**The Challenges of Using FRT in court cases(b/c of of algorithmic bias)**

* **Focus on the surveillance apparatus as a whole**
  + **A world of cameras is being built around us.**
  + **Government has been using this in court cases**

**Topic:**

* Facial recognition technology has hugely grown in popularity in recent years. What are the ethical implications of more and more of the world using this technology, surveillance footage, and their security systems to track humans? Specifically, how could racial biases contained in facial recognition models create ethical problems when the technology is implemented?
* Research side: Redo (ish) previous iterations of facial recognition racial bias studies. Compare and contrast results; are the differences concerning?

**Initial Notes:**

* Specific setting as our motivation
* Keep your analysis on open source models
  + survey/analysis of outcomes from existing facial recognition models
* Algorithmic bias on specific attributes
  + How this bias can impact the application setting around which we motivated our study?
* Paper should focus on one big “ethical concern”
  + Algorithmic biases → racial biases
    - How could things deepen (racial) bias through their interpretations
* Paper:
  + First have some original experiments

Paper:

* First have some original experiments
  + Gender shades
  + Traffic intersections using FRT
* Introduce our experiments
* Tying out experiments to initial motivation
* Primary application: privacy and security? in X settings?

**Introduction**

* What is facial recognition technology (FRT)?
  + What it is and why its growing in popularity
* Customer service as the application setting – Explain how FRT is being used for:
  + Security and surveillance: loss prevention, shoplifting detection
  + Customer identification: loyalty programs, personalized service

**Technical Background**

* How FRT works
  + Detection → encoding → matching
* Use cases in security (loss prevention) and customer experience (loyalty programs)

**Original Experiments(Literature review):**

**Gender Shades:**

Algorithmic biases can manifest across many attributes, but they are particularly pronounced in the realm of race. This raises serious concerns about the ethical implications of FRT, and how consumers may be affected. One of the most influential studies to highlight these disparities is the *Gender Shades* Study by Buolamwini and Gebru (2018). Researchers from MIT and Stanford University conducted a study evaluating three leading gender classification systems – Microsoft, IBM, and Face++ – using the Pilot Parliaments Benchmark, a dataset intentionally balanced across gender and skin tone. In their results, they found that the three program’s error rates had racial bias concerns: while light-skinned men never had an error rate higher than 0.8%, dark-skinned females had error rates as high as 34.7% in one case. This huge difference in accuracy not only points to systemic bias, but also the intersectionality of race and gender and resulting algorithmic unfairness. The researchers attributed these disparities to the unbalanced training data these algorithms had been trained on, which overwhelmingly included and favored light-skinned individuals. Thus when deploying FRT in high-stakes environments, such as customer service, it is essential to address the misidentification that could lead to unfair surveillance, exclusion, and more consequences.

Algorithmic biases in FRT can also be rooted in the datasets used to train them. Researchers at a major US technology company claimed an accuracy rate of more than 97% for a race-recognition system they designed, using the Labeled Faces in the Wild(LFW) dataset to evaluate its performance. However, while LFW has been considered the gold standard benchmark for face recognition, it was estimated to be considerably unbalanced, composed of 77.5% male and 83.5% white subjects. The skewed datasets used to train FRT models are a common pitfall in the development of this technology – high underrepresentation of women, POC, and other marginalized groups creates potential for misidentification, error, and systematic failures of these models. These disparities are not just technical flaws, but ethical concerns, especially when FRT is deployed in high-stakes settings.

**Paper that goes against gender shades**

* Use this as a transition – fact that models today should be able to be more accurate. So are they? → our experiment

**Algorithmic bias:**

* Concept: the systematic and repeatable errors in computer systems that create unfair outcomes, particularly in marginalized groups
* Examples:
  + Gender bias in hiring algorithms
  + Age bias in credit lending
  + Socioeconomic bias in insurance risk assessments
  + **Facial recognition:** mention biases in accuracy, lighting sensitivity, and demographic disparities
* Transition to racial bias
  + (explain) While algorithmic biases can occur along many attributes, **racial bias in facial recognition** has particularly sever implications for privacy, safety and civil rights

Model overview:

* OpenCV (Haar Cascades): Face detection, common in low-cost surveillance
* DeepFace: Race classification and face verification, useful for membership verification
* Face\_recognition(dlib-based): Face encoding and matching for customer access control

**Application setting motivation**

* Customer service & security – how racial bias in FRT affects real-world settings like:
  + Retail stores: higher rates of false positives for theft among certain demographics
  + Airports and stadiums: denial of entry or heightened surveillance based on facial analysis

**Real-World implications**

While the original experiments and our results illuminate the algorithmic biases and disparities in FRT, real-world incidents showcase the tangible consequences of these biases.

