above the threshold) and impostor scores (matching score between a couple of images from different subjects below the threshold). Those scores generate two types of errors: the false matching errors, where an impostor score is above the threshold, and the false non-matching errors when a genuine pair generate a score below the threshold.

Face analytics

As previously detailed in the biometric technologies section, soft biometrics represent one of the trending categories currently. In 2011, K. Rickanet Jr et B. Barbour announced the concept. Contrary to face recognition itself, which extracts features to establish a person's identity, the latter tends to generate descriptive metadata from an image such as expression, pose, shape, age, sex, etc. used for classification, which can be a two-class problem: gender (Male, Female) or multiclass problem: Ethnicity (white, black, Hispanic, Asian), Age(baby, teen, adult, old)[20]

It found its application in commercial advertisements to display user content based on gender, age, race, or emotions. Law enforcement uses it to detect child pornography by identifying the type of partial skin(young or adult)[20], [21]. Combined with facial recognition complex traits, better performance is achieved.

For the process represented in Figure 2, a detection phase and normalization occur, and then the method is then used to extract the features. The descriptors need to be highly distinctive and, at the same time, lower the computation load and the sensitivity to noise, illumination, scaling, rotation, and skew. This first step is the same as face authentication or identification. Some intermediate steps might reduce the number of used features to lighten the computational complexity and increase the accuracy. Finally, a classifier is used for prediction. [22]



Figure 2.Face analytic processing flow, image from[23]

In this process, we can notice the absence of a database for recognition or decision-making based on the threshold presented in face recognition.

Face recognition and face analytics have become two essential domains associated with existing technologies enhancing their capabilities, gaining time, and improving user experience. Hence, governments and businesses have invested in more research for better systems. Given that any technology comes with its downside and implements its creator's bias, questions regarding the accuracy, fairness, privacy, and justice tainted biometric systems.

Chapter 3

Literature Review

In their experiment, J. Buolamwini and T. Gebru [19] created a PPB dataset benchmark that includes intersectional demographics from Africa to Europe regarding balance in gender used to evaluate commercial classifiers MSFT, Face++, and IBM. Their result showed that all the classifiers perform better for men than females. When breaking down into skin tone groups, it showed that the outcome for darker subjects was biased, especially that darker females were less accurately classified. The error score is 34.7% for darker females, while the lighter-skinned male maximum error score is 0.8%. The classifiers' intersectional demographic analysis shows an error rate of 23%, 36%, and 33.1% for darker females for MSFT, Face++, and IBM.

Adding to all the unfairness and racial and social issues when the paper came out, the news went viral and could not be stopped any longer. The concerned companies had to take a step back to improve the system and fill the gap. Several research and experiments followed because the biases had to be addressed for the sake of the population. Our thesis fits in the same context. However, several titles that appeared following the gender shades project have referred to it as face recognition, while the gender shades experiment is about gender classification from faces.

Even though face recognition has many challenges that affect its performance, the process and analysis are complex and involve many factors. Referring to the Introduction, face recognition performs a match between two images to generate a score. This process implies an enrollment step to constitute the database. Next, the decision to accept or reject the match, whether for verification or identification, is based on a predefined threshold. Based on the distribution scores and the threshold, the following information is derived to evaluate the system's performance - false positive rate, false-negative rate, true positive rate, and true negative rate. For the gender detection classifier, the model is trained to detect the category male or

females from each face at the time. The flow process and the input dataset are different for both systems.

The matching process requires a database with multiple instances of a subject, including the probe, compared to all other images in the gallery. In the classification process, we input one image in the model and get our prediction result. This simple difference of how the two methods start clearly shows that the output from one system cannot be directly associated with the second without any evidence.

Gender classification bias

Many other investigations followed the gender shades report trying to provide an approach to explain the error gap between ethnicity and proposed improvements to correct the biases. Before the outbreak of discussions following the gender shades paper, in 2015 [22], Carcagnì et al. carried on an experiment under a controlled and real-world environment to evaluate the accuracy and robustness of a soft biometric classifier based on 1) the features extractor (LBP, CLBP, HOG, SWLD), 2) training dataset (balanced or unbalanced), 3) use of scaling or non-scaling approaches in the framework including data reduction step (LDA) and finally SVM as the classifier. Their experiment regarding gender prediction, which is the interest of this thesis, concluded that a framework built with a CLPB descriptor, using an LDA projector, and trained with a balanced dataset non-scaled input achieved better accuracy in a controlled or real-world environment. This result showed that many factors could explain the error gap noticed in the classification algorithms, and one needs evidence before any assessment.

I. Serna, A. Peña, A. Morales, and J. Fierrez [24] dig deeper to look at the deep inside network to assess how bias impacts the activation of gender detection. They inserted color bias in the MIST dataset and trained the model. Their key finding was 1) when the training data was unbalanced, they obtained higher activation

function for the dominant color, and when tested with uniform color distribution, the overall performance decreased. 2) The second model trained without bias produced similar activation functions for the different colors, reduced the performance gap between testing groups, and improved overall accuracy. They showed that the activation level is susceptible to ethnic attributes and revealed that the biases are heavily encoded in the models' last layers, which is a hidden behavior during the learning process. Hence, they proposed a novel method, inside bias, for earlier detection of biases through layer activation and suggested a heterogeneous dataset for training.

Furthermore, Kim et al. [25] came up with a framework of multi-accuracy auditing and post-processing to improve predictor accuracy across identifiable subgroups (Multi-accuracy boost). They replicated the gender shades experiment to test the effectiveness of their algorithm. They trained an inception-ResNet-v1 using the CelebA data set, which gives 98% accuracy, which confirmed the initial research. Then they applied the multi-accuracy boost using the PPB data set, which has a balanced representation across gender and race. They used ridge regression and data derived from a VAE trained on a celeb A dataset using facet, which reduces the dataset's size to 855 for the audit and 415 individual images. The result showed a significant improvement in the error rate after post-processing, where the error for Darker Female dropped from 39.8 to 12.5, and in general, the error rate for darker skin tone passed from 18.8 to 7.3 and light skin tone from 2.2 to 0.9. After retraining the whole dataset, the classification error was 2.2, and a female was 3.8 compared to a male 0.9. The techniques help improve the accuracy without harming the population that is already accurately classified.

Later on, Radji et al. [26] reported the new performance metrics from the gender shade of the targeted companies IBM, Microsoft, and Face ++ and investigated new targets, Amazon and Kairos. The result showed that the first targeted companies improved significantly on the darker females' cohort and thus reduced the gap between the genders from 17.7%-30.4% initially to 5.7%- 8.3% within seven

months. IBM implied that the training dataset was the main factor of improvement. Nevertheless, the former companies displayed the same significant disparities between gender and performed worse for darker females. The authors conclude that better work could normalize automation systems and reduce unfairness with prioritization and government and public pressure.

Face Recognition bias

Face recognition has been known to face several challenges in image quality and variation in illumination and the processing methods used to operate them. Many algorithms have been tested to assess the impact of demographics on accuracy. Klare et al. in 2012 evaluated six different recognition systems, three commercial, two nontrainable, and one trainable [27]. They used eight different cohorts based on gender (Male, female), ethnicity (black, white, Hispanic), and age group (18-30, 30-50, 50-70) as input. As a result of their experiment, it was determined that the commercial and the nontrainable algorithms have lower matching accuracy for females, black, and those in the age group 18-30.

Similarly, Cook et al. examined the effect of demographic factors on eleven commercial face algorithms[28]. They showed that both efficiency (transaction times) and accuracy are affected by multiple covariances such as gender, age, eyewear, height, and skin reflectance. Regarding skin tone, the results showed it has the most robust net linear effect on the performance and that darker skin is associated with lower efficiency and accuracy.

Krishnapriya et al. further analyzed the False Match (FMR) and False Non-Match Rate (FNMR) to evaluate face recognition accuracy systems relative to race in the Morph dataset. Their result shows that the African American curve has a higher False-Match rate than the Caucasian curve from all four matches. Regarding the ROC curve, even though it is not the best tool to evaluate the system's performance, it was noticed that from two recent-matcher Cots-B and Resnet, the African

American True Positive Rate (TPR) is higher than the Caucasian one. The general pattern is that the African American cohort has an impostor distribution toward higher (better) match scores and has a genuine distribution toward higher (better) match scores. Thus, the African American cohort has a higher false match rate (FMR) and a lower false non-match rate for a given decision threshold than the Caucasian mate for a given matcher. The experiment extends by applying the ICAO-complaint on images to get the same quality to reduce the gap between the two cohorts. To conclude, despite the disadvantage of African Americans' FMR, the d-prime from some matches shows that the distribution is equal regarding both cohorts [29].

J. G. Cavazos et al. work discussed the challenges of data-driven factors (image quality, image population statistics, and algorithm architecture) and scenario modeling factors (threshold decision and demographics constraints). Their experiment tested four algorithms (A2011, A2015, A2017b, A2019) with Asian and Caucasian as testing datasets. From their investigation, a race bias was noticed at a low false acceptance rate and demonstrated that overall accuracy (threshold-independent) and accuracy at a pre-determined threshold led to different results. They stated that it is difficult to assess all the factors that could impact a biometric recognition system. A general assessment of bias for face recognition is unfeasible without measuring each scenario, algorithm, race, and dataset [30].

Albiero et al. investigated the cause of more significant gaps in face recognition accuracy between men and women. They develop five speculated reasons: 1) the facial expression – females exhibit a broader range of facial expressions. At the same time, males appear neutral through their photo shoots which led to higher similarity compared to the formers. 2) head pose – an off-angle pose affects more females than males if the camera is adjusted to a male's height. 3) forehead occlusion – females mostly have their hair occlude the forehead and eyes, and removing the forehead occlusions improves the d-prime for the genders. 4) facial makeup – two subjects with the same makeup style are likely to have a high

matching score, and the same person pair with and without makeup could generate a lower matching score than if the pair has the same makeup condition. When tested, females without makeup have a higher matching score as well as a high impostor rate. 5) balanced training dataset – even though they trained the algorithm with an explicitly balanced gender dataset, the female impostor and genuine distribution are closer together than the male distributions. However, even with all these factors excluded, females are still at a disadvantage compared to males. The authors suggested the need to dig further into face morphology between men and women [31].

Instead, Alasadi et al. proposed an adversarial in-depth learning-based approach to maintain the accuracy in face recognition on disparities in the demographic population. Their experiment used a framework to maximize the images' quality and minimize the network's ability to infer the demographic properties. They created a network with two competitive tasks to match faces. One path took low-resolution images (two convolutional layers followed by flattening concatenating layers). The second one is the high-resolution image (two convolutional layers and Maxpooling, followed by flattening and a set of fully connected nodes). The output must tell if the input belongs to the same person or not for the first branch; the second is male or female. They used Celeb A and UMD faces with the focus on gender as a sensitive attribute. The author used a GAN-Based architecture to reduce the disparities in accuracy across different genders across accuracy, true positive rate, and false-positive rate [25].

Whether it is a gender classification system or face recognition, one factor could not answer biases. It will take fine-grained analysis based on a case basis to start the discussion.

Relationship between face recognition and gender classification errors

To date, few related works are associated with our main question, to know if there is a possible correlation between gender classification and face recognition.

Qui et al. paper was the first to investigate the question and found out that there is a varying relationship across the different demographic groups when analyzing face recognition and gender from face errors. They experimented with three classifiers (open source, Amazon, and Microsoft) and two face matching algorithms [32].

They looked at the variation between the numbers of pairs when one, both, and any images participate in gender error. Their investigation showed that images that resulted in gender classification errors recorded fewer examples of false matching than false non-matching. The gender classification errors represent an insignificant proportion of the impostor distribution. We will follow up on their work by providing a new way of analyzing the correlate errors generated from both systems. Our contribution is to analyze the skin tone effect on the performance of the biometric system.

Chapter 4

Experiment set-up

Our experiment aims to answer the questions mentioned above. Do errors in face analytic carry over to face recognition? Do these errors involve subjects mostly with dark skin tone? There has not been enough research to clarify those questions yet. This experiment will provide the link between errors in face analytics and face recognition relative to skin tone.

For our experiment, we first evaluate the PPB dataset with an open-source algorithm to confirm the result from gender shade. We go further by classifying the skin type of each subject based on the ITA angle. The goal is to show if more errors are from subjects classified between the three to six skin types. Once our results align with the prior research, we extend our work to a larger dataset suitable to perform face recognition. Face recognition requires two or more images per subject to complete the matching process and result in True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The PPB dataset contains one image per subject that does not satisfy the matching process requirements.

The extension of our work used the MORPH dataset, which has an intersectional demographic and several images per subject for face recognition and facial analysis.

In this second experiment for gender classification, we used three classifiers, one Open-source and two commercials' APIs (Amazon, Microsoft) and the open-source algorithm (Arcface) for the recognition process. We then analyzed the correlated errors from both systems and investigated the skin tone influence using the automated skin tone classifier.

Gender Classification (GC) algorithms

For the gender classification experiment, we provide detail of the open-source algorithm since the commercials are proprietary and close algorithms.

The open-source Convolutional Neural Network (CNN) with 50 layers algorithm [34] was implemented in PyTorch. The gender classification model trained using ResNet-50 modified ArcFace with parallel acceleration on both features and centers proposed by [33]. The introduction of a subcenter number to ArcFace releases the intra-class compactness constraint[34] and weighted binary cross-entropy as the loss function. Each gender is weighted according to the number of samples. The following collection of datasets was used for training and validation sets: AAF[35], AFAD[36], AgeDB[37], CACD[38], IMDB-WIKI[39], IMFDB[35], MegaAgeAsian[40], and UTKFace[41].

Face Recognition (FR) matcher

Face recognition is performed using ArcFace, a state-of-the-art deep CNN matcher. The instance of ArcFace used here corresponds to a set of publicly available weights trained on the MS1MV2 dataset, a publicly available, "cleaned" version of MS1M. The impostor distributions for African American male, African American female, Caucasian male, and Caucasian female cohorts of the MORPH3 dataset are sampled at 1-in-10,000 (high-similarity tail) FMR threshold settings to analyze the impact of skin tone on false matches. The 1-in-10,000 FMR threshold for Caucasian males is taken as the baseline for all other cohorts because it is the demographic that usually gives the lowest false match rate, which is also in agreement with NIST's approach.

Skin Tone (SK) classifier

The skin type assessment is based on an automated skin tone rating algorithm implemented using the Individual Typology Angle (ITA) measurement. It facilitates skin tone determination directly from the images and has been adopted as a practical technique to categorize skin color in many studies[42]. The selected skin pixels from the image are converted to CIE Lab color space to obtain the L and b values, where L represents the luminance or lightness, and b represents the chromaticity coordinate from blue to yellow. ITA calculated according to the equation below:

$$ITA = \frac{\arctan\left(\frac{(L-50)}{b}\right) * 180}{\pi}$$

The ITA values computed for the image are classified into six skin types I (sk-1: lighter) to VI (sk-6: darker) in categories, namely: very light, light, intermediate, tan, brown, and dark. Here we customize the ITA ranges to minimize the overlap and achieve better consistency with the automated approach. The customized ranges are: Sk-1: $ITA \geq 50$, sk-2: $25 \leq ITA < 50$, sk-3: $0 \leq ITA < 25$, sk-4: $-25 \leq ITA < 0$, sk-5: $-50 \leq ITA < -25$, sk-6 $ITA < -50$.

PPB dataset

The PPB dataset, including 1,255 individual subjects, was created by Buolowini et al. to address the non-existence of a well-balanced dataset regarding gender, race, and skin color[7]. It includes parliament members from three countries of Europe (Iceland, Sweden, Finland) and three countries of Africa (Rwanda, Senegal, South Africa). The dataset is well balanced, with 55% males and 45% females. The number of male or female subjects from each country is listed in Table 1 below.

