WHITEPAPER

Mitigating Demographic Bias in Facial Presentation Attack Detection

Methodologies for Algorithm Development in Support of Responsible AI Principles

IDR&D

Introduction

The use of biometrics has become common in personal, commercial, and government identity management applications. These systems use artificial intelligence ("AI"), primarily in the form of machine learning, to recognize individuals based on their unique physical or behavioral characteristics.

However, there has been a concern in the public, media, and academia over the existence of systemic bias in systems that use face biometrics for automated decision making. With an increasing utilization of biometrics for identity verification and authentication, the need for consistent accuracy across races, ages, and genders is paramount. Where there is bias, there is the potential for individuals to experience different biometric false match and false non-match error rates based on their appearance and demographics. For example, an individual of a particular race might be more likely to be incorrectly matched to a criminal suspect as a result of a biometric search. That person might experience more difficulty in using their biometrics to authenticate or be subject to a higher suspicion of a fraud attempt. Best practices can be applied to algorithm development that mitigate demographic bias in support of Responsible AI principles.

Bias is not limited to facial matching. It also applies to AI-driven facial presentation attack detection, which is used to prevent spoofing attacks such as the presentation of printed photos, video replays, and 2D and 3D masks on biometric systems. Collectively, these are called "presentation attacks," and anti-spoofing countermeasures are referred to as presentation attack detection ("PAD")[1]. The ISO standard 30107-3 specifies the definition of attacks and methods to assess the performance of PAD solutions. PAD is essential where facial recognition is used for unsupervised processes such as remote onboarding, mobile login, and physical access control. This is because biometric facial matching systems readily match with pictures and videos of a person, enabling a bad actor to spoof the biometrics, e.g. with printed-on-paper or digitally-displayed photos. PAD methods that employ signals indicating the absence of an attack as opposed to an attack have been referred to as "liveness detection", and the terms are often used interchangeably.

This paper clarifies the meaning of demographic bias in AI-based facial PAD systems, provides examples of methods AI developers use to remove bias, and shows the results of independent lab testing of ID R&D's PAD software. The paper shares performance improvements that were observed as a result of putting Responsible AI in place for products being used in large-scale deployments.

[1] https://www.iso.org/standard/53227.html

Demographic Bias: An Issue for Many AI Applications

"AI bias" is the bias in decisions made by machine learning algorithms. Biases can be caused by imbalances in the underlying training data, or by the errors that creep into algorithm development as a result of human prejudices and false assumptions.

In training, biases can result from the underrepresentation of certain demographics as well as the use of biased classification decisioning data. Consider the use of an AI algorithm to screen university applicants. If the algorithm is trained on historical admission decisions, a range of human biases that intentionally or unintentionally favor certain profiles could be built into the algorithm. An example is training a model based on students from specific geographic locations. In geographies with a high correlation to ethnicity, the output will inadvertently favor certain ethnicities over others.

One of the most visible examples of AI bias is the US Court's use of a decision support tool called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) to predict the likelihood of a defendant becoming a repeat offender. The model used has been found to predict twice as many false positives for black offenders than for white offenders.

Another well-known case is Amazon's now-defunct use of an AI recruiting tool that was biased against women. It was found that the algorithm used to vet candidates was trained on older data consisting of a disproportionate number of male applicants during a time when the tech industry consisted predominantly of men.

Defining Bias For Presentation Attack Detection

Knowing the general meaning of bias in AI systems, the question for PAD is what is the nature of bias in a presentation attack detection system?

Let's start with the two types of errors that occur in a PAD system. One type of error is the false negative, called an Attack Presentation Classification Error, when the system fails to detect a spoof attack. In a spoof attack, a person is not suffering from discrimination because, by definition, the attack is not a genuine person. Therefore, attack errors are not relevant to measuring bias.

The other type of error is the false positive, where the PAD system identifies a valid, live person as a spoof. This type of error is called a Bona Fide Presentation Classification Error in PAD systems. The rate at which a system produces Bona Fide errors is called the Bona Fide Presentation Classification Error Rate ("BPCER"). A Bona Fide error creates customer annoyance and inconvenience, and this is also where bias can occur in PAD systems. For example, if the system produces Bona Fide errors for Black people at a higher rate than White people, or males at a lower rate than females who wear hijabs, then the system is biased and thus unfair. To be considered unbiased, a PAD system must exhibit a BPCER that is equal across race, age, and gender.

