Acta Scientific COMPUTER SCIENCES

Volume 7 Issue 7 October 2025

Research Article

Analyzing Bias in Face Recognition and Addressing it with the Perspectum Tool

Aditya Chiduruppa[1] and Praveen Kumar Pandian Shanmuganathan[2]*

[1]High School Junior, Lexington High School, Boston, MA, United States of America
[2]Computer Vision Scientist, Machine Learning Researcher, Florida Institute of Technology, Pittsburgh, PA, United States of America

***Corresponding Author:** Praveen Kumar Pandian Shanmuganathan, Computer Vision Scientist, Machine Learning Researcher, Florida Institute of Technology, Pittsburgh, PA, United States of America.

Received: September 22, 2025
Published: September 29, 2025

Abstract

Facial recognition technology is increasingly utilized for user identification and authentication in various applications. However, several variations across the faces are presented in these systems. This made me wonder how important and prevalent biases are within these systems, namely across races, genders, and beyond. This study initially investigated biases in facial recognition algorithms, focusing on variations in accuracy based on ethnicity, skin tone, and gender. The research comprises two experimental phases. In the first phase, a dataset of facial images from 70,000 individuals, categorized into 11 distinct skin tone groups, was analyzed to evaluate recognition accuracy. Results revealed significant disparities across skin tones, confirming the presence of inherent algorithmic bias. In the second phase, an alternative analysis using the Illinois DOC dataset compared recognition accuracy and False Acceptance Rates (FAR) between African American and Caucasian groups under varying thresholds. To address these biases, Perspectum, a novel metric, will be developed to quantify errors and biases within facial recognition systems in a way that is practical for an enduser making important decisions in a border patrol or law enforcement scenario. Perspectum will be a metric that provides a number between 0 and 100 which can be seamlessly integrated into existing face recognition systems, offering a practical solution to mitigate bias.

Keywords: False Acceptance Rates (FAR); United Healthcare; Facial Recognition Technology (FRT)

Background and Literature Review

Facial recognition technology (FRT) has become increasingly integrated into daily life, from unlocking devices to aiding law enforcement. However, its potential to exacerbate racial biases, especially in policing, has raised significant concerns. Studies reveal racial disparities, such as 41% of Black Americans reporting being stopped due to their race [1], which highlights the need to understand how FRT may contribute to or mitigate such inequities. Notably, FRT has faced challenges in high-stakes situations, like the case involving the CEO of United Healthcare, where the technology initially failed to identify a suspect due to image quality issues [2].

The effectiveness of FRT has been questioned since its earlier uses post-9/11, where the technology failed to prevent false alarms in Tampa, Florida [3]. Moreover, the wrongful arrest of Robert Julian-Borchak Williams due to flawed FRT has spurred cities like San Francisco to ban its use in legal proceedings [4,5]. Research indicates that FRT systems show higher rates of false positives for individuals with darker skin tones, attributed to the underrepresentation of diverse groups in training datasets [5].

Debates persist on whether FRT algorithms should account for race. Some argue for “blinding” models to racial data to reduce bias, while others suggest incorporating race to improve accuracy. Lighting and shading variations complicate these discussions, as some researchers advocate for including skin tone in models, while others warn it could perpetuate bias [6-8]. Studies on “blinding” reveal that while removing racial identifiers didn’t significantly change accuracy, errors varied across demographic groups, suggesting that dataset imbalances, not just awareness of race, drive bias [9].

In focusing on racial bias, research often contrasts

African Americans and Caucasians, finding that African Americans experience higher false positives, particularly under moderate tolerances. These findings reinforce the need for fairness in FRT systems. Critics argue for a more thorough evaluation of AI systems, emphasizing data diversity, improved algorithms, and ethical concerns regarding biased technology in law enforcement. To ensure equitable FRT, ongoing research, better data collection, and transparency in development are critical.

A few different reasons and solutions behind the differences in performance have been researched and proposed. Dark skinned females have the worst facial recognition performance out of any demographic. However, both skin color and hair have been tested out to not be the causes of this. Lip and cheek structure were shown to be the cause of the differences in performance. Additionally, presence makeup on the lips and eyes were presented by the model as a strong indication of a female face [10]. Using color-theoretic methods to systematically lighten or darken skin tones, experiments demonstrated little change in classification accuracy, indicating that skin tone alone wasn't the main cause. Instead, performance disparities likely stemmed from facial morphology differences and correlated features (e.g., makeup cues), suggesting that improving fairness requires addressing broader dataset and model biases beyond just skin color [11].