Table 1. PPB dataset split by gender and country

| <i>PPB Data</i> | <i>Male</i> | <i>Female</i> |
|-----------------|-------------|---------------|
| <i>FL</i> | <i>114</i> | <i>83</i> |
| <i>SA</i> | <i>255</i> | <i>181</i> |
| <i>SW</i> | <i>186</i> | <i>163</i> |
| <i>SE</i> | <i>80</i> | <i>64</i> |
| <i>IL</i> | <i>33</i> | <i>30</i> |
| <i>RW</i> | <i>26</i> | <i>40</i> |
| <i>Total</i> | <i>694</i> | <i>561</i> |

MORPH dataset

The Morph dataset is a large dataset with a cross demographic cohort and gender, African American, Caucasian, Male, and Female. The MORPH dataset was initially collected to support research in facial aging. MORPH contains mugshot-style images that are nominally frontal pose, neutral expression, and acquired with controlled lighting and an 18% gray background. We curated the MORPH 3 dataset to remove duplicate images, twins, and mislabeled images. The curated version contains 35,276 photos of 8,835 Caucasian males (28%), 10,941 images of 2,798 Caucasian females (9%), 56,245 pictures of 8,839 African American males (44%), and 24,857 images of 5,929 African American females (19%).

As displayed in Table 2 below, the gender classification results in a very minimal percentage compared to the face matching output. The prior will generate the number of results equal to the number of the sample present in the dataset while the latter will generate million to billions comparison for the non-mated matching

Table 2. GC and FR result samples

| <i>Morph cohorts</i> | <i>Gender classification</i> | <i>Face matching non-mated pairs</i> |
|-----------------------------|-------------------------------------|---|
| <i>AAF</i> | <i>24,857</i> | <i>308,840,189</i> |
| <i>AAM</i> | <i>56,245</i> | <i>1,581,426,316</i> |
| <i>CF</i> | <i>10,941</i> | <i>59,813,525</i> |
| <i>CM</i> | <i>35,276</i> | <i>622,042,698</i> |

Chapter 5

Experiment results

This section focuses on our experiment results. We started the gender classification and the skin tone analysis on the PPB dataset to confirm results from previous works. Following this, we used the MORPH dataset to continue our investigation.

PPB gender classification (GC) results

We ran the gender classification twice on the PPB dataset. First, we used a model trained with the morph dataset plus all previously mentioned datasets in the algorithm description section. We used the raw picture from the dataset as input. Secondly, we used a model trained without the morph dataset and with cropped PPB images as input.

Overall, the experiment result present in Table 3 shows a better performance for the 1st test (89% accuracy) than the 2nd test (77%). The cross-gender accuracy offers 95% accuracy for males and 83% for females for the former. In comparison, the last experiment recorded 96% for males and 54.4% for females, a significant error gap of 41.6% between males and females. The former has a difference of only 12%.

Table 3. Gender prediction result from the 1st and 2nd experiment

| <i>PPB GC Accuracy</i> | Total ACC | Male | Female | Error gap |
|-----------------------------------|------------------|-------------|---------------|------------------|
| <i>1st test</i> | 89% | 95% | 83% | 12% |
| <i>2nd test</i> | 77% | 96% | 54% | 42% |

The performance difference between the two experiments can be explained partly by the dataset's quality used to train the system. In the first experiment, adding the morph dataset to the training plays a significant role in the improvement; the

balance between male and female and African American and Caucasian representation in the dataset enhances the system's environment. Hence, using the PPB dataset for the testing input, which comprises parliament members from Europe and Africa, the system recognized most of the patterns. In the second experiment, however, most of the training datasets are mainly from an Asian population (AAF, AFAD, MegaAgeAsian) and the rest from celebrities where the representation of all demographics is limited. Figure 3 shows the accuracy and error plots of the two training models.

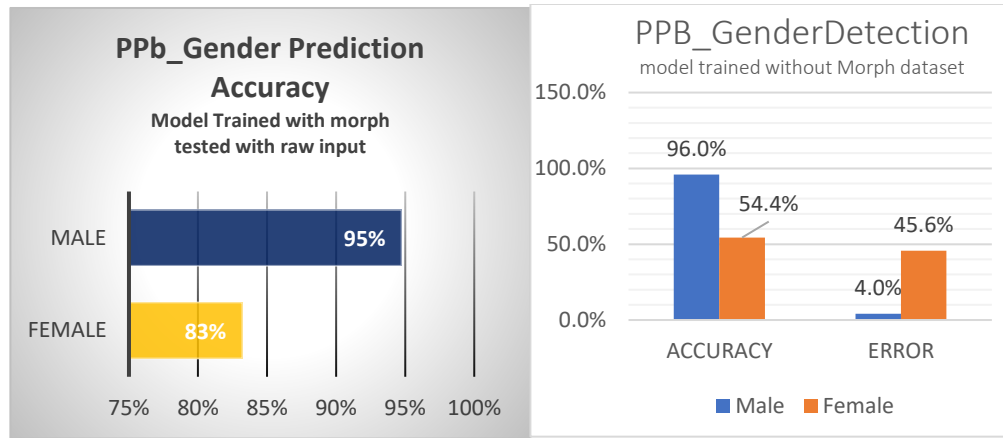


Figure 3. PPB GC results "with Morph used in training" and result "without Morph used in training"

Our PPB gender prediction test results compared to the gender shades [7] work show the same pattern (Figure 4) encountered where all three commercial algorithms have higher accuracy for males than females. Moreover, our second test corroborates one of the popular assessments about the training dataset with the significant error gap in our second test—the more various and richer the training dataset, the more accurate the system.













| Gender Classifier | Female Subjects Accuracy | Male Subjects Accuracy | Error Rate Difference |
|---|--|--|--|
|  Microsoft | 89.3%  | 97.4%  | 8.1%  |
|  FACE++ | 78.7%  | 99.3%  | 20.6%  |
|  IBM | 79.7%  | 94.4%  | 14.7%  |

Figure 4. Accuracy table from gender shade

We kept the second test result where the model is trained without the morph dataset for consistency for our further analysis in this thesis. This dataset will be used for the test experiment later in this work.

PPB GC with Skin tone classification

In the following section, we assess the skin type distribution of the result from the second experiment to evaluate which skin color registered the most errors. The count distribution of the accuracy and error numbers is summarized in the table below.

- **PPB Female cohort**

For the female cohort, out of 561 total subjects, the distribution across the skin tone shows (Table 4) that the lighter skin tone subjects (sk-1 to sk-3) are more represented (80.2% of the total female images) than darker skin tone subjects 19.6% (sk-4 to sk-6).

A closer look at the accuracy and errors level in the distribution for each skin tone reveals that all the darker skin types have higher incorrect predictions than correct predictions compared to lighter subjects. For example, skin type sk-5 with a total of 25 images records 22 inaccurate predictions (88%), and only three images were correctly predicted; the same case is valid with skin type SK-6, which has a total of 18 shots with 95% errors.

Table 4. PPB female classification raw results

| PPB Female | Sk-1 | Sk-2 | Sk-3 | Sk-4 | Sk-5 | Sk-6 | Total |
|--------------|------|------|------------|------|------|------|-------|
| Correct GC | 15 | 121 | 183 | 29 | 3 | 1 | 305 |
| Incorrect GC | 5 | 34 | 92 | 38 | 22 | 17 | 256 |
| Total | 20 | 155 | 275 | 67 | 25 | 18 | 561 |
| Light : 450 | | | Dark : 110 | | | | |

However, when assessing the relative frequency of each skin type from the total number of correctly and incorrectly predicted subjects, the result yields a different interpretation. The more the skin type represented, the higher the frequency of either accuracy or errors. The skin types sk-1, sk-3, and sk-4 (

Table 5) have dominant accuracy and errors, following the curve represented in the dataset.

Table 5. PPB female relative frequency results

| PPB_Female | Sk-1 | Sk-2 | Sk-3 | Sk-4 | Sk-5 | Sk-6 |
|-----------------------------|-------|-------|-------|-------|-------|-------|
| Relative_Frequency_accuracy | 0.049 | 0.396 | 0.6 | 0.095 | 0.009 | 0.003 |
| Relative_Frequency_Error | 0.019 | 0.132 | 0.359 | 0.148 | 0.085 | 0.066 |

The analysis based on each skin type's total number and the distribution between correct and incorrect predictions corroborates the prior studies and the hypothesis that darker skin tone disadvantages more than lighter skin tone. However, the relative frequency of each skin type out of the total number of correct or incorrect predictions without considering their initial representation in the dataset can conclude that lighter subjects have a higher error rate than darker.

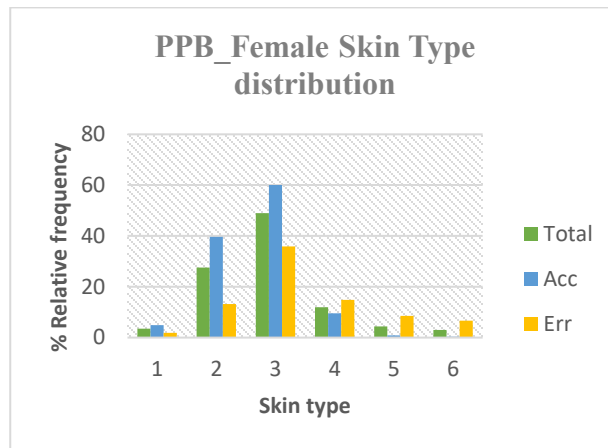


Figure 5.PPB female comparative skin tone plot

As is shown in Figure 5 above, we can see that the skin types sk-4, sk-5, and sk-6 have their error rate (yellow bar) higher than their accuracy (blue bar).

The skin tone cross-comparison shows that sk-2 and sk-3 are dominant regarding the accuracy, where sk-3, sk-4, and sk-2 are dominant in terms of errors. The skin type distribution from the errors rate is represented in Figure 6 below.

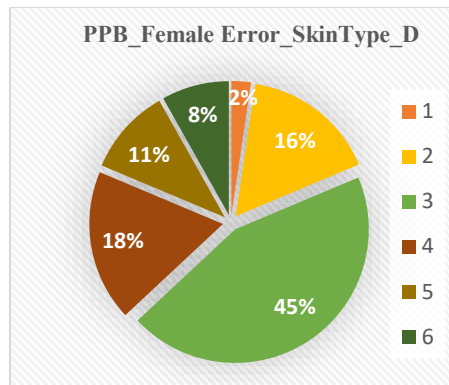


Figure 6.PPB female errors per skin tone

- **PPB Male Cohort**

For the male cohort (Table 6), most subjects are concentrated in skin type II (142 total images), III (297 images), and IV (132 images). Similarly, as for the female cohort, lighter subjects are more represented than darker subjects.

Table 6.PPB Male Classification raw results.

| Male | Sk-1 | Sk-2 | Sk-3 | Sk-4 | Sk-5 | Sk-6 | Total |
|---------------------|-------------|-------------|-------------|-------------|-------------|-------------|--------------|
| Correct GC. | 37 | 134 | 283 | 130 | 44 | 37 | 666 |
| Incorrect GC | 2 | 8 | 14 | 2 | 2 | 0 | 28 |
| Total | 39 | 142 | 297 | 132 | 46 | 37 | 694 |
| | 471 | | | 215 | | | |

Only 28 images were misgendered out of the 694 total images for the entire male cohort regarding the false predicted subjects. The error rate analysis (Table 7) across that skin type showed that skin type sk-3 got 50 % of the error rate (14 misclassified images out of the 28 total incorrect images), followed by skin type sk-2 with 28% (8 wrong photos). The remaining six misgendered images were distributed equally (7.1%) between the skin types sk-1, sk-4, and sk-5 (2 incorrect images each). Skin type sk-6 scored zero errors in this case. The result showed that lighter subjects recorded an 85.1% error rate higher than the darker skin type (14.2%).

Table 7.PPB Male Classification relative frequency results

| PPB_Male | Sk-1 | Sk-2 | Sk-3 | Sk-4 | Sk-5 | Sk-6 |
|------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| %Accuracy | 5.5 | 20.1 | 42.4 | 19.5 | 6.6 | 5.5 |
| % Error | 7.1 | 28.5 | 50 | 7.1 | 7.1 | 0 |

The vertical analysis for each skin type showed the following error rate: for lighter skin type (sk-1 – 5.1%, sk-2 – 5.6%, sk-3– 4.7%) and toward darker skin tone, we

got 1.5% for type sk-4, 4.3 % for sk-5, and 0 error rate for the skin type sk-6, which is the lowest.

Figure 7 below shows that the skin tone comparison shows that the error bar(gray) has higher accuracy (orange bar) across the lighter skin types and slighter higher for sk-5; however, the skin types sk-4 and sk-6 had the lowest error rate.

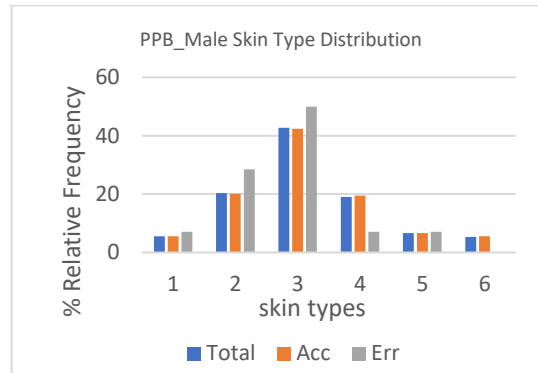


Figure 7.PPB Male comparative skin tone plot.

In addition to the first observations where lighter subjects recorded more errors than darker subjects, we can see the same pattern if we further compare the skin type based on their total samples. For example, the sk-1 (with 39 images) compared to sk-6 (46 photos) and sk-4 (37 images) showed a 5% error rate for sk-1, 4.3% for sk-5, and 0% for the last one; similarly, the skin type sk-2 (142 images) with the only difference of 10 images from sk-4 (132 images) has scored 5.6% where the latter error is 4.3%.

Finally, to compare our result to the one from gender shades as represented below, we add up the skin tone type from 1 to 3 as lighter skin and from 4 to 6 as darker. We got a 79.5% error rate for females with darker skin type while the error rate was only 26.7% for lighter skin tone compared to the male cohort, where we got an 85.1% error rate for lighter subjects compared to 14.2% for darker males. The result confirms the conclusion from the gender shade where darker females have a

significant disadvantage across gender and skin tone classification. However, the male cohort results in our experiment follow the same pattern noticed with the classifier Face ++ in gender shade where darker subjects performed better than lighter subjects (male and female).

MORPH dataset GC results

In this second part, we extend our work using a new dataset, add one open-source classifier result, and explore two commercial classifiers results (Amazon and Microsoft APIs). We run the gender classification with the Morph Dataset as input. As previously mentioned, the fact that the PPB dataset has only one image per subject prevents us from using it for face matching. Hence the Morph3 dataset with approximately 3 to 6 images per subject is suitable for gender classification and face matching for the rest of the investigation and further analyzing the correlation between the gender classification and the face matching result based on the skin tone distribution.

The gender correct classification and misclassification on the morph dataset from the three algorithms present in Table 8 below shows the higher accurate prediction for Microsoft API, followed by the Amazon algorithm and the Open-source algorithm.

The African American cohort (Male and Female) ranks higher in incorrect prediction than the Caucasian demographic groups for the three classifiers. For cross-gender comparison, Males performed better than Females within the same demographic groups, and African American females recorded the highest incorrect prediction.

Table 8. Morph dataset GC result with three classifiers.

| Classification Results on Morph Cohorts | Open-source | | Amazon | | Microsoft | |
|---|----------------|------------------|----------------|------------------|----------------|------------------|
| | Correct gender | Incorrect gender | Correct gender | Incorrect gender | Correct gender | Incorrect gender |
| AAF | 20676 | 4181 | 23098 | 1759 | 23927 | 926 |
| AAM | 55090 | 1155 | 55192 | 1053 | 55805 | 405 |
| CF | 10022 | 919 | 10710 | 231 | 10829 | 111 |
| CM | 34993 | 283 | 35105 | 171 | 35203 | 41 |

Detailed analysis of the open-source GC results

Since the open-source algorithm has a more incorrect prediction, we focus on it for more investigation. Overall, the classification output from the GC on the morph dataset performed very well (Table 9), with a total accuracy of 94.5% and a 5.1% error rate. However, the Morph demographic cross-comparison follows the curve seen with the PPB dataset. The African American female cohort holds 4th place by ranking the accuracy result, with 83.2%. The Caucasian female's cohort recorded 91% for 3rd place, then African American Males cohort with 97.95%, and the Caucasian Males with the accuracy of 99.2%.