However, just being equal is insufficient; the BPCER must also be sufficiently low to ensure a good user experience. For example, If the BPCER is 25% across all demographics, it may be unbiased, but at 25% BPCER, the genuine user for all demographics experiences rejection one out of four times, which in practice generates significant user complaints.**Therefore, for a PAD system to be considered unbiased, ID R&D requires that the PAD system must 1) be statistically equal across demographics and 2) must have a low absolute BPCER.** In other words, a lack of bias is only considered as having been achieved where there is no sacrifice in user experience.

Guidelines for Mitigating Bias in Support of Responsible AI Principles

To achieve an unbiased system, ID R&D prioritizes Responsible AI principles encapsulated in two primary tenets that follow.

Responsible AI Tenet #1:
Study And Encourage Responsible Machine Learning Techniques

The lack of diversity in training data is a well-known issue in machine learning. Data sets often have biased distributions of demographics (gender, race, age, etc.), and machine learning models are trained to exploit whatever correlations exist in this data, leading to discrimination against under-represented groups. Machine learning engineers can apply specific techniques to reduce the effect of imbalanced data. Some of these techniques are listed below:

- **Random oversampling/undersampling**. One of the most effective methods used by researchers to avoid sampling bias is using a random oversampling or undersampling approach. This method addresses imbalanced data sets to ensure that classes with minority representation are not overlooked. In random undersampling, examples from the majority class are randomly deleted from the training data set until a more balanced distribution is reached. In oversampling, examples from the minority class are duplicated to supplement the training data.
- **Weighted loss.** A weighted loss methodology explicitly penalizes two types of errors (i.e. false-negative and false-positive errors) in a binary classification problem. This loss is therefore useful for imbalanced classification in different tasks in which the samples of one class are rare compared to another and should be penalized more if misclassified.
- **Sampling with dynamically adjusted weights.** The motivation for choosing higher weights for some groups is to emphasize groups with lower performance. However, we expect the relative ranking of groups to change during the course of the training: the worst-performing group is assigned the highest weight at first, which leads to improvements for this group, and after a certain number of training steps another group might become the worst-performing group.
- **Data synthesizing.** Data synthesizing can also help with bias mitigation. Generative style-based architectures are able to capture fine-grained aging patterns by conditioning on multi-resolution age-discriminative representations. Simpler face morphing or recoloring techniques can also be used in data augmentation.

Responsible AI tenet #2: Identify Gaps And Correct Issues In Data Sets

Machine learning models have three basic limitations that must be considered:

1. The accuracy of a model is largely dependent on the number of training examples. It is inherently harder for a model to learn categories that lack ample training data.
2. Being statistical in nature, machine learning models aim to deliver accurate predictions for the majority of examples. Therefore, under-represented groups in a data set may be overlooked -- ultimately ignoring the diversity and propagating the bias.
3. Machine learning with constraints is a challenge across the industry that frequently results in increased fairness where errors are equally distributed but at the expense of overall model accuracy. In other words, there may be a tradeoff between fairness and accuracy.