Considering the 'own race bias' present in humans, a phenomenon where humans tend to be better at identifying members of their own race, models were created to be representative of models from the east asian and western sides of the world. These models were found to do better on people from their areas of origin. After testing the models on a new dataset, they were found to perform better on the new majority present in that dataset, which was caucasians [12]. When specifically looking into gender bias in facial recognition, the conclusion was drawn that there was a need for increased representation of female subjects in facial recognition training data for better accuracy. When the balance of data was struck, there were far more equal results, without degradation of other categories [13].

Additionally, while ROC curves at equivalent FMRs may appear similar across cohorts, these do not reflect real-world settings where a single threshold is used operationally. Image quality (measured via ICAO compliance) differed between demographics and partially contributed to these discrepancies, but did not fully explain them [14].

The practice of "actionable auditing"—systematic testing and public reporting of performance disparities in commercial face recognition and gender classification systems . Through the creation of a specially balanced dataset (with diverse gender and skin-tone representation) and the application of color-theoretic and benchmark analysis methods, she demonstrated stark biases—particularly high error rates for dark-skinned women—and advocated for transparency, policy reforms, and improved dataset practices to mitigate such inequalities [15].

When evaluating 200 face recognition algorithms, significant demographic disparities were found, particularly in false positive rates, which were higher for African American and Asian individuals compared to Caucasians. False negatives were also affected by image quality, especially for darker-skinned faces. However, the most accurate modern algorithms showed minimal demographic differences, suggesting that fairness improves with better-designed models and standardized image capture [16].

When looking into gender bias, Attempts to mitigate using different models have been somewhat successful, finding that certain models like VGG16 and ResNet50 do help to mitigate the effects, however they were still imperfect [17].

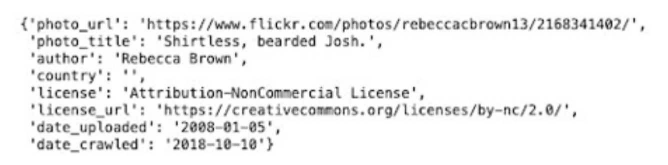
A PNAS study compared professional forensic face examiners, untrained "super-recognizers," and deep learning face recognition algorithms on challenging image-matching tasks. It found that expert humans and modern algorithms achieve similar high accuracy, but combining human and algorithm judgments yields the best performance, suggesting that human–machine collaboration can significantly improve forensic face identification reliability [18].

Bias still exists and could benefit from more ways to address it, including the human side of judgement.

Dataset

The FFHQ dataset is a rich source of diverse images, featuring individuals of varying ages, genders, and races, making it ideal for studying facial recognition biases. The dataset's high resolution and diverse conditions, including varying lighting, accessories, and image quality, provide a realistic testing environment for facial recognition algorithms. The inclusion of facial landmarks further supports advanced analysis, ensuring precise alignment for comparison [19]. However, geographic metadata inconsistencies limited the ability to infer ethnicity or skin tone directly, prompting the switch to image-based skin tone analysis using the STONE Library.

The IDOC (Illinois Department of Corrections) dataset, while less diverse than FFHQ, offers a substantial sample size, particularly in terms of racial representation. With 37,991 African American and 20,992 Caucasian subjects, the dataset allows for targeted analysis of these two racial groups, which are critical for understanding performance disparities in facial recognition systems. The inclusion of physical attributes such as weight, height, and eye color provides additional context for analyzing potential correlations between facial features and recognition accuracy. Although the lack of facial landmarks and repeated subject images limits some of the analysis, the large number of images and detailed demographic data make this dataset an essential resource for evaluating how facial recognition systems perform across different racial and ethnic groups. Together, these datasets complement each other by offering a balance of diversity, image quality, and demographic representation, enabling more comprehensive insights into facial recognition system performance and bias. The combination of these datasets supports the study of both general and specific facial recognition challenges, allowing for nuanced investigations into how demographic factors, such as skin tone and racial background, impact recognition accuracy.



```
{'photo_url': 'https://www.flickr.com/photos/rebeccacbrown13/2168341402/',
 'photo_title': 'Shirtless, bearded Josh.',
 'author': 'Rebecca Brown',
 'country': '',
 'license': 'Attribution-NonCommercial License',
 'license_url': 'https://creativecommons.org/licenses/by-nc/2.0/',
 'date_uploaded': '2008-01-05',
 'date_crawled': '2018-10-10'}
```

Figure 1: Sample Metadata of a user in the FFHQ dataset.

Hypothesis

This project aimed to investigate how variations in skin tone affect the performance of widely used facial recognition algorithms, addressing two key questions: How accurately can algorithms distinguish between individuals with similar skin tones? How effectively can they differentiate between individuals with varying skin tones? Given the global reliance on facial recognition technology in security, authentication, and law enforcement, it is critical to assess whether these systems exhibit biases across different skin tones and ethnicities. Subtle differences in skin color, often overlooked during algorithm design, can significantly impact recognition accuracy, potentially leading to false matches or identification errors. Additionally, while a model might not refer to Skin Color, the classification would also account for facial features associated with certain races and skin colors, that otherwise we would not be able to isolate.