The result confirmed the general hypothesis: males perform better than females, and Caucasians have less error rate regarding the race demographic. The table below summarized the results of the experiment.

Table 9. Morph dataset GC result with the Open-source classifier.

| MORPH | Total Images | Correct GC | Incorrect GC | Accuracy | Error |
|--------------|---------------|---------------|--------------|---------------|--------------|
| AAF | 24857 | 20676 | 4181 | 83.2% | 16.8% |
| CF | 10941 | 10022 | 919 | 91.6% | 8.4% |
| AAM | 56245 | 55090 | 1155 | 97.95% | 2.05% |
| CM | 35276 | 34993 | 283 | 99.20% | 0.80% |
| Total | 127319 | 120781 | 6538 | 94.9% | 5.1% |

Even though the morph dataset is already grouped into four different race demographics, we cannot infer a conclusion based on the labeling of each group without assessing the skin types of each image.

Open-source GC result with Skin type classification

- **African American Female (AAF)**

The African American Female (AAF) cohort has 24,857 total images. The skin types of distribution are summarized in Table 10 below, assessing the correct and incorrect prediction count and the entire image representation for each skin type.

Table 10. AAF Open-source GC result per skin tone

| AAF | Sk-6 | Sk-5 | Sk-4 | Sk-3 | Sk-2 | Sk-1 |
|------------------|--------------|------|------|--------------|------|------|
| Correct | 354 | 4072 | 8466 | 6034 | 1639 | 107 |
| Incorrect | 185 | 1090 | 1506 | 1076 | 309 | 15 |
| Total | 539 | 5162 | 9972 | 7110 | 1948 | 122 |
| | Dark : 15673 | | | Light : 9180 | | |

The total skin tone distribution of the Africa American Female (AAF) cohort shows that the darker subjects are more represented (63%) than, the lighter subjects (39%). The output result follows the data representation (

Figure 8), with an average of 60% per darker subject and 30 percent for the lighter skin subject, as shown in the table below.

Looking at the individual skin types, the skin types sk-4, sk-3, and sk-5 scored the highest accuracy at 41%, 29%, and 20%, respectively. In terms of error rate, those are the top three as well.

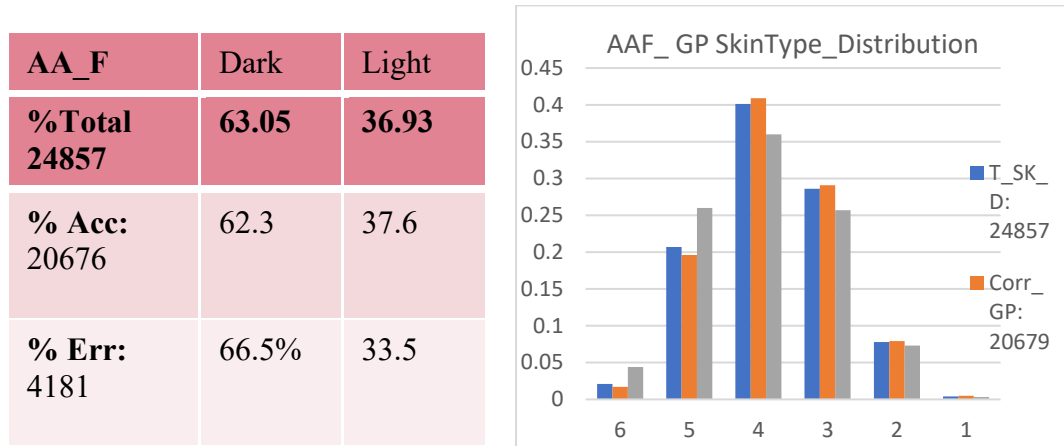


Figure 8. AAF Comparative skin tone table and plot

From this first cohort, the results are proportional to the total representation across the skin tone. As we can see on the graph above, the highest one for both accuracy and errors is skin type IV, representing 40% of the total images in the cohort. However, the cross-comparison showed that skin types VI and V have their error rate higher than their accuracy rate, which means they perform worse than the rest of the represented skin types.

- **Caucasian Female (CF)**

The Caucasian female cohort has 10,941 total images split into skin tones (Table 11) result in 3.5% darker subjects and 96.5% lighter subjects' representation. The gender classification performed well, with an overall accuracy of 91.8%.

Table 11.CF Open-source GC result per skin tone

| C_F | Sk-6 | Sk-5 | Sk-4 | Sk-3 | Sk-2 | Sk-1 |
|------------------------|------|------|------|---------------|------|------|
| Correct : 10022 | 2 | 58 | 288 | 846 | 3576 | 5252 |
| Incorrect : 919 | 0 | 5 | 30 | 76 | 328 | 480 |
| Total | 2 | 63 | 318 | 922 | 3904 | 5732 |
| Dark: 383 | | | | Light : 10558 | | |

The relative comparison across the skin types showed that the more represented is the data, the higher the accuracy or error rate, and the relative frequency seems equal for both accuracy and error rate (Figure 9).

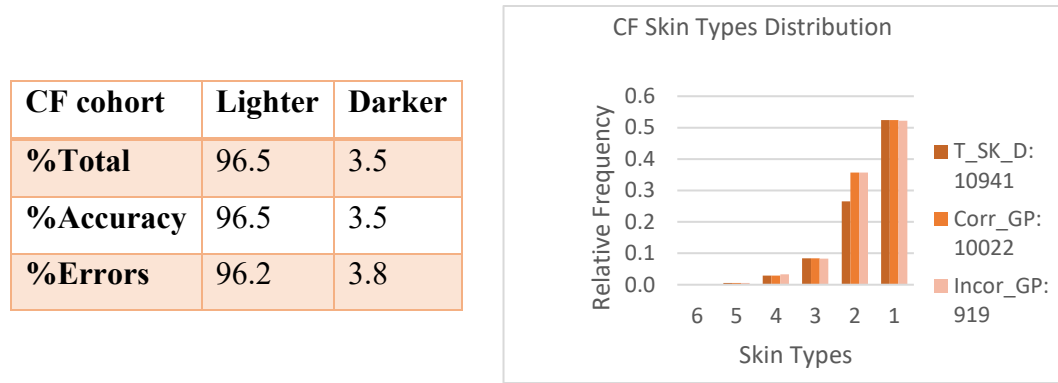


Figure 9. CF comparative skin tone table and plot

Overall, even though lighter subjects perform well due to their more significant representation, the darker subject rate did not perform poorly. The result follows the pattern of the data as we noticed that: darker subjects with 3.5% total representation got 3.8% errors out of the total of 919 incorrect predictions and 3.5% accuracy out of 10,022 correct predictions; similarly, lighter subjects represent in total 96.5% with 96.2 errors and 96.5% accuracy.

- **African American Male**

The African American Male with 56,245 total subjects recorded overall 97.95% accuracy with a 2.05% error rate. The complete skin tone representation raw numbers (Table 12) yield 72% for darker skin tone and 28% for lighter skin tone, and the accuracy rate follows the same proportion. The skin types sk-2, sk-4, and sk-5, are more represented and scored higher accuracy and higher error rate as noticed from the female's cohort.

Table 12. AAM Open-source GC result per skin tone

| AAM | Sk-6 | Sk-5 | Sk-4 | Sk-3 | Sk-2 | Sk-1 |
|---------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| <i>Correct GC</i> | 3123 | 15788 | 20750 | 12585 | 2690 | 139 |
| <i>Incorrect GC</i> | 61 | 404 | 464 | 189 | 28 | 9 |
| <i>Total</i> | 3184 | 16192 | 21214 | 12774 | 2718 | 148 |
| | 40590 | | | 15640 | | |

The comparison of errors rate in terms of the two skin tone (Figure 10) groups showed that darker skin tone subjects generated a higher error, 80.4%, than lighter subjects, 19.6%.

| AAM | Dark | Light |
|---------------|-------------|--------------|
| %Total | 72 | 28 |
| %Acc | 72 | 28 |
| %Err | 80.4 | 19.6 |

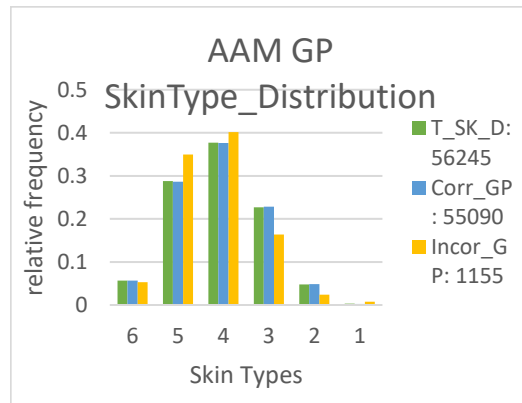


Figure 10. AAM Comparative skin tone table and plot.

- Caucasian Male**

The Caucasian male cohort has 95% lighter subjects and 5% darker subjects. With an overall accuracy of 99%, it outperformed the rest of the demographic. Regarding the skin tone, as previously seen with the Caucasian female, the result follows the curve of the total representation of the data. Lighter subjects perform very well with 95% accuracy while the darker subject has 5% accuracy; for errors, we got 91% for the light skin type and 9% for the dark skin type (Figure 11).

Table 13. CM Open-source GC result per skin tone

| | Sk-6 | Sk-5 | Sk-4 | Sk-3 | Sk-2 | Sk-1 |
|---------------|-------------|------|------|---------------|-------|-------|
| CM_Acc | 22 | 420 | 1287 | 2588 | 12319 | 18356 |
| CM_Err | 0 | 2 | 23 | 38 | 92 | 128 |
| SK_T | 22 | 422 | 1310 | 2626 | 12411 | 18484 |
| | Dark : 1754 | | | Light : 33521 | | |

| CM cohort | Lighter | Darker |
|------------------|---------|--------|
| %Total | 95 | 5 |
| %Accuracy | 95 | 5 |
| %Errors | 91 | 9 |

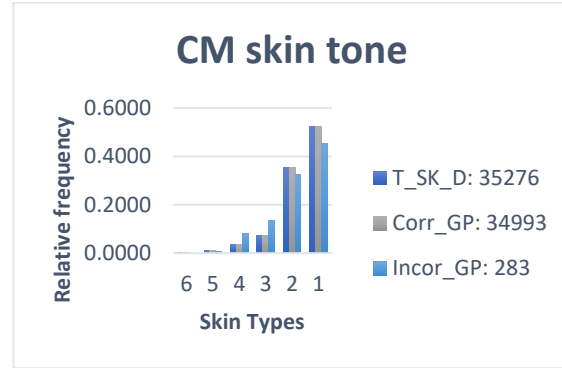


Figure 11. CM Comparative skin tone table and plot.

Comparison across the demographic cohort

The observation across all the demographics showed that the accuracy and error rate for the skin tone distribution tend to follow the data representation in each cohort. However, a close look at each plot showed that skin type sk-5 to sk-6, the error bar is often higher than the accuracy bar for all cohorts.

Table 14 below is the cross-comparison between the entire Morph dataset with categories lighter and darker subject and the distinction between male and female; we can see that lighter subjects outperformed darker subjects. In these two subcategories, females scored more errors than males.

Table 14. Open-souce GC cross skin tone comparaisn.

| Morph Dataset | Accuracy | Error |
|------------------|----------|-------|
| Lighter subjects | 58% | 4% |
| Darker subjects | 32% | 32% |
| Lighter Males | 60% | 1% |
| Lighter Females | 55% | 7% |
| Darker Males | 37% | 1% |
| Darker Females | 27% | 6% |

Face matching (FM) results on the Morph dataset

For the face matching result, we provide a brief description of the process and present the resulting distributions. The focus is to analyze the impact of gender classification on its outcome. We showed the distribution plots for the non-mated and mated similarity scores before moving on with our primary investigation.

As a reminder on the face matching experiment process, a probe image is matched against all the samples in the database, which will generate a comparison score. As is shown on the distribution plot below for African American and Caucasian groups generated from the ArcFace matching algorithm, on one side, when the probe is matched against an image of another subject, we have the impostor score. On the other side, we have a genuine score matching the quest against another shot of the same subject. From the score, a decision (accept or reject) is based on the threshold and set according to the application field of the biometric system. If the score generated from the comparison pair is higher than the threshold, the decision Accept is applied; otherwise, the subject is rejected. When the score generated from the impostor comparison is higher than the threshold, which is supposed to be lower, a wrong Accept decision is made, which is an error named False Positive.

Similarly, when an authentic pair generates a lower score than the threshold, the probe is falsely rejected, False-negative. From those two types of errors, the performance and the accuracy of the biometric systems are evaluated. Hence the goal is to set a threshold to minimize those errors. Figure 12 shows the d-prime, which is the distance between the authentic and the impostor—the higher the d-prime, the better the system's performance.

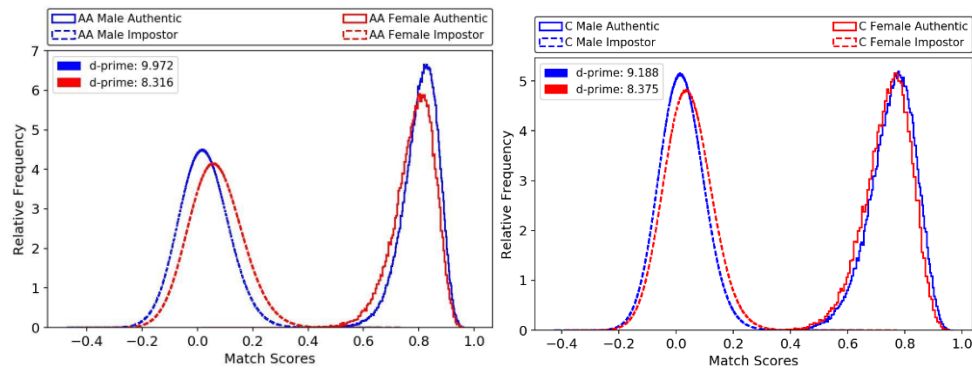


Figure 12.Face Matching distribution score

Chapter 6

Gender classification errors and face matching analysis

From the similarity scores produced in the face matching, we conducted analyses to answer the following question: **Do images that result in gender misclassification have more decisive matches in the non-mated pair distribution, which could yield a high number of false matches?**

For the test, we segment the pair distribution into six bins, with each bin having a width of 1 standard deviation. We then determine the number of comparisons that fall within each bin. If the gender misclassified images strongly influence the false match rate of a face recognition algorithm, on the one hand, they would need to produce higher similarity scores for non-mated comparisons and therefore be concentrated at +2 std and +3std. On the other, they would produce a higher similarity score for mated comparison at -3 and -2 std to strongly influence the false non-match rate.

GC errors placement across FM non -mated score distribution

We evaluated the errors from the three classifiers by analyzing their occurrence within the standard deviation range of the impostor distribution for each of the four demographic cohorts.

We display the initial impostor distribution for each cohort. The African American female (AAF) has a total of 308, 840,189 comparisons pairs slightly skewed with 51% data below the mean and 49% above the mean split within each standard deviation range as follow: (-1std: **35%**, 1std: **33%**, -2std: **14%**, 2std: **13%**, -3std: **2%**, 3std: **3%**). The African American Male (AAM) has a total of 1,581,426,316 pairs with a similar pattern as AAF with more value at negative one std: (-1std: **35%**, 1std: **33%**, -2std: **14%**, 2std: **13%**, -3std: **2%**, 3std: **3%**). The Caucasian Female (CF) with a total of 59,813,525 pairs and the Caucasian Male (CM),

622,042,698 have the same percentage within the range (-1std: **35%**, 1std: **33%**, -2std: **14%**, 2std: **13%**, -3std: **2%**, 3std: **3%**).