In law enforcement, facial recognition misidentifications have led to wrongful arrests of Black individuals. For instance, in January 2020, Robert Williams, a Black man from Detroit, was wrongfully arrested after an FRT system misidentified him as a shoplifting suspect[[1]](#footnote-0). Despite clear discrepancies between Williams and the suspect in surveillance footage, human oversight could not make up for the algorithm’s error; Williams spent 30 hours in jail and posted a $1,000 bond before this grave mistake was corrected. The heavy reliance on technology in law enforcement often leads to uncritical acceptance of its results, even when those results are flawed. When facial recognition systems are racially biased, their deployment in high-stakes contexts like policing is especially dangerous. Algorithmic errors can lead to real-world harm, with human oversight often failing to catch these mistakes.

The risks associated with algorithmic biases in FRT don’t necessarily stop at misidentifications; they can be compounded by corporate policies that use these technologies for selective enforcement. One example is Madison Square Garden’s controversial use of FRT in 2022 to enforce an “exclusion list” that bans lawyers whose firms are engaged in litigation against the company[[2]](#footnote-1). Lawyers were denied entry from attending events at venues of MSG Entertainment, including Radio City Music Hall and MSG itself. Lawyer Kelly Conlon was denied entry to MSG after being identified through facial recognition software. Although she had no direct involvement in litigation against the company, just her association with her firm was enough to trigger the ban. This case exemplifies how with any algorithmic biases in FRT, corporate policies and exclusionary practices can magnify them. This sets the precedent that FRT is not unbiasedly used for security, identification, etc, but instead can also be leveraged as a selective and controlling technology. In customer service settings particularly, the consequences could be very detrimental. If businesses adopt similar policies of exclusion, algorithmic biases could be used to justify this discriminatory practice. The corporate leveraging of FRT highlights the need for increased oversight to prevent the deepening of biases in these models.

* Police could use FRT to misidentify their race
  + Having this bias present in that technology can be dangerous

**Our Data Analysis Insights**

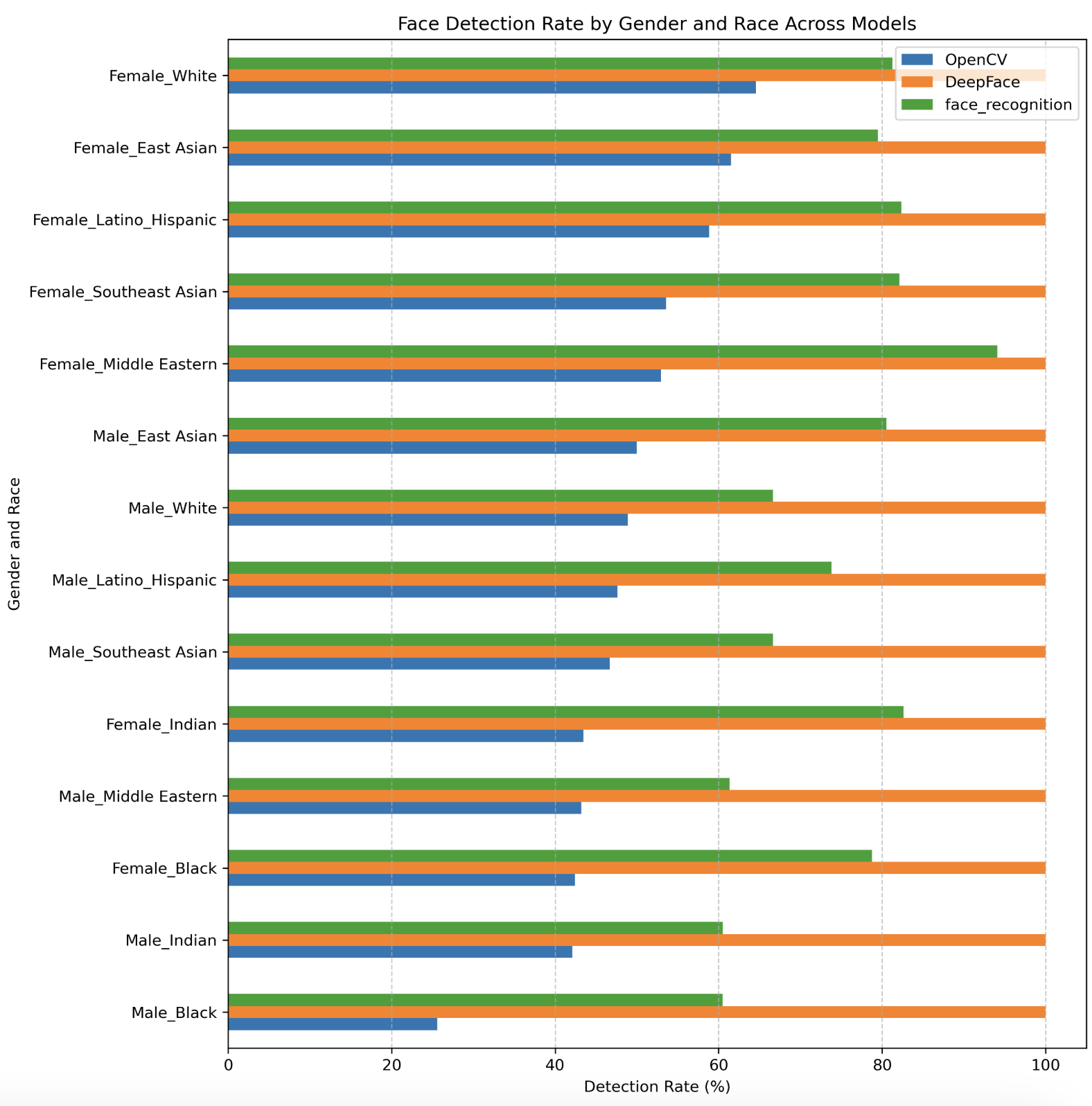
Our analysis compares the facial recognition capabilities of several open-source software: OpenCV, the face\_recognition python library, and DeepFace. [ How do they work/ what are the differences between them]. We used random samples of the FairFace dataset to evaluate each model’s detection rate of different races. The results are presented in the tables below:

Model overview:

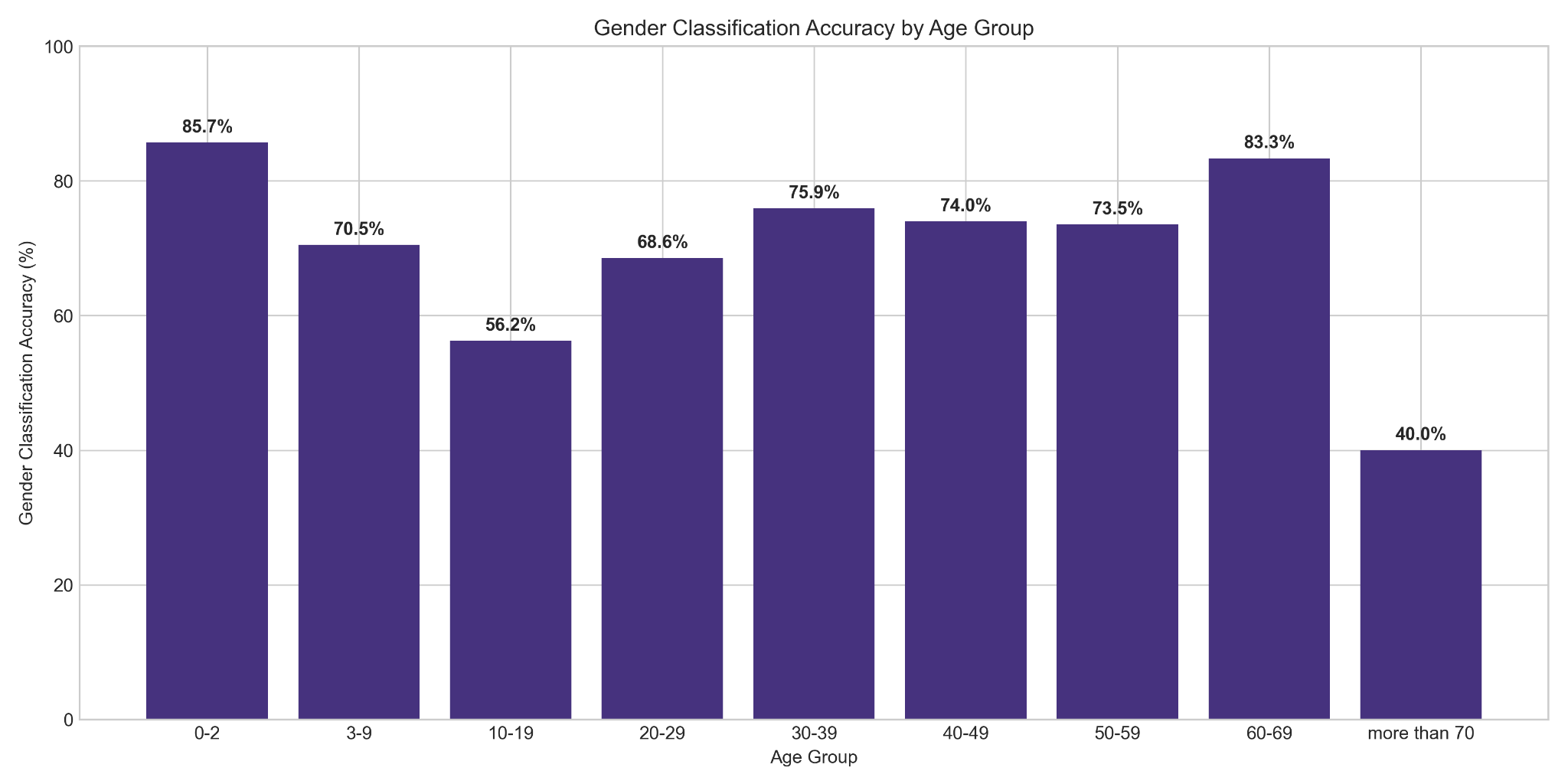
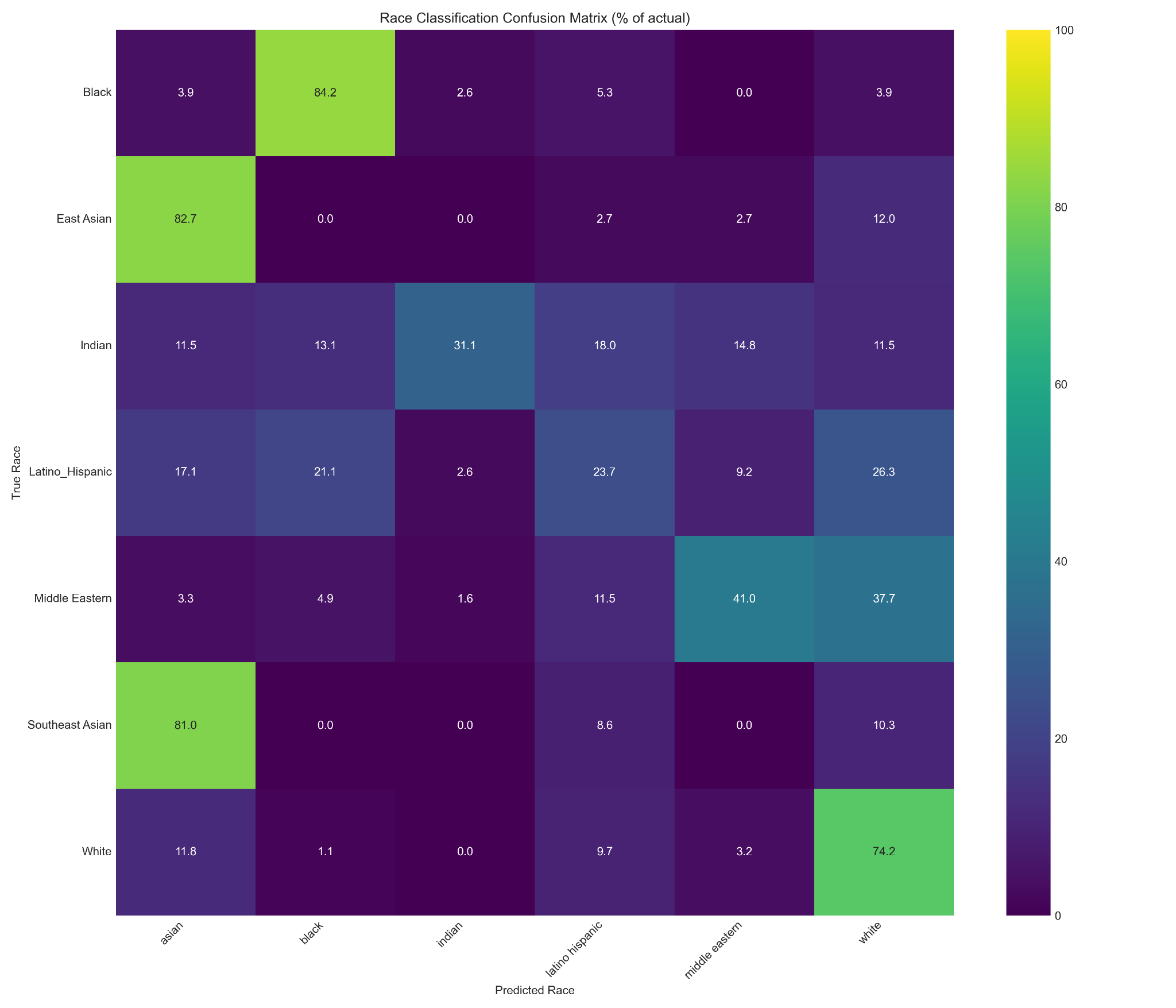
* OpenCV (Haar Cascades): Face detection, common in low-cost surveillance
* DeepFace: Race classification and face verification, useful for membership verification
* Face\_recognition(dlib-based): Face encoding and matching for customer access control
* The models show different performance on different races.
* More systematic fairness metric
* Link bias to real life