Figure 2: Sample Image from the FFHQ dataset.

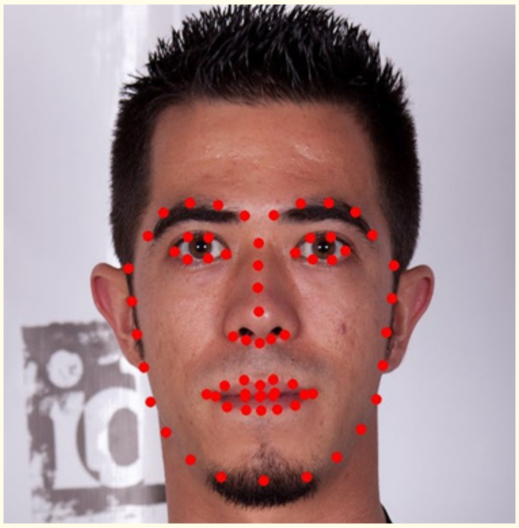


Figure 3: Sample Image from the FFQH dataset with landmarks plotted

To explore these issues, I utilized the FFHQ (FlickrFaces-HQ) dataset, which contains 70,000 high-resolution images from diverse demographic groups. Skin tones were categorized using the STONE (Skin Tone Classifier) Library, which classifies faces into 11 categories ranging from darkest to lightest. This granular classification enabled a detailed analysis of recognition performance across a broad spectrum of skin tones. Widely used facial recognition algorithms were evaluated through the Face-Recognition Library, focusing on their ability to correctly identify or differentiate between faces within and across skin tone groups. Key performance metrics—true positives, true negatives, and false positives—were used to measure system accuracy. While a true positive indicates a correct match between identical faces, a false positive occurs when two different faces are incorrectly identified as the same person, a critical issue in applications like law enforcement. False negatives were not observed in this study due to the dataset's lack of repeated subjects.

To further analyze racial disparities, I focused on African American and Caucasian subjects using preclassified metadata. The algorithm's performance was tested within each racial group, and instances of misidentification (false positives) were compared to identify patterns of bias. This allowed for an examination of whether recognition accuracy varied significantly between these groups, shedding light on racial disparities in algorithmic performance. I specifically chose African Americans and Caucasians due to both the large differences in features and skin tone, and the already existing racial disparities that could be further affected by algorithmic bias.

The hypothesis driving this research was that facial recognition systems might perform less accurately for individuals with darker skin tones due to the historical underrepresentation of these groups in training datasets. Conversely, individuals with lighter skin tones, often overrepresented, were expected to show higher recognition accuracy. By testing across the full spectrum of skin tones, this study aimed to uncover disparities and evaluate whether the algorithms were skewed toward better recognition of lighter-skinned individuals, as suggested by prior studies.

Understanding these biases is critical given the widespread deployment of facial recognition in contexts such as policing, border control, and personal authentication. This research aims to contribute to ongoing efforts to improve fairness and equity in AI systems, ensuring they work equally well across all demographic groups. By quantifying biases and identifying performance gaps, the findings highlight the need for more inclusive training datasets and algorithmic improvements to mitigate disparities and ensure ethical applications of facial recognition technology.

Preprocessing

I initially attempted to infer skin tones using geographic metadata from the FFHQ dataset, assuming that regional demographics could serve as a proxy for skin tone. However, due to incomplete or missing location data, this method proved unreliable. Consequently, I turned to image-based approaches, specifically using facial landmark detection to sample skin color from the bridge of the nose, a region less prone to shadows or occlusions. I initially used the Fitzpatrick scale to classify skin tones into six categories, but this approach faced several limitations. The Fitzpatrick scale's range was too narrow to accurately capture the diversity of skin tones in the dataset, and factors like inconsistent lighting, varying image quality, and interference from accessories further complicated accurate classification.

To address these challenges, I adopted the STONE (Skin Tone) Library, which classifies skin tones into 11 categories (ranging from A for darkest to K for lightest), providing a more nuanced and continuous scale. This significantly enhanced the classification accuracy, offering a finer distinction between tones compared to the Fitzpatrick scale. However, the dataset exhibited significant representation imbalances, with darker skin tones (A, B, C) being underrepresented, particularly with only a few dozen to a few hundred images in these categories. Mid-range tones were overwhelmingly dominant, comprising tens of thousands of images. This imbalance highlights broader issues within facial recognition datasets, where lighter skin tones tend to be overrepresented, skewing algorithmic performance and reinforcing biases in these systems.