We analyze the percentage of match scores that involve a gender misclassified image within each bin. Because face recognition match comparisons involve two images, we first examine comparisons made when precisely one of the two images was misclassified relative to gender, displayed in Table 15. Secondly, we look at cases where both images produced a gender misclassification present in Table 16.

Our analysis finds comparisons involving gender misclassified images represented in each bin from -3 std to +3 std of the non-mated pair distribution represented by percent below.

The analysis showed that we have a highly concentrated number at 68% and 95% of the distribution for the three classifier algorithms. In the first case, when one image was involved in the classification errors, the total comparisons at negative three and negative two standard range to the positive side, and we have more similarity score engaged in the prior than the latter. **Hence the influence expected from the gender classification on the false matching error cannot be verified.**

Table 15. One impostor image involved in GC errors split by std range

| Percentage per std bins where one impostor image involved in gender er Open-source (OP), Amazon (AM), Microsoft t(MI) | | | | | | | | |
|--|---|------------|-------|-------|-----------|-----------|-----------|----------|
| Non-mated | Total Comparisons involving in misclassification errors | | -3std | -2std | -1std | 1std | 2std | 3std |
| AAF 308,840,189 | OP (28%) | 86,425,898 | 1.7% | 14.6% | 36.3 % | 32.9 % | 12.2 % | 2.3 % |
| | AM (13%) | 40,619,357 | 1.7% | 14.6% | 36.6 % | 32.9 % | 12.0 % | 2.3 % |
| | MI (7%) | 22,154,632 | 1.9% | 16.0% | 38.2 % | 31.7 % | 10.5 % | 1.7 % |
| AAM 1,581,426,316 | OP (4%) | 63,621,386 | 1.8% | 14.0% | 35.7 % | 32.8 % | 13.0 % | 2.8 % |

| | | | | | | | | |
|--------------------------|-------------------|-------------------|------|-------|-------|-------|-------|------|
| | AM (4%) | 58,111,030 | 1.9% | 14.3% | 35.9% | 32.5% | 12.7% | 2.7% |
| | MI (1.4%) | 22612814 | 1.7% | 14.2% | 36.3% | 32.7% | 12.5% | 2.6% |
| CF 59,813,525 | OP (15.4%) | 9,205,517 | 1.7% | 13.5% | 35.2% | 34.0% | 13.1% | 2.6% |
| | AM (4.1%) | 2,472,585 | 1.7% | 14.1% | 36.5% | 33.6% | 11.9% | 2.2% |
| | MI (2.0%) | 1,201,511 | 1.8% | 14.9% | 37.5% | 33.0% | 10.9% | 1.8% |
| CM 622,042,698 | OP (1.6%) | 9,900,557 | 1.9% | 14.5% | 36.9% | 32.6% | 11.8% | 2.2% |
| | AM (0.96%) | 6002026 | 1.9% | 14.4% | 36.9% | 32.9% | 11.7% | 2.1% |
| | MI (0.23%) | 1,444,359 | 1.7% | 14.4% | 38.0% | 33.1% | 11.0% | 1.8% |

As displayed in the table above, for all three classifiers and all cohorts at one and two standard deviations, the pairs involved in the gender misclassified images have a higher percentage below the mean than above the mean. At three standard deviations, we have approximately equal shares at the negative and positive side with more at the positive three std slightly higher than the negative for all. One exception for the Microsoft classifier with the African American Female, which has a minor occurrence at +3 std than at -3std, has the lowest percentage among all the cohort. This observation points out that the female black demographic can outperform other demographics, and the result from one algorithm cannot predict the outcome for another one.

Table 15 displays the distribution when the two images in the comparison pair are involved in the classification errors. Results show a shift to the positive side of the mean, which implies that only when both images participate in the classification errors generate a higher similarity score and can impact the false match rate.

Table 16. Two impostor images involved in GC errors split by std range

| Percentage per std bins where the two images in the pair involved in gend Open-source (OP), Amazon (AM), Microsoft (MI) | | | | | | | | |
|--|--|-----------|-------|-------|-------|-------|-------|-------|
| Non-mated | Total pair Comparisons involving in 2 images in misclassification errors | | -3std | -2std | -1std | 1std | 2std | 3std |
| AAF 308,840,189 | OP (3%) | 8,728,809 | 1% | 11% | 34% | 36% | 15% | 3% |
| | AM (0.5%) | 1,541,883 | 1% | 10% | 33% | 37% | 16% | 4% |
| | MI (0.1%) | 425,108 | 1% | 10% | 33% | 37% | 16% | 3% |
| AAM 1,581,426,316 | OP (0.04%) | 665,101 | 0.5% | 6.5% | 25.7% | 36.8% | 22.7% | 7.8% |
| | AM (0.03%) | 552,158 | 0.4% | 5.1% | 22.9% | 36.7% | 25.3% | 9.7% |
| | MI (0.01%) | 80,929 | 0.3% | 4.8% | 22.1% | 35.3% | 26.2% | 11.2% |
| CF 59,813,525 | OP (0.7%) | 420,354 | 0.9% | 9.6% | 31.4% | 36.9% | 17.2% | 4.0% |
| | AM (0.04%) | 26,248 | 1.0% | 8.9% | 30.6% | 38.1% | 17.4% | 4.1% |
| | MI (0.01%) | 6,019 | 1.2% | 10.7% | 32.5% | 37.1% | 15.1% | 3.3% |
| CM 622,042,698 | OP (0.006%) | 39,799 | 1.2% | 10.7% | 32.5% | 35.5% | 16.0% | 4.1% |
| | AM (0.002%) | 14,494 | 0.9% | 9.6% | 32.3% | 35.4% | 17.0% | 4.7% |
| | MI (0.0001%) | 803 | 0.7% | 8.8% | 32.5% | 34.9% | 18.1% | 5.0% |

From those two analyses, the main observations are: 1) the impostor distribution includes the gender classification error across its entire range and follow the initial representation pattern presented initially, especially when one image is involved in the error, 2) The representation of the gender classification at positive two and three

standard deviation range is minimal especially when both images from the pairs involved, hence will have less impact on the false match once a threshold is set.

Further, we focus on each cohort, especially the African American cohorts, and provide a more in-depth analysis regarding the effect of gender classification on the face matching across the entire impostor and the false match area and skin tone analysis.

African American Female (AAF): GC errors with the non-mated distribution

This section looked at the AAF non-mated similarity score and the participation of one image of the comparison pairs in the gender classification errors. Figure 13 below shows the AAF impostor curve distribution with an average of 0.066 slightly skewed to the right and the standard deviation of 0.096. The variation range at one std $[-0.03, 0.163]$ around the means shows 68% of the data. The two standard deviations above and below the means $[-0.127, 0.26]$ delimited by the dotted orange line include 95% of the observations. The second graph shows the median and split the data at 25% and 50%, which can have a balanced number of images at each side for the analysis. However, to maintain the data integrity, we used the first distribution with the average and standard deviation.

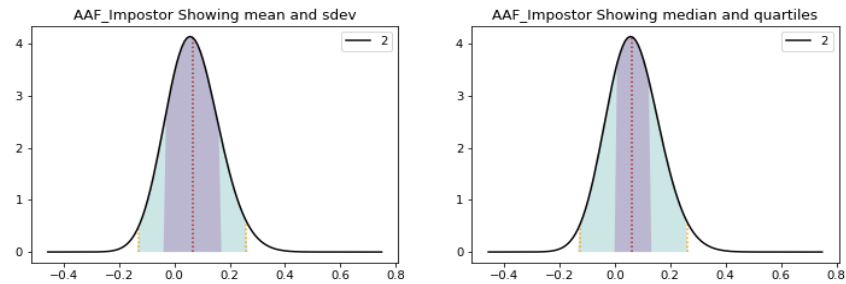


Figure 13. AAF Impostor distribution on the left the mean and std range, on the right median and quartiles

As mentioned above, for each std bin, we assessed the initial number of non-mated distributions and pairs that do not participate in the classification errors. Then, the comparison pairs involved in gender classification error with one and two images. We present the plot in Figure 14 for better visualization.

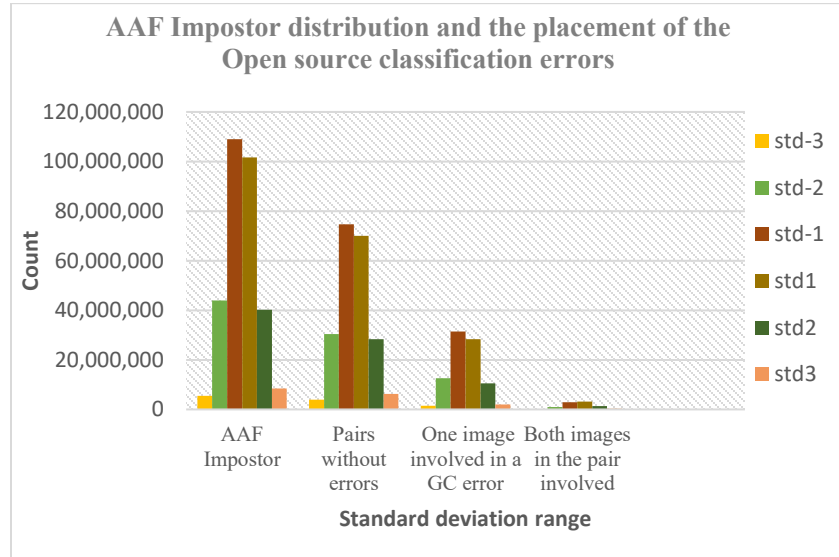


Figure 14.AAF impostor distribution with GC errors placement

Figure 14 above shows the four parts, including the initial AAF non-mated data within the bins. For the first three, we can observe a similar histogram with more concentration at std-2, std-1(the negative side) than std2, and std1(the positive side), in contrast with the last bins where we have more concentration at the std3 (positive side) than std-3 (negative side)

On the contrary, in the last plot with both pairs involved, the positive side with the three bins takes over the opposing side, which hypothetically will influence the matching errors. However, the following histogram, Figure 15, presents the contrast between the four categories involved with the open-source classifier and shows how minimal the last category compares to the first three.

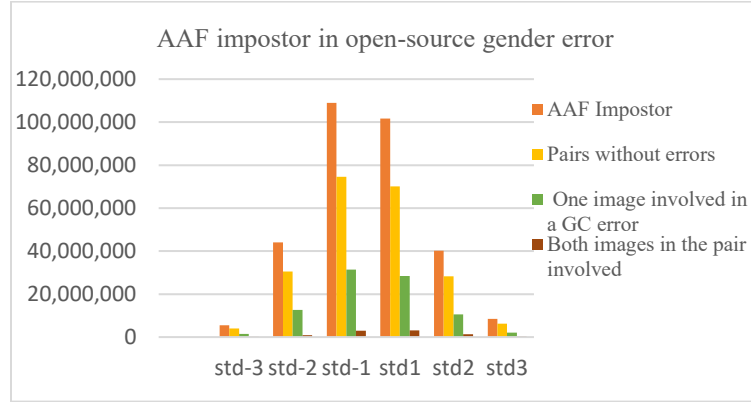


Figure 15. AAF open-source comparative plot between the impostors and the pairs involved in GC errors

Looking at the plot from the commercial algorithms (Figure 16) where we have better accuracy than the open-source one, we can notice that the bar where two images are barely seen at +2 and +3 std deviation.

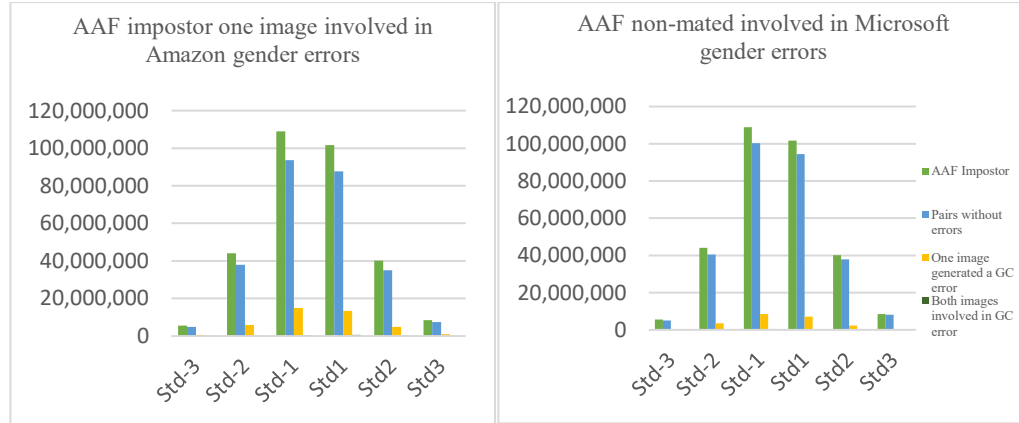


Figure 16. AAF Amazon(left), Microsoft(right) comparative plot between the impostors and the pairs involved in classification errors

The African American Female dataset result shows that at positive two and three standard deviations, we have fewer images participating in the classification error

and present in the face matching similarity score that could yield to the false matching error.

Following, we look at the false matching error at the 1-in-10,000 Caucasian male threshold to access the classification error variation on that specific zone.

AAF false match errors analysis involving in the GC errors

We have already demonstrated that the impact of the gender classification error is insignificant; we emphasize it by analyzing the variation of the false match error when involved with the classification error. As shown in Table 17 below, the initial false match rate is set as a baseline. We noticed a decrease of the False match for the three classifiers algorithm when one image participates in the gender errors. This observation confirms once more that the false match errors cannot be inferred from the gender errors.

Table 17. AAF False match errors with one image involved in the classifier's errors

| | | Num of comparisons | Num of errors | FMR | FMR (%) | Error |
|------------------------------------|--------------------|---------------------------|----------------------|------------------|----------------|-----------------|
| AAF (all image comparisons) | | 308,840,189 | 945908 | 0.0030628 | 0.30 | baseline |
| AAF w/ 1 GC Err | Open source | 86,425,888 | 203993 | 0.0023603 | 0.24 | Dec |
| | Amazon | 40,619,357 | 96265 | 0.0023699 | 0.24 | Dec |
| | Microsoft | 22,154,632 | 32840 | 0.0014823 | 0.15 | Dec |

In the second case, when both images from the mismatching comparisons errors appeared in the classification errors, we stated that some positive increases were noticed across the distribution but remain very small. The result from the false match variation, in this case, shows a decrease refer to Table 18 for the open-source

algorithm and an increase for the commercial one. The inconsistency proves that we cannot derive a definitive conclusion in this case.

Table 18. AAF False match errors with two images involved in the classifier's errors

| | | Num of comparisons | Num of errors | FMR | FMR (%) | Error |
|------------------------------------|--------------------|---------------------------|----------------------|------------------|----------------|-----------------|
| AAF (all image comparisons) | | 308,840,189 | 945908 | 0.0030628 | 0.30 | baseline |
| AAF w/2GC Err | Open source | 8,728,809 | 20916 | 0.0023962 | 0.24 | Dec |
| | Amazon | 1,541,883 | 6615 | 0.0042902 | 0.43 | Inc |
| | Microsoft | 425,108 | 1439 | 0.003378 | 0.34 | Inc |

We repeated the same analysis with the rest of the cohorts to see if they show the same result before moving to skin tone analysis.