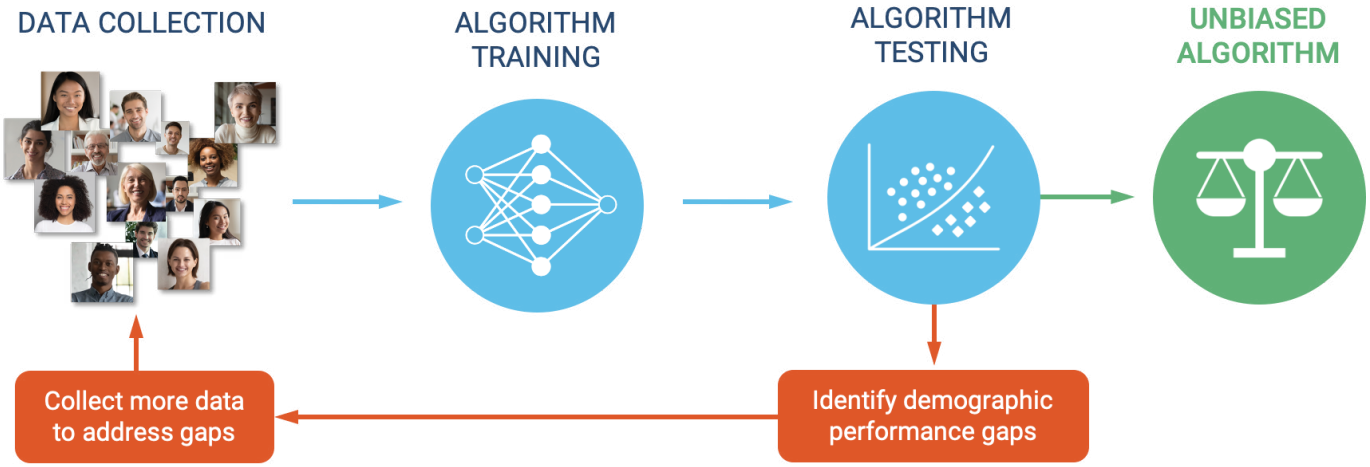


Figure 1. Facial anti-spoofing algorithm development pipeline with focus to mitigate demographic bias.

Methodologies: Putting Best Practices into Action

ID R&D adheres to the principles of Responsible AI. The following are examples of implementations of some of the best practices as outlined above.

Consider a data set containing the facial images of people of different ages. They can be split into the following age groups in years: 0-14, 14-25, 25-40, 40-55, 55-75.

Age group	0-14	14-25	25-40	40-55	55-75
Number of images	100	2000	3000	2500	250

The first and last groups have a relatively low number of samples. Using the default training procedure will lead to trained networks with acceptable performance only on the middle groups that have more samples.

Methodology #1: Adding Real or Synthetic Data

Of the many candidate approaches to improving model fairness, the most impactful but often most challenging way is by adding more data, either synthetic or real.

The most traditional way to enrich a training data set is to collect as much real data as possible. Although it's a difficult process, it is the most effective. The more accessible that real and well-balanced training data is, the easier it is to train a fair model.

Adding synthetic data can also be extremely effective. Modern GANs (generative adversarial network) facilitate synthesis of different images based on specific requirements. An example is using a GAN network to "change" the age of a person by training the network to do so based on large volumes of data. From real data, a GAN can be trained to generate synthetic versions of images of people in an age group where more samples are needed.

GANs also can be used to close more challenging gaps, such as caused by people with facial hair or wearing makeup or glasses. The variety of modifications that can be applied to creating a synthetic face image is enormous and offers a virtually infinite number of ways to enrich training data.

Despite the breadth of a GANs capabilities, it has some drawbacks. It can be extremely difficult to train such a network because the process can be unstable and require a lot of work to label training data (e.g. sunglasses, beard, makeup) and then validate the results.

Methodology #2: Oversampling

The default training procedure randomly samples data without considering the group counts. To make the network unbiased given the distribution of the data, the procedure should be modified to account for the groups with less data, and the most straightforward way to do so is to **oversample** the samples of the smaller groups. Thus, instead of random sampling from the training data set, we will sample images such that each training iteration network considers the same number of images from each group.

A related "focal loss" approach gives the samples in the smaller groups higher weight during calculations, which effectively forces the network to work hardest on the samples in those groups that are hardest to classify.

Methodology #3: Data Augmentation

Further bias reductions can be achieved with augmentation of the original data. This is particularly effective when data is in short supply. For instance, it is possible to rotate, flip, compress, or otherwise process an image and then include it in the data set as a new sample.

This approach brings the added benefit of increasing the ability of the network to generalize. For example, a training data set might consist predominantly of vertically oriented faces (i.e. portrait versus landscape). But in the real world, users may accidentally submit a rotated face. This scenario can be accommodated if during network training, rotated images are added to the network. It is also possible to apply harder augmentations to the under-represented groups to make the difference between data even larger. Although these methods help us to reduce bias, they are still less effective than data-driven approaches.