Although I initially considered using the Fitzpatrick scale for classification, its limitations—particularly the poor RGB value mapping, inability to isolate points without possible interference from objects and obstructions, and broad, generalized categories—made it unsuitable for this study's requirements [20]. The STONE Library's more detailed categorization system offered a far better foundation for sorting and comparing skin tones in the images. Initially, I considered using the Fitzpatrick scale to classify skin tones, but it posed several challenges. There was no reliable method to map precise RGB values to its categories, and the scale's limited range and generalized classifications proved insufficient for my needs [20]. Instead, I adopted the STONE library, which offers a more detailed and nuanced categorization, using its defined buckets as the basis for comparing and sorting the images. Additionally, STONE did appear to not be accurate to skin tone all the time, experiencing issues when the face had alternative shading, and generally not always being perfectly accurate to skin tone when the real tone differed than what appeared on photo.

In the first FFHQ dataset, we faced an issue of darker skin tones being very underrepresented, as we did not have an even distribution in skin tones. In the second phase, we moved away from purely using skin tones, to using a dataset with clearly prelabeled metadata on race, the IDOC dataset. The FFHQ dataset had this issue in the first phase, but in the second phase, there was a much larger amount of both races provided, with tens of thousands of images for both races provided in IDOC. For facial recognition analysis, I explored a range of libraries. OpenFace, an early consideration, was ultimately impractical due to compatibility issues with modern Python versions and outdated commands [21]. Although I could have invested significant effort to resolve these issues, it would have been inefficient. Instead, I opted for the Face-Recognition Library, which is more modern, actively maintained, and compatible with current Python versions [22]. This library provided a robust API to generate face encodings and compare them efficiently. The Face-Recognition Library encodes facial features into 128 floating-point values, which are compact yet highly detailed, making them ideal for machine-learning applications. Its architecture and ease of use allowed me to quickly overcome the challenges posed by OpenFace, enabling smoother progress in the analysis.

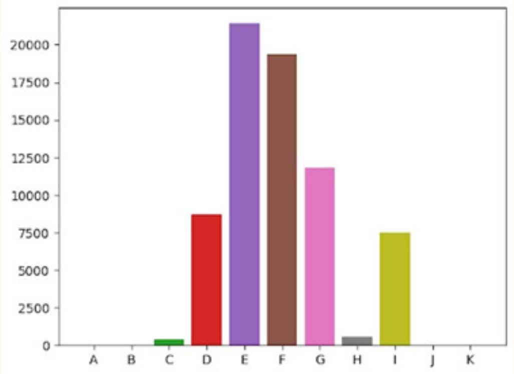


Figure 4: Graph of skin color commonality. A has 30 images, B has 16 images, C has 363 images, D has 8715 images, E has 21408 images, F has 19339 images, G has 11799 images, H has 587 images, I has 7487 images, J has 2 images, and K has 7 images. Variances are due to distribution of data as analyzed by STONE.

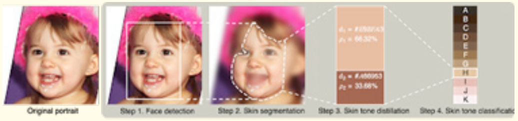
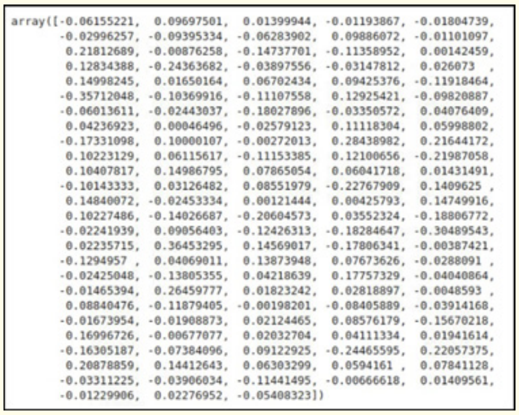

Original portrait
Step 1. Face detection
Step 2. Skin segmentation
Step 3. Skin tone distillation
Step 4. Skin tone classificatio
A
B
C
D
E
F
G
H
I
J
K

Figure 5: Sample of skin tone identification and chart of skin tones.