African American Male (AAM): GC errors with the non-mated distribution

Like the AAF non-mate data, the AAM distribution (Figure 17) has a skewed distribution with 51.5% below the mean and 48.5% above the mean. The impostor scores generated an average of 0.023 with a 0.089 standard deviation

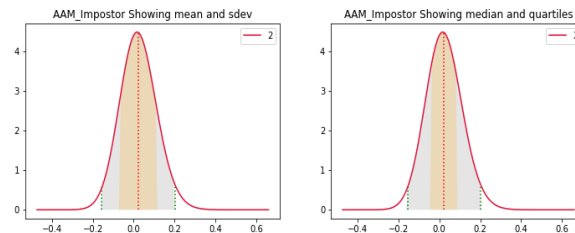


Figure 17. Impostor distribution showing on left the mean and std range, on the right, the median, and quartiles

From the initial tables (Table 15 and Table 16), we already know that the male cohort has better accuracy than the female cohort. For example, the percentage of the misclassification in the non-mated distribution for the open-source is 4% for AAM, which is very small compared to 28% for AAF (Table 15) when one image is involved and 0.04% versus 4% (Table 16) when two pictures involved, and this is even lower for the commercial algorithms. When splitting those low percentages of errors into bins beside the initial representation of the distribution and the pairs without mistakes, we see that it is barely noticeable (Figure 18) and cannot influence the false matching output.

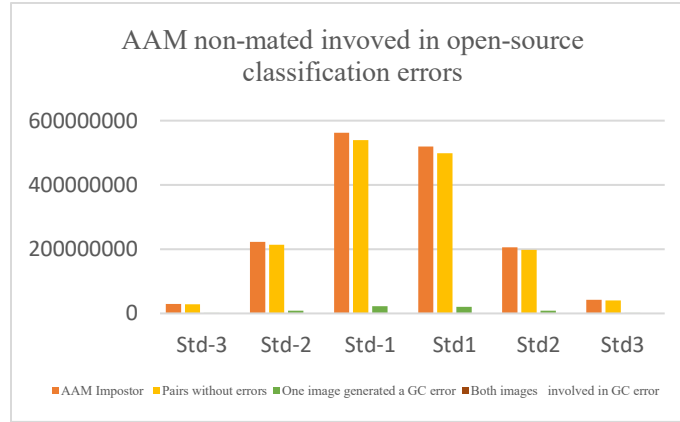


Figure 18. AAM open-source comparative plot between the impostor and the pairs involved in GC errors

AAM False matching error analysis involved in the GC errors

On the male side, the false match rate is 0.039% which is less than on the female side; however, the variation of the incorrect match when one image from the matching errors involved in GC errors shows an increase for the three classifiers. When both pictures in pairs are involved in the classification errors, we have the same pattern as the AAF, where the rate decrease for the open-source algorithm and increases for the commercial algorithm.

Although the male error rate is minimal, we have a disparity between the gender, and African American females are not always the ones that perform worse in every situation. Table 19 below shows the false match analysis for the AAM cohort.

Table 19. AAM FM errors with one and two images involved in GC errors

| | | Num of comparisons | Num of errors | FMR | FMR (%) | Error |
|------------------------------------|--------------------|---------------------------|----------------------|------------|----------------|-----------------|
| AAM (all image comparisons) | | 1,581,426,316 | 626080 | 0.00039 | 0.039 | baseline |
| AAM w/ 1 GC Err | Open source | 63,621,386 | 31580 | 0.00049 | 0.049 | Inc |
| | Amazon | 58,111,025 | 29377 | 0.00050 | 0.050 | Inc |
| | Microsoft | 22,612,814 | 10445 | 0.00046 | 0.046 | Inc |
| AAM w/2GC Err | Open source | 665,101 | 1691 | 0.00254 | 0.25 | Dec |
| | Amazon | 552,158 | 1818 | 0.00329 | 0.33 | Inc |
| | Microsoft | 80,929 | 439 | 0.00542 | 0.54 | Inc |

We now look at the Caucasian distribution to see if the previous observations extended to the Caucasian demographic.

Caucasian Female (CF) GC errors with the non-mated distribution

The Caucasian female cohort scores generate an average of 0.041 with a standard deviation(std) of 0.083. The std range point at one [-0.042,0.125] and two [-0.126, 0.209] above and below the mean are used to plot the distribution (Figure 19), and within each bin, we generate the count of images participated in gender classification error

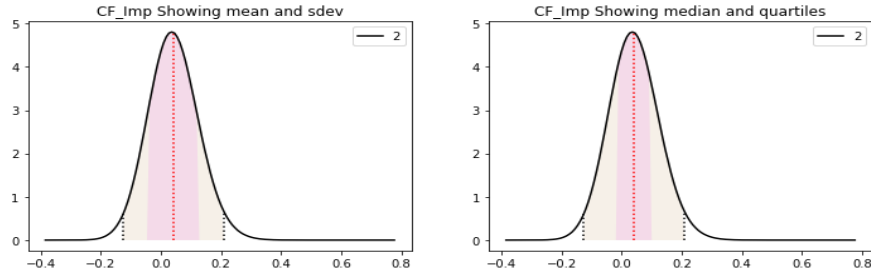


Figure 19. CF Impostor distribution showing on the left the mean and std, on the right median and quartiles

Referred to Table 15 and Table 16, the CF comparison pairs involved in the classification errors represent a tiny dataset. More errors are generated with the open-source algorithm where we have 15.7% for one image from the non-mated pairs participate in the GC errors and 0.7% when both images from the pairs are involved. Amazon generated 4% in the first case and 0.04 in the second case, and Microsoft has the best performance with barely 2% for the prior and 0.01% for the latter. Similar to the African American cohort, the errors follow the data pattern within the range shown in Figure 20 below. The positive two and three standard deviations have fewer errors than the negative side, and a slight flip is noticed when the two images involved are barely seen on the plot.

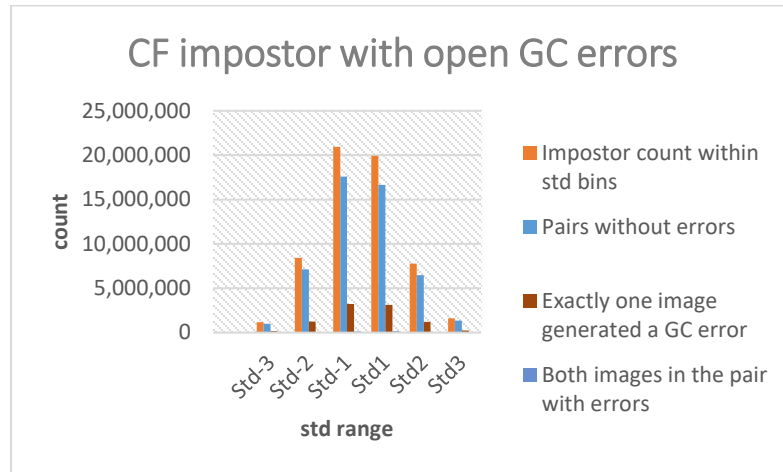


Figure 20. CF open-source comparative plot between the impostor and the pairs involved in classification errors

CF False matching error analysis involved in the GC errors

Narrowing down the result to the three standard deviation ranges above the mean with the false match errors at (1 in 10000) Caucasian male threshold, we have a 0.04% false match rate set as the baseline for the comparison rate involved classification errors. We can notice in

Table 20 below, same as with the AAF cohort, a slight continuous decrease from the Open source to Microsoft algorithm when one image is involved in the GC errors, and the second category, we have very little data. We considered only the errors generated from the open-source since it crosses the bar of 300-errors. There was a slight increase in the false match compared to the African American Cohort, where we decreased FMR for the open-source even though both images were involved.

Table 20. CF FM errors with one and two images involved in GC errors

| | Num of comparisons | Num of errors | FMR | FMR (%) | Error |
|--|--------------------|---------------|-----|---------|-------|
| | | | | | |

| | | | | | | |
|-----------------------------------|--------------------|-------------------|--------------|---------------|--------------|-----------------|
| CF (all image comparisons) | | 59,813,525 | 23928 | 0.0004 | 0.04 | Baseline |
| CF w/ 1 GC Err | Open source | 9205517 | 3598 | 0.00039 | 0.039 | Dec |
| | Amazon | 2472585 | 733 | 0.00029 | 0.029 | Dec |
| | Microsoft | 1201511 | 294 | 0.00024 | 0.024 | Dec |
| CF w/2GC Err | Open source | 420354 | 302 | 0.00071 | 0.07 | Inc |
| | Amazon | 26248 | 16 | 0.00061 | 0.06 | N/A |
| | Microsoft | 6019 | 3 | 0.00049 | 0.05 | N/A |

Caucasian Male (CM) GC errors with the non-mated distribution

We close the analysis with the Caucasian male cohort. The CM impostor distribution (Figure 21) has an average of 0.02 and std of 0.08. the std range at one and two-point [-0.138 -0.059 0.098 0.177]. The distribution curve and all the detailed plots within each range are summarized below.

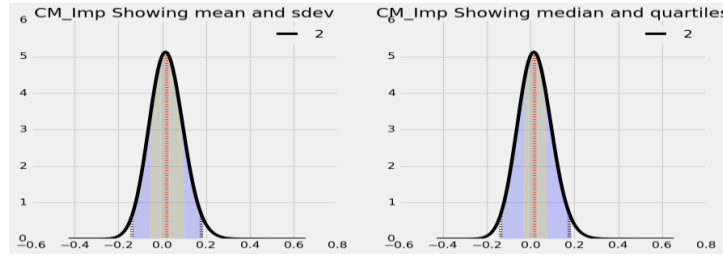


Figure 21. CM Impostor distribution showing on left the mean and std, on right median and quartiles.

Caucasian males have the best accuracy in terms of face matching and gender classification among the four cohorts. The participation of the comparison pairs in the classification errors is minimal. We have barely 2% with the open-source classifier and 0.23% for the Microsoft API when one image participates in the gender errors and 0.006% for the open-source and 0.0001% for Microsoft in case of

two images involved in the gender errors. Figure 22 shows the distribution where the bars engaged in the classification errors are almost invisible.

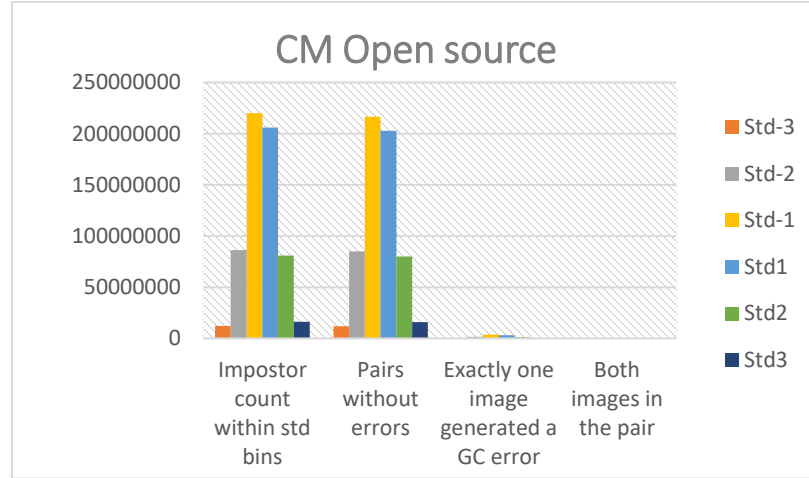


Figure 22. CM impostor distribution with GC errors placement

CM False matching error analysis involved in the GC errors

In terms of false match comparison shown in Table 21 below, for one image involving the gender errors, the open-source and amazon showed an increase similar to the African American male result, but a decrease for the Microsoft API. In the two images involving, the number of errors is too small to be evaluated.

Table 21. CM FM errors with one and two images involved in the GC errors

| | | Num of comparisons | Num of errors | FMR | FMR (%) | Error |
|----------------------------|-------------|--------------------|---------------|----------|---------|----------|
| CM (all image comparisons) | | 622042698 | 57097 | 0.000091 | 0.009 | baseline |
| CM w/ 1 GC Err | Open source | 9900557 | 1040 | 0.00011 | 0.01 | Inc |

| | | | | | | |
|-----------------------------|--------------------|---------|-----|----------|--------------|------------|
| CM w/2GC Err | Amazon | 6002026 | 650 | 0.00010 | 0.01 | Inc |
| | Microsoft | 1444359 | 107 | 0.000074 | 0.007 | N/A |
| | Open source | 39799 | 14 | 0.00035 | 0.035 | N/A |
| | Amazon | 14494 | 2 | 0.00013 | 0.013 | N/A |
| | Microsoft | 803 | 0 | 0 | 0 | N/A |

To summarize the results from this section, we have seen that the gender classification errors fall into every standard deviation bin across all the Morph cohorts. The placement of the GC errors across the range follows the representation of the initial non-mate distribution. On the one hand, when one image involves the GC errors, we have more concentration toward the lower similarity scores than the higher similarity score. However, when two images are affected, the number of errors increases toward the higher similarity score.

Regarding the false matching and the gender classification errors analysis, even though the Male cohorts have lower incorrect match error than the females, the false match rate increases when the comparison pairs involved with the GC errors contrast to the female side.

Overall, the impact of gender classification errors on false matching is minimal. The few inconsistencies noticed across the cohort and the classifiers prove that inferred the face matching output from the gender classification result would be inaccurate.

In the following section, we continued our analysis by assessing the skin tone of images involving gender classification and the face matching non-mate comparisons. This analysis will give a clear view of the skin tone's influence on the errors. We looked at the skin tone distribution across the standard deviation range and analyzed the false match area.

Chapter 7

Skin tone (SK) factor in errors analysis

The second part of our investigation is to assess the evidence of the skin tone impact on the correlated distribution of the gender classification and the face matching. We tend to answer the question: Are resulting similarity scores for comparisons involving images with darker skin-tone ratings more concentrated at +2 std and +3 std of the non-mated pair distribution?

As we did before with the gender classification only in Experiment results, we apply the same automated skin-tone rating algorithm produced by KKS et al. on those images of match scores that involve a gender misclassified image for each skin tone within each bin. We start with our first case, where only one image is involved in classification errors and when the comparison involves two images.

AAF SK assessment

AAF one image involved in GC errors

Table 22 below includes the skin type percentage of the relative frequency within each bin for the open-source classification error embedded in the face matching.

Table 22.AAF SK relative frequency for one image involved in open-source GC errors

| Skin | std-3 | std-2 | std-1 | std1 | std2 | std3 |
|-------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Sk-1 | 0.3% | 0.4% | 0.4% | 0.3% | 0.3% | 0.2% |
| Sk-2 | 8.0% | 8.0% | 7.8% | 7.2% | 6.4% | 5.6% |
| Sk-3 | 23.6% | 24.7% | 25.6% | 26.2% | 26.3% | 26.1% |
| Sk-4 | 32.9% | 34.3% | 35.5% | 36.7% | 37.7% | 38.7% |
| Sk-5 | 30.1% | 28.0% | 26.3% | 25.2% | 24.9% | 25.0% |
| Sk-6 | 5.1% | 4.7% | 4.4% | 4.3% | 4.4% | 4.4% |

We plot the information to represent better each skin type's variation across the bins and a second plot to compare the same skin types between bins.

From Figure 23 below, skin types per standard deviation range plot, we can observe that sk-3, sk-4, and sk-5 are the most represented across the distribution with a dominant skin type 4. We noticed that at the negative standard deviation range, skin types 4 and 5 dominate, while on the positive side, skin types 4 and 3 generate the higher impostor score. From our hypothesis, if we have a higher concentration of the darker skin tone at the two and three positive standard deviations, meaning the skin tone 4, 5, and 6, we can consider the skin effect. Surprisingly, toward the higher similarity score, the lighter skin type 3 takes over the skin type 5. Across the entire distribution, additionally, we have more images classified with skin type 2 than skin type 6.