Results: Examples of Bias Reduction Success

ID R&D first launched its facial liveness product, IDLive® Face, in August 2019. Upon analyzing the product for demographic gaps in the earlier releases, ID R&D found that the performance for some races and age groups was better than others. The team took a number of actions to address this behavior and continues to apply the aforementioned methods to achieve fairness across demographics. The following are two examples of the impact of our bias reduction efforts.

Example 1 - Lower BPCER in Under-represented Categories

ID R&D conducted a detailed analysis of the categories with lower performance. For the category of race, data scientists looked at the groups Black, Asian, Caucasian, and Indian.

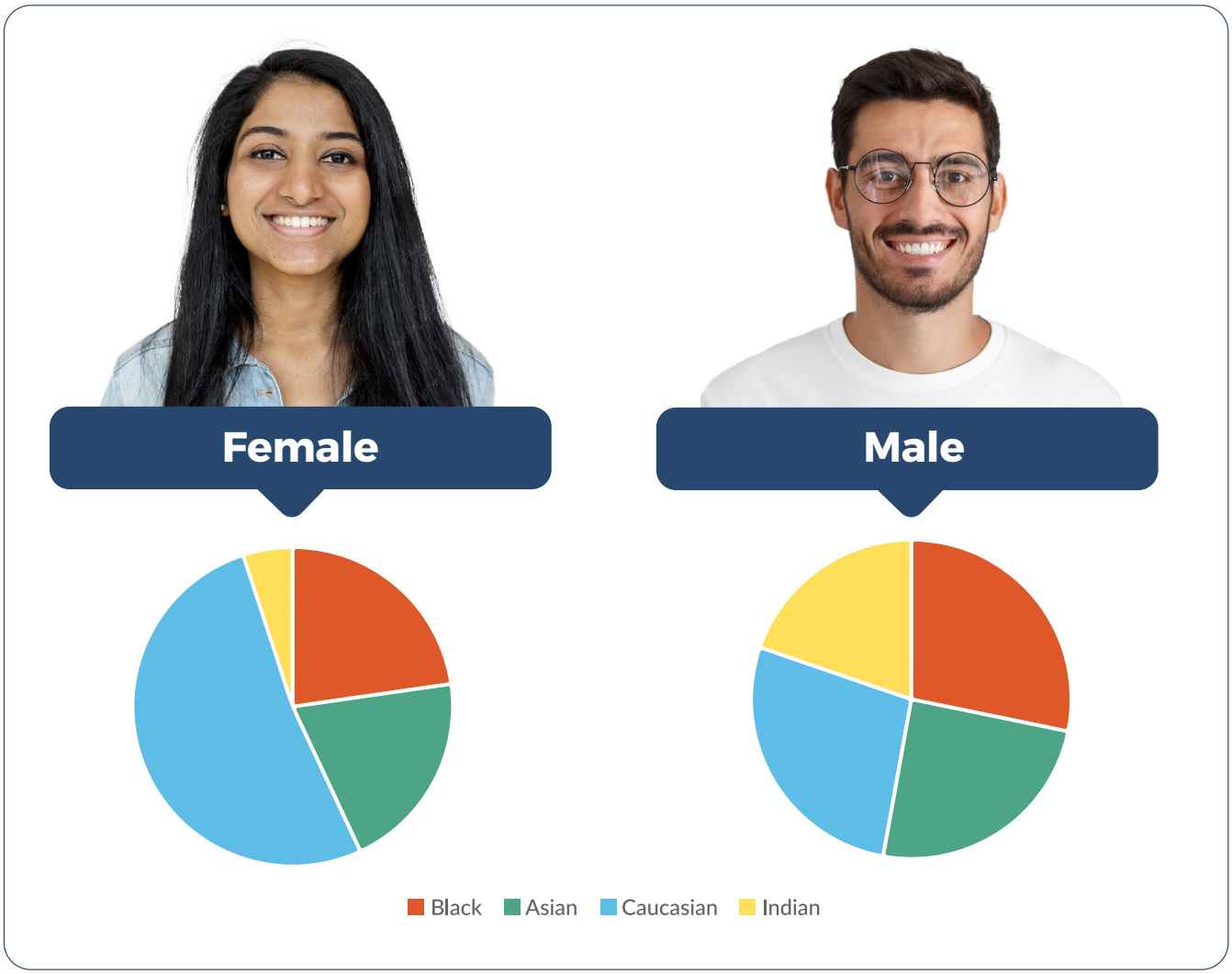


Figure 2. Distribution by race and gender.