Experimental methodology

For both experiments, I calculated face encodings using the Face-Recognition Library, processing 99.65% of the FFHQ dataset (69,753 usable images). The process involved detecting faces, generating unique numerical representations (encodings), and comparing them to compute similarity scores and key metrics like True Positives, False Positives, and True Negatives. Initially, I used the face_distance function to compute Euclidean distances but switched to the more efficient compare_faces function, which directly returns a Boolean result for matches, using a tolerance of 0.5 to balance sensitivity and specificity.



```
array([-0.06155221,  0.09697501,  0.01399944, -0.01193867, -0.01804739,
       -0.02996257, -0.09395334, -0.06283902,  0.09886072, -0.01101097,
        0.21812689, -0.00876258, -0.14737701, -0.11358952,  0.00142459,
        0.12834388, -0.24363682, -0.03897556, -0.03147812,  0.026073  ,
        0.14998245,  0.01650164,  0.06702434,  0.09425376, -0.11918464,
       -0.35712048, -0.10369916, -0.11107558,  0.12925421, -0.09820887,
       -0.06013611, -0.02443037, -0.18027896, -0.03350572,  0.04076409,
        0.04236923,  0.00046496, -0.02579123,  0.11118304,  0.05998802,
       -0.17331098,  0.10000107, -0.00272013,  0.28438982,  0.21644172,
        0.10223129,  0.06115617, -0.11153385,  0.12100656, -0.21987058,
        0.10407817,  0.14986795,  0.07865054,  0.06041718,  0.01431491,
       -0.10143333,  0.03126482,  0.08551979, -0.22767909,  0.1409625 ,
        0.14840072, -0.02453334,  0.00121444,  0.00425793,  0.14749916,
        0.10227486, -0.14026687, -0.20604573,  0.03552324, -0.18806772,
       -0.02241939,  0.09056403, -0.12426313, -0.18284647, -0.30489543,
        0.02235715,  0.36453295,  0.14569017, -0.17806341, -0.00387421,
       -0.1294957 ,  0.04069011,  0.13873948,  0.07673626, -0.0288091 ,
       -0.02425048, -0.13805355,  0.04218639,  0.17757329, -0.04040864,
       -0.01465394,  0.26459777,  0.01823242,  0.02818897, -0.0048593 ,
        0.08840476, -0.11879405, -0.00198201, -0.08405889, -0.03914168,
       -0.01673954, -0.01908873,  0.02124465,  0.08576179, -0.15670218,
        0.16996726, -0.00677077,  0.02032704,  0.04111334,  0.01941614,
       -0.16305187, -0.07384096,  0.09122925, -0.24465595,  0.22057375,
        0.20878859,  0.14412643,  0.06303299,  0.0594161 ,  0.07841128,
       -0.03311225, -0.03906034, -0.11441495, -0.00666618,  0.01409561,
       -0.01229906,  0.02276952, -0.05408323])
```

Figure 6: Sample face encoding.

The experiments consisted of intra-class (same skin tone) and inter-class (different skin tones) comparisons across 11 skin tone categories (A to K), with 66 sets of pairwise comparisons (11 intra-class and 55 inter-class), involving thousands of image pairs. This was computationally intensive, requiring significant processing time and storage. The second experiment focused on African American and Caucasian subsets, using a dataset of nearly 70,000 images.

The first experiment primarily focused on the false positive rate (Imposter Score), which is crucial for evaluating the reliability and fairness of facial recognition systems. In the second experiment, I used the False Acceptance Rate (FAR), calculated as the ratio of false positives to the sum of false positives and true negatives, to evaluate performance differences between the African American and Caucasian subsets. I normalized the data to account for varying group sizes—skin tone categories ranged from 17 to tens of thousands of images, while African American (40,000 images) and Caucasian (20,000 images) subsets required normalization of FAR.

Results from the first experiment showed that no comparison set exceeded a 1% false positive rate, but false positives were higher among individuals with similar skin tones, indicating difficulty in distinguishing between closely matched tones. The second experiment revealed performance disparities across tolerance levels: at lower tolerances (0.1–0.4), both subsets performed similarly with near-zero FARs. However, at higher tolerances (e.g., 0.9), the African American and Caucasian FARs were similar (1.0 vs. 0.94), but significant disparities emerged at midrange tolerances. At a tolerance of 0.5, the Caucasian FAR was nearly zero, while the African American FAR was 0.07. At tolerance 0.6, the gap widened (0.01 vs. 0.4), peaked at 0.7 (0.13 vs. 0.81), and remained notable at 0.8 (0.59 vs. 0.97).

In both experiments, due to a lack of multiple instances for a single subject, no false negatives were measured for either dataset.

These findings underscore persistent bias in facial recognition systems, particularly at mid-range tolerances, with the system performing worse for African American subjects compared to Caucasian subjects. This highlights the need for further improvements in the technology to address these disparities and ensure fairer and more accurate performance across all demographics.