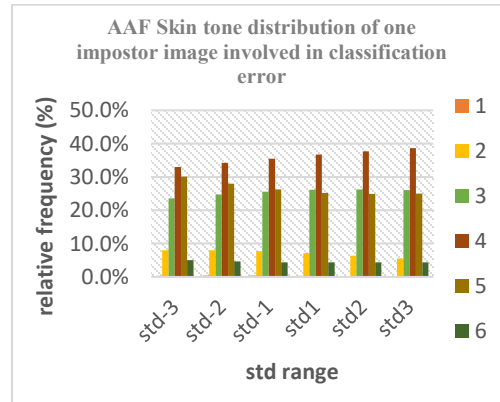


Figure 23. AAF SK distribution plot per std range

For the commercial algorithms, we observe another pattern (Figure 24). Across the entire range, the dominant skin types are 4 and 5, and the variation is noticed between those two where skin type 5 dominates on the negative side, and skin type 4 dominates on the positive side. Skin type 3 stays behind all over the range, and in this case, skin type 6 have a higher percentage than skin type 2

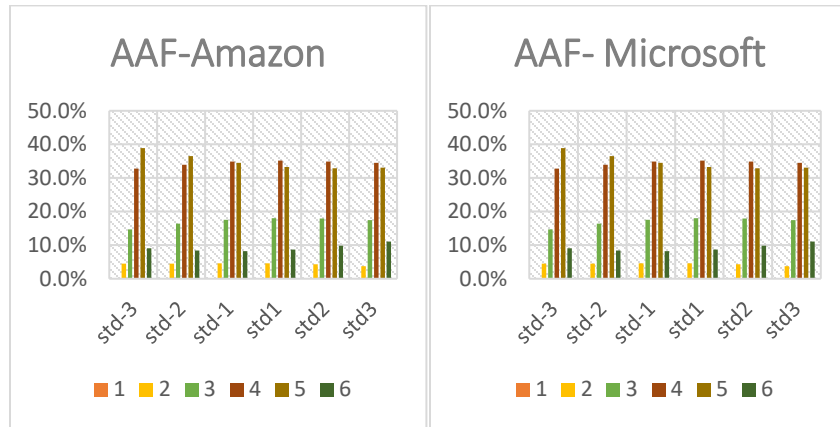


Figure 24. AAF Skin tone Amazon(left), Microsoft(right) SK distribution plot per std range

The second plot, Figure 25 below, shows the comparison of the same skin type across the different bins. The plot shows that only skin type 4 presents a continuous increase from the lower std range to the higher range followed by skin type 3, while skin types 5 and 6 show a decrease toward the higher similarity score.

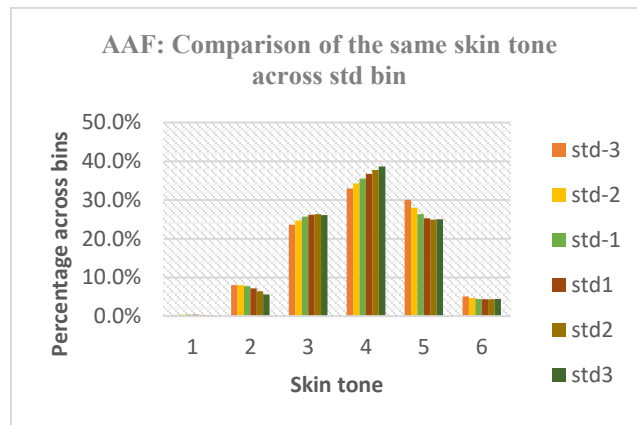


Figure 25. AAF Skin tone open-source comparative std plot per SK.

For the commercial, the low similarity score has more concentration for skin type 5, while some increase with the skin types 4 and 6 within the higher similarity score (Figure 26).

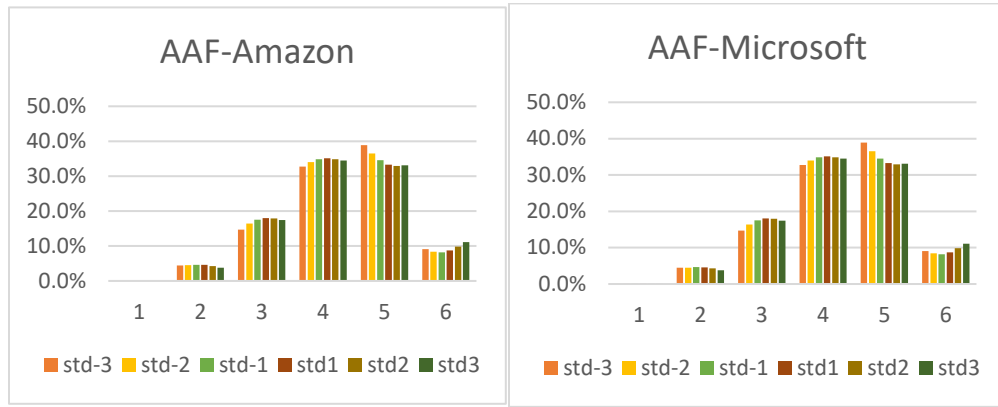


Figure 26. AAF Skin tone Amazon(left) Microsoft(right) comparative std plot per SK.

AAF SK assessment: two images involved in GC errors

We mentioned a flip to the higher similarity score for the two images participating in the classification errors, which remains negligible. We looked at four categories for the skin tone assessment: both images have the same skin tone, one skin tone difference, two skin tone differences, and greater or equal to three skin tone differences.

From the result present in Figure 28, we see that the one difference category dominates for the three classifiers, making it difficult to assess to evaluate the skin tone impact. However, the pairs of dominant skin tones are around the skin types, 4, 3, and 5, as in on the first case.

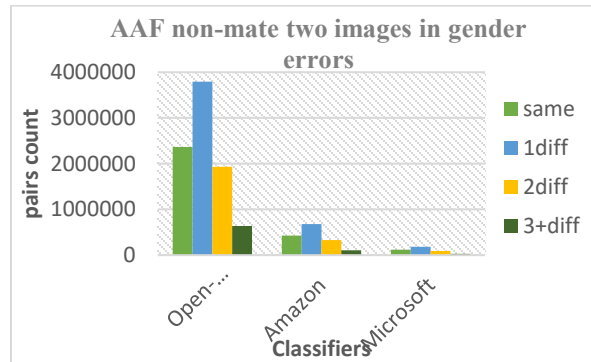


Figure 27. AAF two images involved in GC errors per classifiers.

Further, we analyzed the total number of skin types for the individual image presented in each pair within each bin and the variation of the false matching rate alongside it.

AAF false match error with one image in GC error

Previously, we looked at the skin tone distribution across the standard deviation range and find some variation with skin tone between skin tone 3 and 5 for the open-source and skin tone 5 and 4 for the commercial classifiers above and below the mean. To analyze the false match variation within the different skin tones, we sum up each skin tone present in the std range. When one image is involved in the classification error, the raw count displayed in Table 23 below showed that the Amazon API result aligns with the open-source with the dominant skin 4, 5, 3, 2, and 6. In contrast, skin type 5 comes first for Microsoft, followed by 4, 3, 6, and 2, respectively.

Table 23. AAF Total impostor involving one image in GC errors per SK.

| AAF: W/1GC: Skin tone | Sk-1 | Sk-2 | Sk-3 | Sk-4 | Sk-5 | Sk-6 | Total |
|-----------------------------|--------|---------|----------|----------|----------|---------|-------------------|
| Open source | 310090 | 6387723 | 22242690 | 31130541 | 22531222 | 3823622 | 86,425,888 |
| Amazon | 69277 | 2424812 | 9075913 | 15125754 | 11730429 | 2193172 | 40,619,357 |
| Microsoft | 47855 | 1004890 | 3876104 | 7704130 | 7608165 | 1913488 | 22,154,632 |

Next, we access the number of each skin tone that exceeds the (1 in 10000) Caucasian male threshold. We display the row count in Table 24, where we see the dominant with skin tone 4 for all three classifiers and the variation for the second place 3, 5, and 2 for the open-source, and for the commercial, we have 5, 3, 6, respectively.

Table 24. AAF FMR involving one image in GC errors per SK.

| AAF T: 0.35 skin tone | Sk-1 | Sk-2 | Sk-3 | Sk-4 | Sk-5 | Sk-6 | Total |
|-----------------------------|------|-------|-------|-------|-------|------|---------------|
| Open source | 387 | 10100 | 52993 | 80597 | 51089 | 8827 | 203993 |
| Amazon | 72 | 4074 | 21034 | 38997 | 26987 | 5101 | 96265 |
| Microsoft | 45 | 1183 | 5752 | 11021 | 10927 | 3912 | 32840 |

Finally, we analyzed the progression of the false match between the total number of errors across the entire distribution and the one that crossed the threshold. Overall, referring to Table 25, none of the false match rates per skin type exceed the baseline. Within skin types, the dominant error rates are around skin tones 3, 4, 5, and 6. For the open-source and Amazon classifiers, we see an increase of the false match from skin tone 3 to 4, which has the highest peak and decreases at skin types 5 and 6, both with the same percentage. In contrast, Microsoft shows the highest rate for skin type 6, followed by skin type 3, then skin types 4 and 5, which have the same false match rate.

Table 25. AAF FMR involving one image in GC errors per SK.

| AAF skin tone 1GC/W1GC | Sk-1 | Sk-2 | Sk-3 | Sk-4 | Sk-5 | Sk-6 | FMR base |
|------------------------------|-------|-------|--------------|--------------|--------------|--------------|-------------|
| Open source | 0.12% | 0.16% | 0.24% | 0.26% | 0.23% | 0.23% | 0.30 |
| Amazon | 0.10% | 0.17% | 0.23% | 0.26% | 0.23% | 0.23% | |
| Microsoft | 0.09% | 0.12% | 0.15% | 0.14% | 0.14% | 0.20% | |

AAF false match error with two images in GC error

In the second case, when both images are involved in the classification errors, the pair of images have either the same skin tone or different skin tones. The result across the entire distribution showed on the last plot that the pairs with varying tones of skin outnumbered the comparison with the same skin tone.

The false match variation showed the same pattern where one and two skin tone differences have higher false matches than the same skin tone rate, as displayed in Table 26 below.

Table 26. AAF FMR involving two images in GC errors per SK.

| AAF W/ 2 GC error: Skin tone | same | 1 diff | 2diff | 3+diff | FMR base |
|------------------------------|-------|--------------|--------------|--------|----------|
| Open-source | 0.42% | 0.52% | 0.44% | 0.42% | 0.30 |
| Amazon | 0.33% | 0.47% | 0.35% | 0.33% | |
| Microsoft | 0.24% | 0.32% | 0.27% | 0.24% | |

We looked at the FMR of the same skin tone analysis; we consider the pairs of images that accumulated at least 300 errors which concern the pairs 3_3, 4_4, and 5_5 for the open-source classifier.

Compared to the false match baseline rate, we have a slight increase in the pairs with the skin type 5_5 shown in Table 27.

Table 27. AAF FMR with same SK for the pairs involved in OP GC errors.

| OP: same skin tone | AAF W/ 2 GC error | T: 0.35 AAF W/ 2 GC error | FMR | FMR(%) | Error |
|--------------------|-------------------|---------------------------|-----------------|--------------|-----------------|
| Total | 2365009 | 9972 | 0.004216 | 0.42% | Baseline |
| 6_6 | 16469 | 230 | 0.013966 | 1.40% | N/a |
| 5_5 | 591852 | 3029 | 0.005118 | 0.51% | Inc |
| 4_4 | 1131694 | 4460 | 0.003941 | 0.39% | Dec |
| 3_3 | 577486 | 2029 | 0.003514 | 0.35% | Dec |
| 2_2 | 47404 | 224 | 0.004725 | 0.47% | N/A |

AAM skin tone assessment

AAM one image involved in classification errors

The skin tone distribution of the AAM non-mated involved in the classification in one image error shows more consistency across the std range than the female case. The dominant skin types for the three classifiers (Figure 28 and Figure 29) are 4, 5, and 3, respectively, while with AAF, we noticed variation within the bins and across the classifiers. Skin type 4 shows a constant increase toward the higher range, while skin type 5 decreases as it goes toward the higher std.

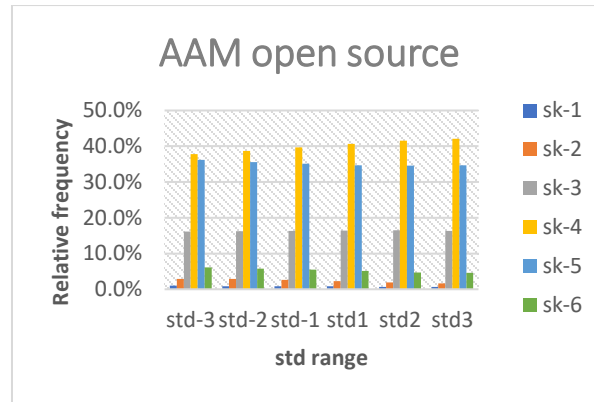


Figure 28. AAM open-source SK distribution plot per std range

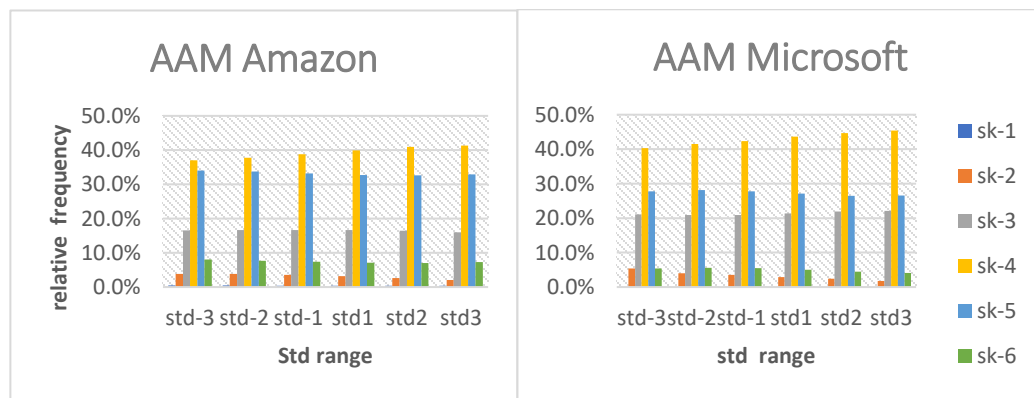


Figure 29. AAM Amazon(left), Microsoft(right) SK distribution plot per std range

The transpose plots (Figure 30, Figure 31) show the variation of the same skin tone within each std range, where we can see that skin type 4 percentage is higher at std2 and sdt3 and skin type 5 is higher at the opposing side.

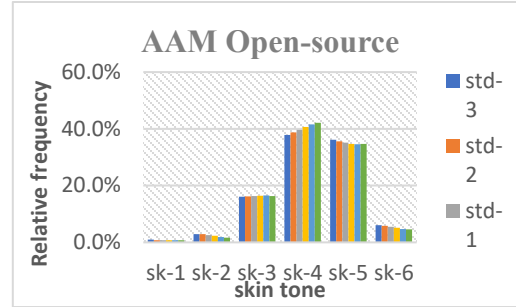


Figure 30. AAM Open-source comparative std plot per skin tone

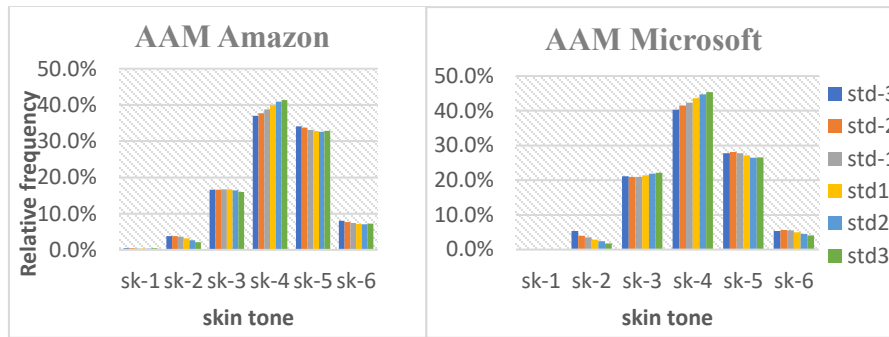


Figure 31. AAM Amazon(left) Microsoft(right) comparative std plot per skin tone

AAM two images involved in classification errors

When both images are involved in the classification errors (Figure 32), the skin tone distribution is like the AAF pairs involved in the classification error where the one skin tone difference outnumbered the rest of the categories.