Knowing the distribution, the following conclusions can be drawn:

- More data needs to be collected for under-represented categories
- During training, the focus should be put on the less represented categories; for example, young and elderly people
- When evaluating the quality of the algorithm to select the best solution, it is worth making a correction for the data imbalance and choosing a model that works equally well in all categories

The actions resulting from this data analysis enabled ID R&D to lower the BPCER in under-represented categories as well as reduce the variance between demographic groups, which you will see later in Figure 3.

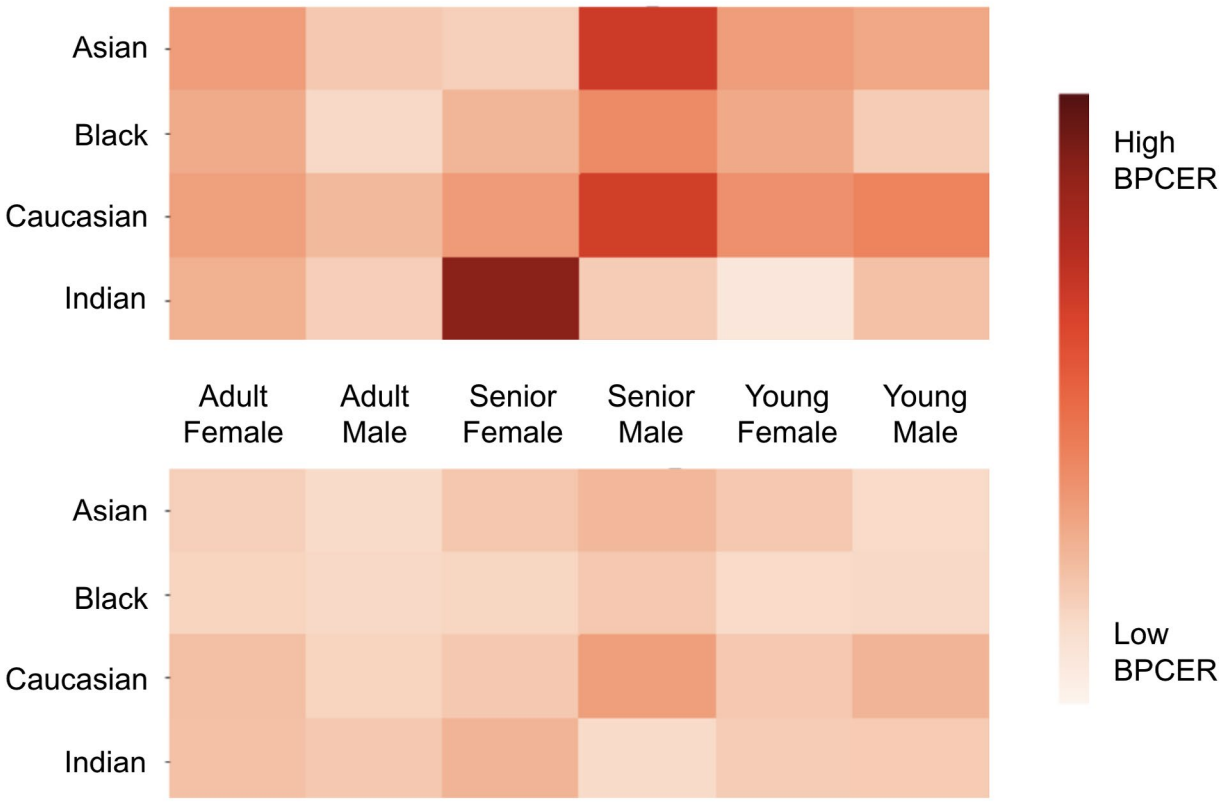


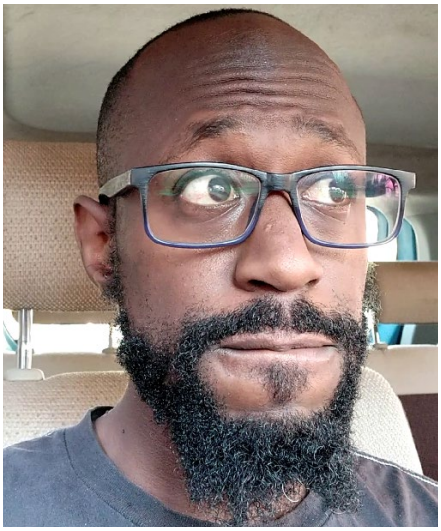
Figure 3. BPCER heatmap before and after applying more data for poor data categories and weighted sampling while training.

Example 2 - Lower BPCER for Sub-Category Features

In addition to general analysis by category, ID R&D also conducted a detailed analysis of the conditions in which the product performed the worst. For categories where the product underperformed, teams analyzed individual images in the data set to determine the areas of the image that the neural network deemed "important" when making its prediction. When mitigating bias, this method helps determine discriminating areas in the image.

A

B



C

D

Figure 4. Heat maps of neural network demonstrate the area of interest of the facial anti-spoofing algorithm: a) original image of a woman in the national headdress, b) heat map of neural network has higher attention on the medium and bottom parts of the headdress, c) original image of a man with mustache and beard, d) heat map of neural network has attention on the bottom part of the face as well as on the part of beard close to the ear.