Results and Need for Perspectum

The experiments reveal significant issues with the reliability and fairness of facial recognition systems. A 1% false positive rate, though seemingly low, can have serious consequences in high-stakes contexts like law enforcement. The increased false positive rate among individuals with similar skin tones suggests that the system struggles to differentiate subtle intra-group variations, indicating potential bias. This is particularly concerning for underrepresented skin tone categories, where small errors can lead to disproportionately negative outcomes. The findings highlight the need for improved model training with diverse datasets to ensure fairness and accuracy.

Graphs of false positive rates across skin tones (A to K) show more frequent misidentifications in extreme lighter tones, likely due to the specific FFHQ dataset imbalances. Subtle variations within similar skin tones were often misidentified, suggesting limitations in the system's ability to handle fine-grained differences. The second experiment showed high accuracy overall, but a pronounced bias favoring Caucasian subjects, highlighting the need to address such disparities.

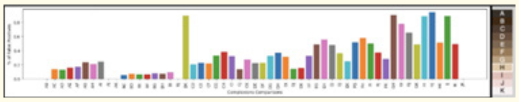


Figure 7: Graph of all of the skin tone comparison's false positive rate, excluding ones that have the two same skin colors compared against each other.

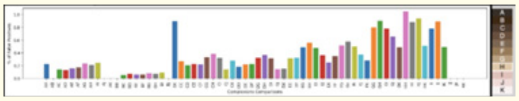


Figure 8: Graph of all of the skin tone comparison's false positive rates.

Perspectum

To address these challenges, I propose "Perspectum", a trustworthiness standard for face recognition systems. Perspectum aims to quantify bias, assess reliability, and provide a transparent framework for interpreting the results of facial recognition analyses. It introduces a trustworthiness score based on metrics such as bias quantification, accuracy disparities, and false-positive rates, offering a standardized way to evaluate system performance across diverse populations. Furthermore, Perspectum will emphasize transparency by standardizing reporting formats, ensuring results are interpretable and actionable.

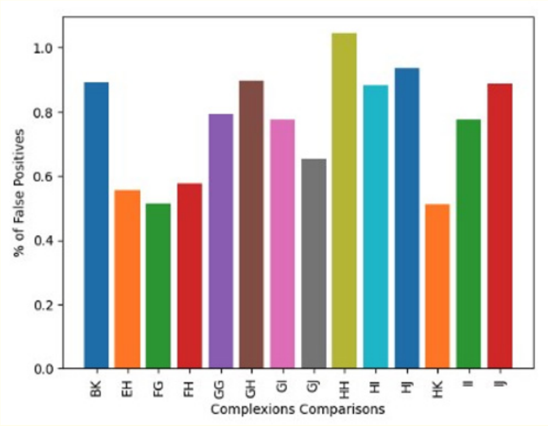


Figure 9: Graph of all skin tone group comparisons with false positive rates of over 0.5%.

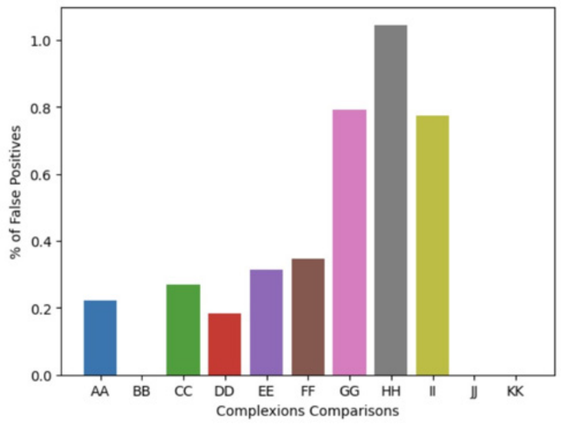


Figure 10: Graph of false positive rate for all intra-skin tone group face comparisons. The images with no false positives are likely because they had a very small number of images, which are 17 for B, 2 for J, and 7 for K.

This framework will be designed for integration into various domains, including law enforcement, enterprise applications, and consumer platforms, to mitigate risks and uphold ethical standards. By providing a certification process for face recognition systems, Perspectum could guide policymakers, legal professionals, and developers in making informed decisions about the reliability and fairness of these technologies. Ultimately, the adoption of such a standard is essential to minimize harm, promote accountability, and ensure equitable outcomes in face recognition applications.

The algorithm and logic behind it is as follows: Whenever a Face Recognition Algorithm becomes available and is ready for implementation, Perspectum will run on top of the Face Recognition algorithm as a tool for the user who is evaluating subjects. Perspectum will have a huge facial dataset that continues to improve and evolve, collecting different possible faces in the world with combinations of different skin tones, racial variations, and other factors (such as gender and age). Once a face recognition algorithm is ready to be used, a predefined dataset from Perspectum will be run through the Face Recognition algorithm to identify the FAR and FRR scores for all the users (FRR is False Reject Rate, with an equation of FP/(TP+FP). Since the current phases have no repeat images for one subject, it cannot be calculated currently), and it will be noted as part of the Perspectum catalog. This is run as part of the calibration process before the actual usage.