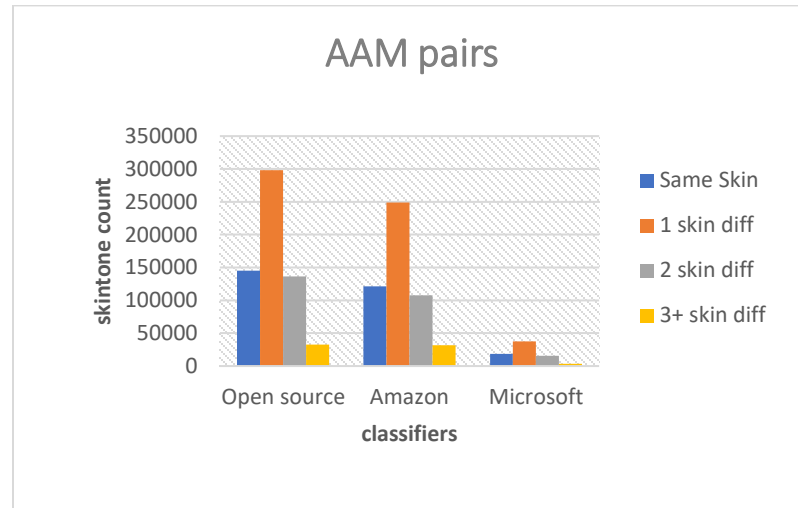


Figure 32. AAM two images involved in GC errors per classifiers

AAM false match error with one image in GC error

As was previously done with AAF, we analyzed the false match error based on the skin tone distribution. The total raw number of the skin types showed in Table 28 below follows the same pattern around 4, 5, and 3, as shown on the plots.

Table 28. AAM Total impostor involving one image in GC errors per SK.

| AAM | Sk-1 | Sk-2 | Sk-3 | Sk-4 | Sk-5 | Sk-6 | Total |
|---------------|------------|-------------|----------------|----------------|----------------|-------------|------------------------|
| Open source | 49573 2 | 154232 7 | 10,410,85 8 | 25,558,70 2 | 22,253,71 1 | 336005 6 | 63,621,38 6 |
| Amazon | 22073 1 | 193148 9 | 9,657,648 | 22,847,10 1 | 19,204,74 9 | 424930 8 | 58,111,02 6 |
| Microso ft | 0 | 725804 | 4,801,856 | 9,715,131 | 6,197,547 | 117247 6 | 22,612,81 4 |

The number of images that crossed the threshold presents the same progression across the skin tone, focusing on skin types 4, 5, and 3 (Table 29).

Table 29. AAM FM involving one image in GC errors per SK.

| AAM T : 0.35 skin tone | Sk-1 | Sk-2 | Sk-3 | Sk-4 | Sk-5 | Sk-6 | Total |
|---------------------------|------|------|-------------|--------------|--------------|------|--------------|
| Open source | 217 | 400 | 5200 | 12801 | 11419 | 1543 | 31580 |
| Amazon | 117 | 421 | 4245 | 11942 | 10117 | 2535 | 29377 |
| Microsoft | 0 | 162 | 2269 | 4635 | 2920 | 459 | 10445 |

The false match rate present in Table 30 shows an overall increased rate compared to the baseline for the dominant skin tone. However, only the Amazon API shows a constant increase from sk-4 to sk-6, which remains small compared to the initial value.

Table 30. AAM variation of FMR involving one image in GC errors per SK.

| AAM FM | Sk-1 | Sk-2 | Sk-3 | Sk-4 | Sk-5 | Sk-6 | Baseline |
|-------------|--------|--------|---------------|---------------|---------------|---------------|---------------|
| Open source | 0.044% | 0.026% | 0.050% | 0.050% | 0.051% | 0.046% | 0.039% |
| Amazon | 0.053% | 0.022% | 0.044% | 0.052% | 0.053% | 0.060% | |
| Microsoft | 0.000% | 0.022% | 0.047% | 0.048% | 0.047% | 0.039% | |

AAM false match error with two images in GC errors

When two images from the comparison pairs are involved in the classification errors, the AAM shows more images with one skin tone difference when we look at the raw numbers (Table 31). Amazon has recorded more pairs participating in the classification errors that crossed the threshold out of the three classifiers.

Table 31. AAM total Imp (left), False match (right) involving two images in GC errors per SK.

| AAM: W/2c Skin tone | Open- source | Amazon | Microsoft |
|------------------------------|-----------------|----------------|---------------|
| same | 208161 | 163713 | 24650 |
| 1diff | 304772 | 248869 | 37334 |
| 2diff | 119271 | 107881 | 15424 |
| 3+diff | 32897 | 31694 | 3521 |
| Total | 665,101 | 552,158 | 80,929 |

| AAM:w/ 2c T: 035 | Open- source | Amazon | Microsoft |
|------------------------|-----------------|-------------|------------|
| same | 653 | 687 | 176 |
| 1diff | 811 | 876 | 223 |
| 2diff | 200 | 213 | 31 |
| 3+diff | 27 | 41 | 9 |
| Total | 1691 | 1818 | 439 |

Contrary to the AAF false match, the AAM false match ratio between the total image involved and those that crossed the threshold scored a higher rate for pairs with the same skin tone than those with different skin tones (Table 32).

Table 32. AAM FMR involving two images in GC errors per SK.

| AAM W/ 2 GC error: Skin tone | same | 1diff | 2diff | 3+diff | Baseline |
|---------------------------------------|--------------|-------|-------|--------|--------------|
| Open-source | 0.31% | 0.27% | 0.17% | 0.08% | 0.039 |
| Amazon | 0.42% | 0.35% | 0.20% | 0.13% | |
| Microsoft | 0.71% | 0.60% | 0.20% | 0.26% | |

We looked at the detailed false match of the images with the same skin tone pairs present in Table 33; we had previously decided to consider the number of errors more significant than 300. The split of the same skin tone pairs involved in the classification errors have only the pairs with skin tone 4-4 that crossed the bar of 300 errors. The false match generated from that pair is slightly over the base threshold.

Table 33. AAM FMR with same skin tone for the pairs involved in open-source GC errors.

| OP: same skin tone | AAM W/ 2 GC error | T: 0.35 AAM W/ 2 GC error | FMR | FMR (%) | Error |
|---------------------------|--------------------------|----------------------------------|-----------------|----------------|-----------------|
| Total | 208161 | 653 | 0.002542 | 0.25 | Baseline |
| 6 6 | 1819 | 2 | 0.00109 | 0.109 | N/A |
| 5 5 | 81359 | 214 | 0.00263 | 0.26 | N/A |
| 4 4 | 107124 | 357 | 0.00333 | 0.33 | incr |
| 3 3 | 17668 | 80 | 0.00452 | 0.45 | N/A |
| 2 2 | 356 | 0 | 0 | 0 | N/A |
| 1 1 | 36 | 0 | 0 | 0 | N /A |

To summarize the findings regarding the skin tone effect for the African American Cohort, the sk4, which is dominant for the male and female, shows an increase within the higher similarity score for the open-source algorithm, when for the commercial, the dominant skin tone sk5 shows a decrease toward the std2 and std3.

Regarding the variation of the false match error, the commercial algorithms generate a higher mismatch for darker skin tone sk6 for both genders, Microsoft for the female cohort, and Amazon for the male mate when one image is involved in the GC errors.

For the two images participating in the false match, the female cohort has more errors, with the one skin tone difference pairs, where the male has more mismatched with the same skin tone pairs. For the same skin tone pairs, the female's cohort showed an increase for the skin tone 5-5, where the males showed an increase for the skin tone 4-4.

CF skin tone assessment

CF one image involved in GC errors

The Caucasian distribution has a significant skin tone from sk1 to sk4 displayed in Table 34. Like the African American cohort, the sk1, the most represented, has more errors towards the higher similarity score where the others decrease.

Table 34. CF SK relative frequency for one image involved in open-source GC errors

| CF OP | std-3 | std-2 | std-1 | std1 | std2 | std3 |
|-------|-------|-------|-------|------|------|------|
| sk-1 | 51% | 51% | 52% | 53% | 54% | 55% |
| sk-2 | 36% | 36% | 36% | 36% | 34% | 33% |
| sk-3 | 8% | 8% | 8% | 8% | 8% | 9% |
| sk-4 | 4% | 4% | 3% | 3% | 3% | 3% |
| sk-5 | 1% | 1% | 1% | 1% | 1% | 1% |
| sk-6 | 0% | 0% | 0% | 0% | 0% | 0% |

The plots from the Caucasian female show the variation of the skin tone within the std range for the open-source and the commercial algorithms. The same pattern observed across the three classifiers (Figure 33, Figure 34).

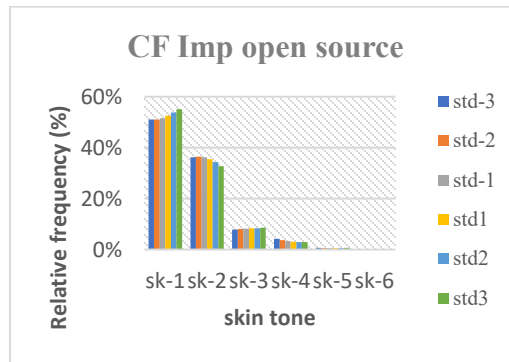


Figure 33. CF Skin tone Open-source comparative std plot.

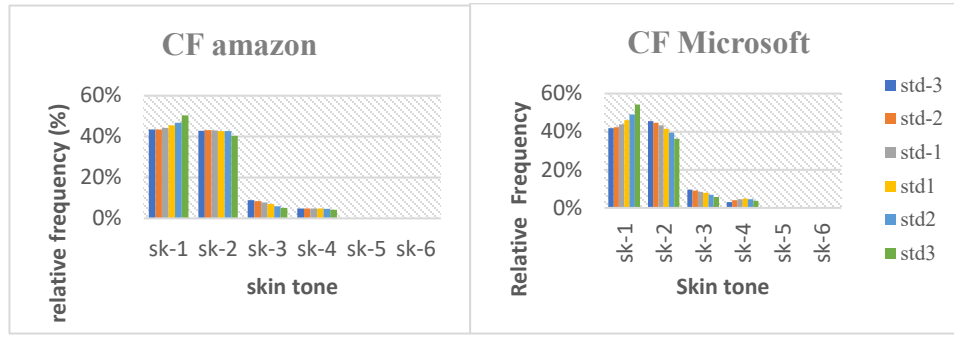


Figure 34. CF Amazon(left) Microsoft(right) comparative std plot per SK.

CF one image involved in GC errors

Looking at the total images for each skin tone, we have a continuous decrease from sk-1 to sk-6 initially (Table 35) and the total errors that crossed the threshold (Table 36).

Table 35. CF Total impostor involving one image in GC errors per SK.

| CF: W/1GC: Skin tone | Sk-1 | Sk-2 | Sk-3 | Sk-4 | Sk-5 | Sk-6 | Total |
|-------------------------|---------|---------|--------|--------|-------|------|----------------|
| Open source | 4808336 | 3285372 | 761257 | 300446 | 50106 | 0 | 9205417 |
| Amazon | 1113281 | 1059588 | 181972 | 117744 | 0 | 0 | 2472585 |
| Microsoft | 541257 | 508732 | 97406 | 54116 | 0 | 0 | 1201511 |

Table 36. CF FM involving one image in GC errors per SK.

| CF : W/1GC : Skin tone T : 0.35 | Sk-1 | Sk-2 | Sk-3 | Sk-4 | Sk-5 | Sk-6 | Total |
|------------------------------------|------|------|------|------|------|------|-------------|
| Open source | 2040 | 1093 | 352 | 92 | 21 | 0 | 3598 |
| Amazon | 464 | 213 | 35 | 21 | 0 | 0 | 733 |
| Microsoft | 196 | 79 | 14 | 5 | 0 | 0 | 294 |

For the false match analysis present in Table 37, the open-source classifier generated more errors. If we consider the 300 errors, the skin tone sk3 close to the darker skin tone has the highest false match rate when one image is involved in the GC errors.

Table 37. CF FMR involving one image in GC errors per SK.

| CF FM | Sk-1 | Sk-2 | Sk-3 | Sk-4 | Sk-5 | Sk-6 | Baseline |
|--------------------|---------------|---------------|---------------|--------|--------|--------|-------------|
| Open source | 0.042% | 0.033% | 0.046% | 0.031% | 0.042% | 0.000% | 0.04 |
| Amazon | 0.042% | 0.020% | 0.019% | 0.018% | 0.000% | 0.000% | |
| Microsoft | 0.036% | 0.016% | 0.014% | 0.009% | 0.000% | 0.000% | |

CF two images involved in GC errors

The CF involving two images in the GC errors generated more errors with one skin tone difference, similar to the AAF distribution. The number that crossed the threshold is very minimal, and we only consider the open-source classifier even though it does not cross the 300-errors (Table 38).

Table 38. CF total Imp (left), False match (right) involving two images in GC errors per SK.

| CF:w/ 2c | Open- source | Amazon | Microsoft | CF:w/ 2c T: 035 | Open- source | Amazon | Microsoft |
|--------------|-----------------|--------------|-------------|--------------------|-----------------|-----------|-----------|
| same | 171046 | 10189 | 2317 | same | 107 | 9 | 1 |
| 1diff | 184322 | 12083 | 2780 | 1diff | 153 | 7 | 2 |
| 2diff | 46567 | 2834 | 673 | 2diff | 31 | 0 | 0 |
| 3+diff | 18419 | 1142 | 249 | 3+diff | 11 | 0 | 0 |
| Total | 420354 | 26248 | 6019 | Total | 302 | 14 | 3 |

The false match rate shows an increase for the one difference skin tone for open source (Table 39), and the assessment of the exact incorrect match for the same skin tone images shows an increase for the skin tone sk2 (Table 40).

Table 39. CF of Open-source FMR involving two images in GC errors per SK.

| CF W/ 2 GC Skin tone | same | 1diff | 2diff | 3+diff | Baseline |
|----------------------------|--------|--------|--------|--------|-------------|
| Open- source | 0.063% | 0.083% | 0.067% | 0.060% | 0.04 |

Table 40. CF FMR with same SK for the pairs involved in open-source GC errors.

| OP: same skin tone | CF W/ 2 GC error | T: 0.35 CF W/ 2 GC error | FMR | FMR (%) | Error |
|--------------------|------------------|--------------------------|-----------------|--------------|-----------------|
| Total | 171046 | 107 | 0.000625 | 0.063 | Baseline |
| 1_1 | 114579 | 72 | 0.000628 | 0.063 | N/a |
| 2_2 | 53202 | 34 | 0.000657 | 0.066 | N/A |
| 3_3 | 2825 | 1 | 0.000353 | 0.035 | N/A |

CM skin tone assessment

CM one image involved in GC errors

The Caucasian male skin tone distribution showed more variation across the classifiers (Figure 35, Figure 36) than in the African American Male distribution. The open-source and the Amazon API have a similar pattern with sk-1 dominance. In contrast, the Microsoft algorithm has more error concentration with skin tone sk-2, showing a decrease in higher similarity scores. The commercial algorithms showed increased errors with the darker skin tone sk-3 to sk-5 toward the std+2 and std+3 range.

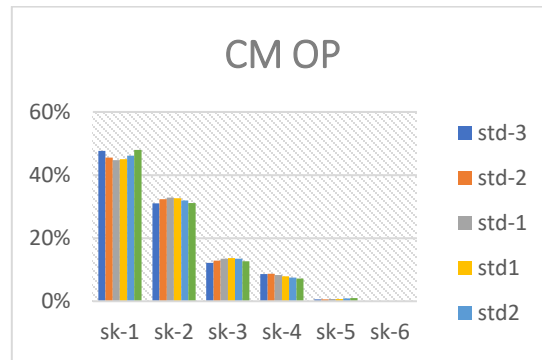


Figure 35. CM Open-source comparative std plot per SK.