Using this method, ID R&D found deterioration in the quality of liveness detection for people wearing hijabs. After finding the flaw, additional data was collected to ensure that the models were more likely to encounter images of people in hijabs, hoods, and other headdresses during training. The team also set up a separate test base with people in hijabs so that during the validation stage, we could focus more attention on these conditions, making sure that the models began to work better on hijabs. Thus, we have reduced the BPCER on hijabs by 20%.

Test Report - ID R&D's Current Bias Performance

According to ID R&D's definition of bias, the ultimate goal is to operationalize software development methodologies in support of Responsible AI principles to achieve as close as possible to 0% BPCER across age, race, and gender.

In support of this goal, ID R&D commissioned the first independent evaluation of bias in facial liveness detection. Bixelab, an independent laboratory fully accredited by NIST for biometric testing, performed a comprehensive assessment of ID R&D facial liveness detection technology for bias in target demographics, including age, gender, and race/ethnicity.

Following error metric calculation methodology has been used for the calculation of Bias:

For N >2, Bias per group

$$Bias_{BPCER} = \frac{1}{(N-1)}\left[\sum_{N} BPCER - BPCER_{min}\right] - BPCER_{(second\ besst)}$$

For N =2, Bias per group

$$Bias_{BPCER} = BPCER_{max} - BPCER_{min}$$

Where N is the total number of attributes in a demographic group. For example, the attributes in *Race* group are *Caucasian*, *African*, *East Asian*, and *South Asian*.

The results of the evaluation demonstrate that the software performs with levels of bias below critical target values. These are illustrated in the following table. The full report with complete details of the test are available from ID R&D.

Group	Bias metric	Critical Value (95% confidence level)	Null Hypothesis Accept: Bias Metric < Critical Value Reject: Bias Metric > Critical Value	Bias: Unbiased if Null Hypothesis is accepted Biased if Null Hypothesis is rejected
Age	0.0008	0.002	Accept	Unbiased
Gender	0.002	0.004	Accept	Unbiased
Race/ethnicity	0.0025	0.0031	Accept	Unbiased

Table: Summary of performance variation measured for each demographic group and corresponding bias decisions