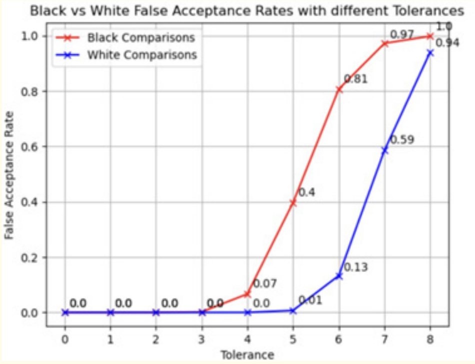
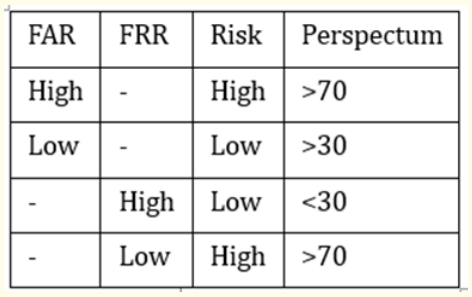


Figure 11: Graph of False Acceptance Rates against Tolerance for black and white comparisons, from 0.1(labeled as 0) to 0.9(labeled as 8).

Now, when a probe (a face that is currently being evaluated) comes into the Face Recognition system during the actual usage, Perspectum will perform a similarity assessment. It compares the probe to all the faces in the predefined dataset, and the most similar match of the probe against the dataset is then picked. This match is then evaluated against the score that was extracted during the FAR and FRR calculations. If the match has a higher FAR, then it is evident that the probe will also have a higher FAR, due to bias in the system, and the person needs to be inspected additionally —although the system might show a lower risk. If the template has a higher FRR, then the person could be flagged as high risk from the face recognition system, however, it is because of the systematic bias and not the face itself.

Upon looking at this bias, Perspectum will then convert the FAR and FRR values to a Perspectum Score that will be provided to the agent who is responsible for clearing the subject. If the FAR and FRR values are outside of a threshold (less than 80% and more than 20%) it will then convert to a Perspectum score between 0 and 100. This score becomes a value that can be acted upon the probe if it is severe or acceptable before making a decision.



FAR	FRR	Risk	Perspectum
High	-	High	>70
Low	-	Low	>30
-	High	Low	<30
-	Low	High	>70

Figure 12: Decision Table for Perspectum Score Creation.

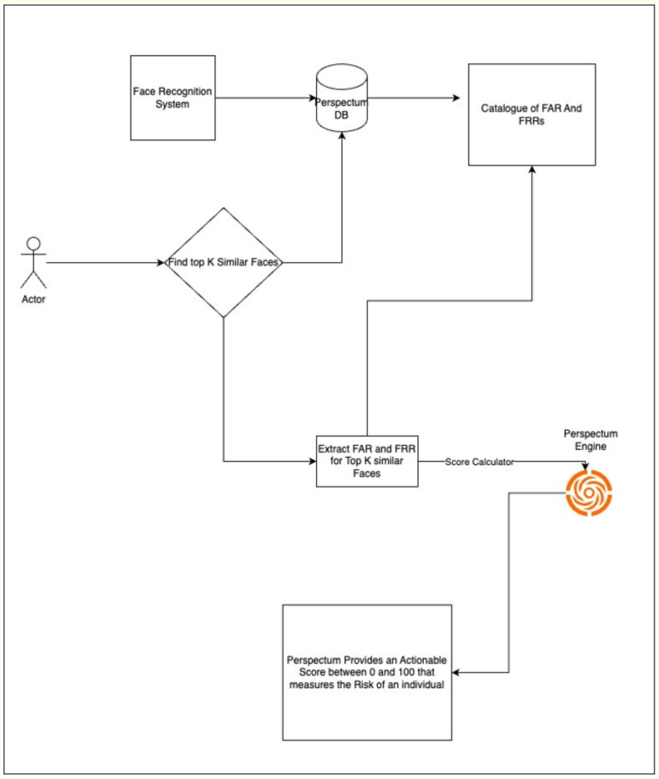

Face Recognition System
Perspectum DB
Catalogue of FAR And FRRs
Actor
Find top K Similar Faces
Extract FAR and FRR for Top K similar Faces
Score Calculator
Perspectum Engine
Perspectum Provides an Actionable Score between 0 and 100 that measures the Risk of an individual

Figure 13: Architecture Diagram for Perspectum.