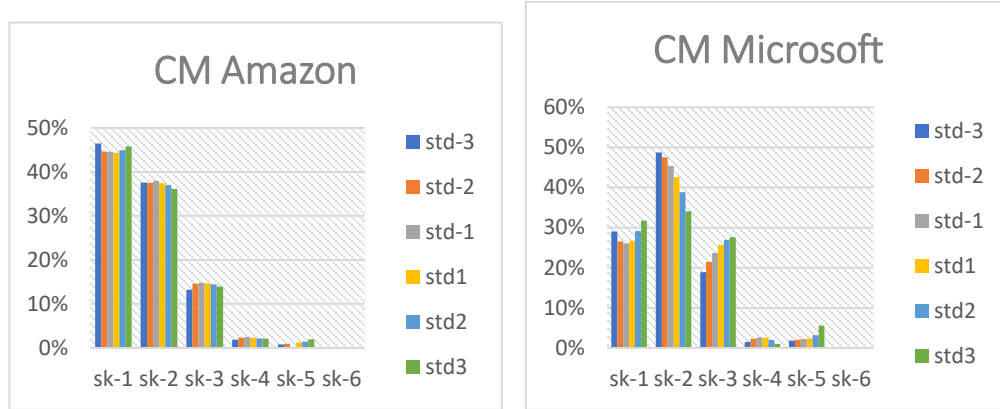


Figure 36. CM Amazon(left) Microsoft(right) comparative std plot per SK.

CM false match error with one image in GC error

We looked at the Caucasian Male total skin tone raw number involving one image in GC errors only since the second category where two images involved have nearly zero data that crossed the threshold for the false match analysis.

The result (Table 41) shows more errors initially from sk1 to sk-4 respectively for open-source and Amazon, and with Microsoft, we have sk-2, followed by sk1 to sk-5, respectively.

Table 41. CM Total impostor involving one image in GC errors per SK

| CM: W/1GC: Skin tone | Sk-1 | Sk-2 | Sk-3 | Sk-4 | Sk-5 | Sk-6 | Total |
|----------------------------|---------|---------|---------|--------|-------|------|----------------|
| Open source | 4478493 | 3218384 | 1329196 | 804696 | 69788 | 0 | 9900557 |
| Amazon | 2667734 | 2246339 | 876498 | 140403 | 48461 | 0 | 5979435 |
| Microsoft | 387558 | 634178 | 352214 | 35233 | 35176 | 0 | 1444359 |

For pairs that crossed the threshold, we only have the open-source errors above the 300 bars with the sk-1 and sk-2, respectively. If we look at the commercial

algorithm errors, Microsoft showed more errors at sk-3 with Amazon; we have more errors sk-5 than sk-4.

Table 42. CM FM involving one image in GC errors per SK.

| CM : W/1GC Skin tone T : 0.35 | Sk-1 | Sk-2 | Sk-3 | Sk-4 | Sk-5 | Sk-6 | Total |
|----------------------------------|------------|------------|------|------|------|------|-------------|
| Open source | 445 | 430 | 80 | 72 | 13 | 0 | 1040 |
| Amazon | 282 | 256 | 73 | 19 | 20 | 0 | 650 |
| Microsoft | 29 | 30 | 37 | 1 | 10 | 0 | 107 |

Referring to Table 43, the ratio shows a slightly higher or equal false match rate for sk-1 and sk-2 for the open-source. The others have too small errors to be examined.

Table 43. CM FMR involving one image in GC errors per SK

| CM FM | Sk-1 | Sk-2 | Sk-3 | Sk-4 | Sk-5 | Sk-6 | Baseline |
|--------------------|---------------|---------------|--------|--------|--------|--------|----------|
| Open source | 0.010% | 0.013% | 0.006% | 0.009% | 0.019% | 0.000% | 0.009% |
| Amazon | 0.011% | 0.011% | 0.008% | 0.014% | 0.041% | 0.000% | |
| Microsoft | 0.007% | 0.005% | 0.011% | 0.003% | 0.028% | 0.000% | |

Overall, the Caucasian cohort has better accuracy, which minimizes the error rate involved with the gender classification. The skin tone classification turns around the skin tone sk-1 and sk-2, representing the dataset's dominant representative.

Chapter 8

Conclusion

Our investigation of the relationship between face analytics, specifically gender classification and face recognition, is based on the media narrative that conflates gender classification analysis with the accuracy of automatic face recognition. We aimed to draw a fine line between the two systems. We provided a review of the different dataset inputs used and processing steps for the two systems. We have a two-domain decision output for one (male or female) and a multidomain decision for the second (one probe against n samples).

Also, with a precise experiment, we analyzed match scores involving gender misclassified images in the imposter distribution and observed these scores to be well distributed throughout. Secondly, the number of comparisons involving gender misclassified images represents only a tiny fraction ($\sim 2\%$) of scores present within the imposter distribution itself. Based on this experiment, we could not conclude that gender misclassified images significantly influence face matching accuracy

The second experiment used automated skin tone ratings to assess whether or not skin tone was a driver behind face processing results. Overall, the error generated tends to follow the data's initial skin tone representation. The African American cohort has their dominant skin tone around sk-4 with more errors toward the higher similarity score and recorded low errors with the darkest skin tone sk-5. Similarly, Caucasians have more skin tone sk-1 with a higher similarity score or sk-2 when we decrease the higher similarity score.

The false match rate for the face recognition system shows higher for the darker skin tone when one gender misclassified image is involved in the comparison. This finding is more pronounced for the Male cohort and the Microsoft API. When two

misclassified images are concerned, we have a higher false match for the darker skin tone with African American female cohort.

Based on our experiment with the morph dataset and the algorithm used, we can conclude that darker skin may be associated with higher errors for a small number of comparisons—but not for the system as a whole. Additionally, higher error rates for darker skin tones are not consistent across all demographics and all algorithms.

Since our investigation used isolated demographic cohorts, future work will focus on gender classification across the cohort and commercial face matching systems to provide further insight on the relationship between gender classification and face recognition based on skin tone.

References

- [1] “Industry 4.0: fourth industrial revolution guide to Industrie 4.0,” *i-SCOOP*. <https://www.i-scoop.eu/industry-4-0/> (accessed Feb. 11, 2021).
- [2] “What is the Difference Between IT and OT? | Coolfire Solutions Blog,” *Coolfire*. <https://www.coolfiresolutions.com/blog/difference-between-it-ot/> (accessed Feb. 23, 2021).
- [3] S. Lohr, “Facial Recognition Is Accurate, if You’re a White Guy,” *The New York Times*, Feb. 09, 2018. Accessed: Mar. 03, 2021. [Online]. Available: <https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html>
- [4] “How is Face Recognition Surveillance Technology Racist?,” *American Civil Liberties Union*. <https://www.aclu.org/news/privacy-technology/how-is-face-recognition-surveillance-technology-racist/> (accessed Mar. 03, 2021).
- [5] “Racial Discrimination in Face Recognition Technology,” *Science in the News*, Oct. 24, 2020. <https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/> (accessed Mar. 02, 2021).
- [6] P. J. Grother, G. W. Quinn, and P. J. Phillips, “Report on the Evaluation of 2D Still-Image Face Recognition Algorithms,” p. 61, 2010.
- [7] J. Buolamwini and T. Gebru, “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification,” in *Conference on Fairness, Accountability and Transparency*, Jan. 2018, pp. 77–91. Accessed: Feb. 02, 2021. [Online]. Available: <http://proceedings.mlr.press/v81/buolamwini18a.html>
- [8] A. K. Jain, A. Ross, and S. Prabhakar, “An introduction to biometric recognition,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 4–20, Jan. 2004, doi: 10.1109/TCSVT.2003.818349.
- [9] S. Z. Li and A. K. Jain, Eds., *Handbook of Face Recognition*. London: Springer London, 2011. doi: 10.1007/978-0-85729-932-1.

- [10] J. A. Unar, W. C. Seng, and A. Abbasi, “A review of biometric technology along with trends and prospects,” *Pattern Recognit.*, vol. 47, no. 8, pp. 2673–2688, Aug. 2014, doi: 10.1016/j.patcog.2014.01.016.
- [11] A. K. Jain and A. Kumar, “Biometric Recognition: An Overview,” in *Second Generation Biometrics: The Ethical, Legal and Social Context*, E. Mordini and D. Tzovaras, Eds. Dordrecht: Springer Netherlands, 2012, pp. 49–79. doi: 10.1007/978-94-007-3892-8_3.
- [12] A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino, “2D and 3D face recognition: A survey,” *Pattern Recognit. Lett.*, vol. 28, no. 14, pp. 1885–1906, Oct. 2007, doi: 10.1016/j.patrec.2006.12.018.
- [13] J. R. Lyle, P. E. Miller, S. J. Pundlik, and D. L. Woodard, “Soft biometric classification using periocular region features,” in *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, Sep. 2010, pp. 1–7. doi: 10.1109/BTAS.2010.5634537.
- [14] “Face Perception - an overview | ScienceDirect Topics.”
<https://www.sciencedirect.com/topics/neuroscience/face-perception> (accessed Feb. 27, 2021).
- [15] O. Pascalis *et al.*, “Development of Face Processing,” *Wiley Interdiscip. Rev. Cogn. Sci.*, vol. 2, no. 6, pp. 666–675, 2011, doi: 10.1002/wcs.146.
- [16] O. Pascalis and D. J. Kelly, “The Origins of Face Processing in Humans: Phylogeny and Ontogeny,” *Perspect. Psychol. Sci.*, vol. 4, no. 2, pp. 200–209, Mar. 2009, doi: 10.1111/j.1745-6924.2009.01119.x.
- [17] W. Freiwald, B. Duchaine, and G. Yovel, “Face Processing Systems: From Neurons to Real-World Social Perception,” *Annu. Rev. Neurosci.*, vol. 39, no. 1, pp. 325–346, 2016, doi: 10.1146/annurev-neuro-070815-013934.
- [18] “History of Face Recognition & Facial recognition software,” *FaceFirst Face Recognition Software*, Aug. 01, 2017.
<https://www.facefirst.com/blog/brief-history-of-face-recognition-software/> (accessed Feb. 16, 2021).

- [19] “Facial recognition history.”
<https://www.thalesgroup.com/en/markets/digital-identity-and-security/government/inspired/history-of-facial-recognition> (accessed Feb. 16, 2021).
- [20] K. Ricanek and C. Boehnen, “Facial Analytics: From Big Data to Law Enforcement,” *Computer*, vol. 45, no. 9, pp. 95–97, Sep. 2012, doi: 10.1109/MC.2012.308.
- [21] K. Ricanek, “Beyond Recognition: The Promise of Biometric Analytics,” *Computer*, vol. 47, no. 9, pp. 87–89, Sep. 2014, doi: 10.1109/MC.2014.236.
- [22] P. Carcagnì, M. D. Coco, D. Cazzato, M. Leo, and C. Distantè, “A study on different experimental configurations for age, race, and gender estimation problems,” *EURASIP J. Image Video Process.*, vol. 2015, no. 1, p. 37, Nov. 2015, doi: 10.1186/s13640-015-0089-y.
- [23] A. González-Briones, G. Villarrubia, J. F. De Paz, and J. M. Corchado, “A multi-agent system for the classification of gender and age from images,” *Comput. Vis. Image Underst.*, vol. 172, pp. 98–106, Jul. 2018, doi: 10.1016/j.cviu.2018.01.012.
- [24] I. Serna, A. Peña, A. Morales, and J. Fierrez, “InsideBias: Measuring Bias in Deep Networks and Application to Face Gender Biometrics,” *ArXiv200406592 Cs*, Jul. 2020, Accessed: Mar. 02, 2021. [Online]. Available: <http://arxiv.org/abs/2004.06592>
- [25] J. Alasadi, A. Al Hilli, and V. K. Singh, “Toward Fairness in Face Matching Algorithms,” in *Proceedings of the 1st International Workshop on Fairness, Accountability, and Transparency in MultiMedia - FAT/MM '19*, Nice, France, 2019, pp. 19–25. doi: 10.1145/3347447.3356751.
- [26] I. D. Raji and J. Buolamwini, “Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products,” in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA, Jan. 2019, pp. 429–435. doi: 10.1145/3306618.3314244.

- [27] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain, "Face Recognition Performance: Role of Demographic Information," *IEEE Trans. Inf. Forensics Secur.*, vol. 7, no. 6, pp. 1789–1801, Dec. 2012, doi: 10.1109/TIFS.2012.2214212.
- [28] 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 Trans. Biom. Behav. Identity Sci.*, vol. 1, no. 1, pp. 32–41, Jan. 2019, doi: 10.1109/TBIOM.2019.2897801.
- [29] K. K. S, K. Vangara, M. C. King, V. Albiero, and K. Bowyer, "Characterizing the Variability in Face Recognition Accuracy Relative to Race," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA, Jun. 2019, pp. 2278–2285. doi: 10.1109/CVPRW.2019.00281.
- [30] J. G. Cavazos, P. J. Phillips, C. D. Castillo, and A. J. O'Toole, "Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?," *IEEE Trans. Biom. Behav. Identity Sci.*, pp. 1–1, 2020, doi: 10.1109/TBIOM.2020.3027269.
- [31] V. Albiero, K. S. Krishnapriya, K. Vangara, K. Zhang, M. C. King, and K. W. Bowyer, "Analysis of Gender Inequality In Face Recognition Accuracy," in *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, Snowmass Village, CO, USA, Mar. 2020, pp. 81–89. doi: 10.1109/WACVW50321.2020.9096947.
- [32] Y. Qiu, V. Albiero, M. C. King, and K. W. Bowyer, "Do Images that Generate Gender Classification Errors Also Cause Face Recognition False Matches?," *M*, p. 9.
- [33] "deepinsight/insightface," *GitHub*.
<https://github.com/deepinsight/insightface> (accessed Mar. 05, 2021).
- [34] "deepinsight/insightface," *GitHub*.
<https://github.com/deepinsight/insightface> (accessed Mar. 05, 2021).

- [35] J. Cheng, Y. Li, J. Wang, L. Yu, and S. Wang, “Exploiting effective facial patches for robust gender recognition,” *Tsinghua Sci. Technol.*, vol. 24, no. 3, pp. 333–345, Jun. 2019, doi: 10.26599/TST.2018.9010090.
- [36] “AFAD-Dataset.GitHub.io by afad-dataset.” <https://afad-dataset.github.io/> (accessed Mar. 05, 2021).
- [37] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, “AgeDB: The First Manually Collected, In-the-Wild Age Database,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 1997–2005. doi: 10.1109/CVPRW.2017.250.
- [38] “Papers with Code - CACD Dataset.” <https://paperswithcode.com/dataset/cacd> (accessed Mar. 05, 2021).
- [39] A. Chatterjee, *imdeepmind/processed-imdb-wiki-dataset*. 2021. Accessed: Mar. 05, 2021. [Online]. Available: <https://github.com/imdeepmind/processed-imdb-wiki-dataset>
- [40] Y.-H. Huang, *b02901145/SSR-Net_megaage-asian*. 2020. Accessed: Mar. 05, 2021. [Online]. Available: https://github.com/b02901145/SSR-Net_megaage-asian
- [41] “Papers with Code - UTKFace Dataset.” <https://paperswithcode.com/dataset/utkface> (accessed Mar. 05, 2021).
- [42] K. S. Krishnapriya, M. C. King, and K. W. Bowyer, “Analysis of Manual and Automated Skin Tone Assignments for Face Recognition Applications,” *ArXiv210414685 Cs*, Apr. 2021, Accessed: May 11, 2021. [Online]. Available: <http://arxiv.org/abs/2104.14685>