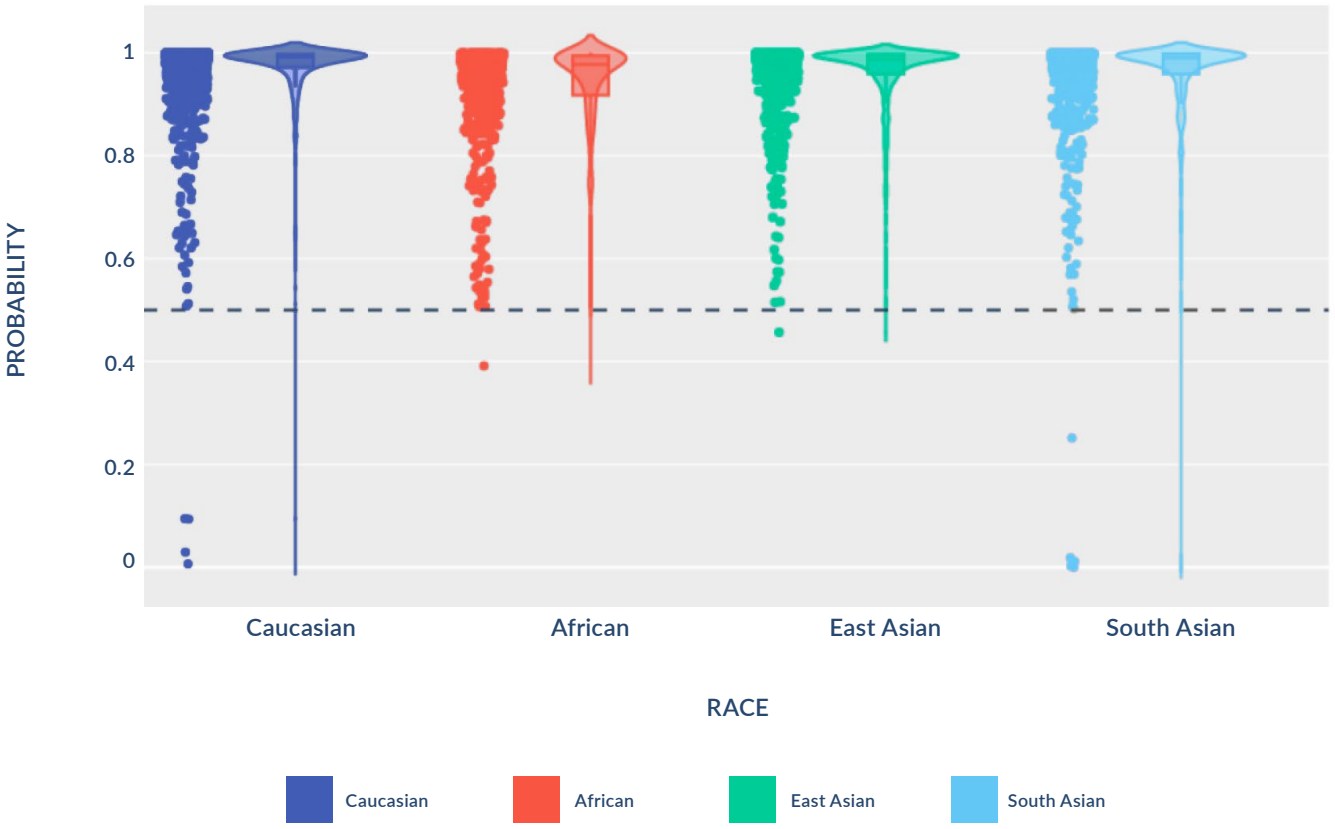


Chart: Violin plots illustrate the probability of false rejection for different races

Conclusion

Bias must be actively addressed for both biometric comparators and for presentation attack detection. Bias is particularly pronounced for face biometrics. Responsible AI principles guide the use of methodologies to remove bias. Independent third party validation proves that applying these principles results in a Presentation Attack Detection system that is fair and generally free of bias.

ID R&D's neural networks were trained on extremely large amounts of image data. By ensuring adequate representation from all demographics, ID R&D PAD algorithms can perform nearly equally across all races and genders. Future work will continue to apply these techniques towards lowering bias and driving down BPCER across all demographics. The work will continue to be adhere tomethodologies that support Responsible AI principles.

Learn more about ID R&D's **passive facial liveness.**

About ID R&D

ID R&D is an award-winning biometrics company on a mission to deliver the next generation of “zero-effort” frictionless authentication and fraud prevention. With extensive expertise in the science of biometrics and the industry’s leading R&D team, we deliver a new breed of AI-driven voice, face and behavioral biometrics, and unique face, voice, and document liveness detection capabilities.

1350 Broadway, Suite 605
New York, NY 10018 USA
info@idrnd.net

www.idrnd.ai