Conclusion and Future Works

The data reveals a significant bias in facial recognition systems, particularly when distinguishing individuals with similar skin tones. This bias increases the risk of false positives, which can lead to racial profiling and discrimination, particularly in law enforcement contexts. Higher false positive rates among certain racial or ethnic groups could lead to unjust surveillance or wrongful accusations, exacerbating existing systemic biases. Additionally, overrelying on skin tone or racial features in algorithms can undermine accuracy and fairness, eroding public trust in the technology.

To mitigate this, reducing the influence of skin tone in facial recognition models is essential. "De-biasing" the models by limiting reliance on racial features or training them on more diverse datasets can improve accuracy across demographic groups. Minimizing the focus on race can also reduce human biases embedded in machine learning systems. As next steps, we would like to conduct further research into other factors, such as facial hair, aging, gender, etc. to analyze their impact and relevance. While achieving perfect balance may be difficult, addressing these biases directly by providing users with clear accuracy indicators can help. Ultimately, improving fairness and accuracy is crucial for ensuring that facial recognition technology serves as a reliable and equitable tool for security, without perpetuating racial disparities.

As our next step, we will perform a rigorous inter and intra class comparison to isolate where the bias originates and where it is more prominent. Intra-class will quantify how skin tone, illumination, pose, facial hair, occlusions and aging affect the genuine and imposter scores for the same individual. We plan to measure intra-class variance, genuine score distribution and imposter score distribution. On the Inter-class analysis, we will look for imposter pair separability across and within skin tone groups to detect where False positives are higher and will report on the same.

Bibliography

1. ACLU. "Racial Profiling".
2. Kimery Anthony. "Limitations of FRT apparent in search for United Healthcare CEO's killer". BiometricUpdate.com (2024).
3. K Gates. "Can computers be racist?". *Juniata Voices* 15 (2015).
4. GM Haddad. "Confronting the biased algorithm: the danger of admitting facial recognition technology results in the courtroom". *Vanderbilt Journal of Entertainment and Technology Law* 23 (2020): 891.
5. K Krishnapriya., *et al*. "Issues related to face recognition accuracy varying based on race and skin tone". IEEE Trans. Technol. Soc. 1.1 (2020): 8-20.
6. M Jmal., *et al*. "Classification of human skin color and its application to face recognition". presented at the MMEDIA 2014: the Sixth International Conference on Advances in Multimedia, IARIA (2014).
7. Y Ban., *et al*. "Face detection based on skin color likelihood". *Pattern Recognition* 47.4 (2014): 1573-1585.
8. B Dhivakar., *et al*. "Face detection and recognition using skin color". Presented at the 2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN), IEEE (2015): 1-7.
9. S Wehrli., *et al*. "Bias, awareness, and ignorance in deep-learning-based face recognition". *AI Ethics* 2.3 (2022): 509-522.
10. V Muthukumar., *et al*. "Understanding unequal gender classification accuracy from face images". ArXiv Prepr. ArXiv181200099 (2018).
11. V Muthukumar. "Color-theoretic experiments to understand unequal gender classification accuracy from face images". presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019).
12. P J Phillips., *et al*. "An other-race effect for face recognition algorithms". ACM Trans. Appl. Percept. TAP 8.2 (2011): 1-11.
13. M Atay., *et al*. "Evaluation of gender bias in facial recognition with traditional machine learning algorithms". presented at the 2021 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE (2021): 1-7.
14. K Vangara., *et al*. "Characterizing the variability in face recognition accuracy relative to race". presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019).
15. J Buolamwini and ID Raji. "Actionable auditing: investigating the impact of publicly naming biased performance results of commercial AI products". Conference on Artificial Intelligence, Ethics, and Society, (2019).
16. P Grother., *et al*. "Face recognition vendor test (fvrt): Part 3, demographic effects". National Institute of Standards and Technology Gaithersburg, MD (2019).

17. T Gwyn and K Roy. "Examining gender bias of convolutional neural networks via facial recognition". *Future Internet* 14.12 (2022): 375.

18. P J Phillips., *et al*. "Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms". *Proceedings of the National Academy of Sciences of the United States of America* 115.24 (2018): 6171-6176.

19. NVlabs. "Flickr-Faces-HQ Dataset (FFHQ)". NVlabs, (2019).

20. W Coleman., *et al*. "Updating the Fitzpatrick classification: the skin color and ethnicity scale". *Dermatology Surgery* 49.8 (2023): 725-731.

21. B Amos., *et al*. "Openface: A general-purpose face recognition library with mobile applications". *CMU School of Computer Science* 6.2 (2016): 20.

22. A Geitgey. "face_recognition". Adam Geitgey (2018).