We started with 3 models. DeepFace quickly emerged as the best result for our experiment

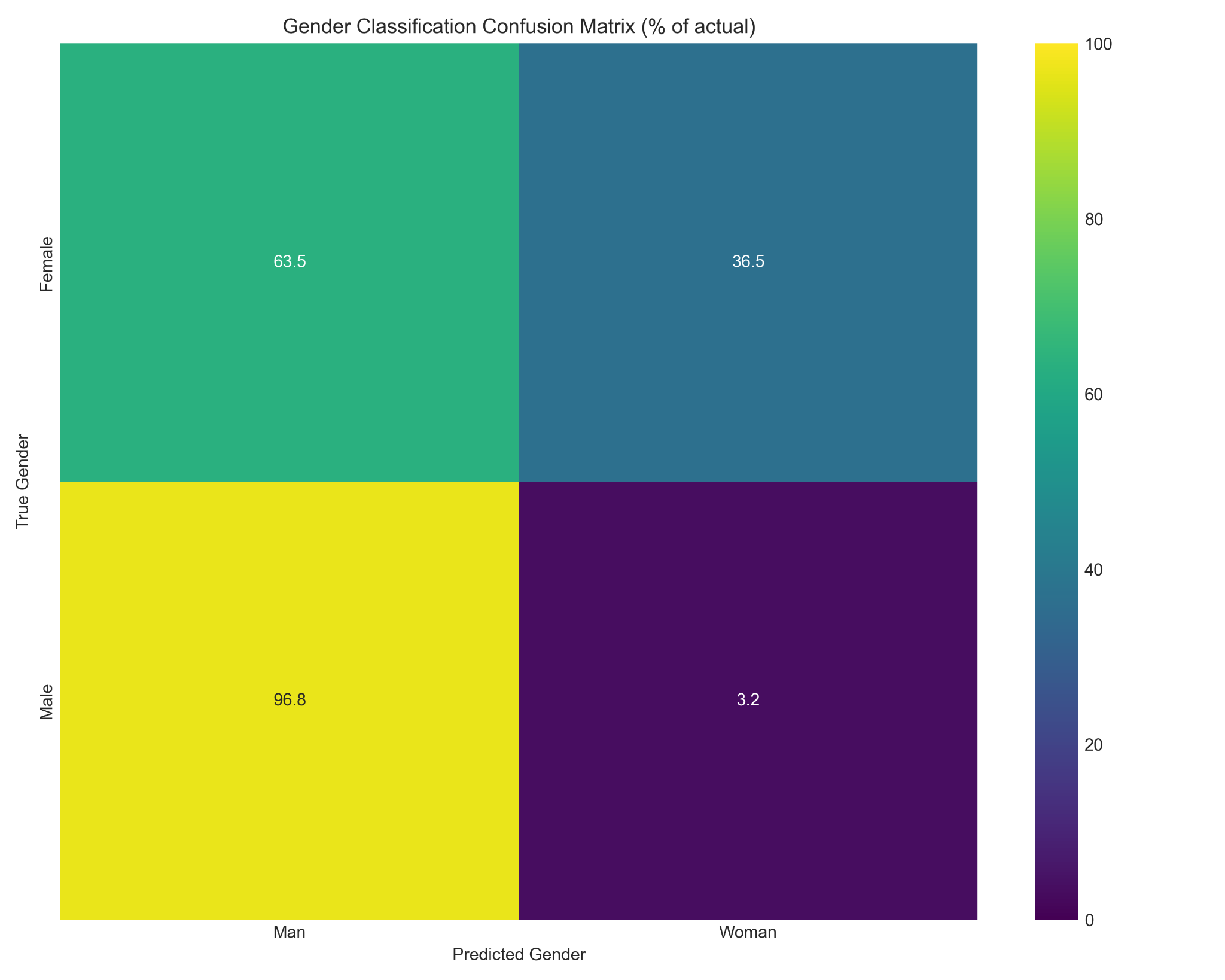
* Commentary on the other 2 models



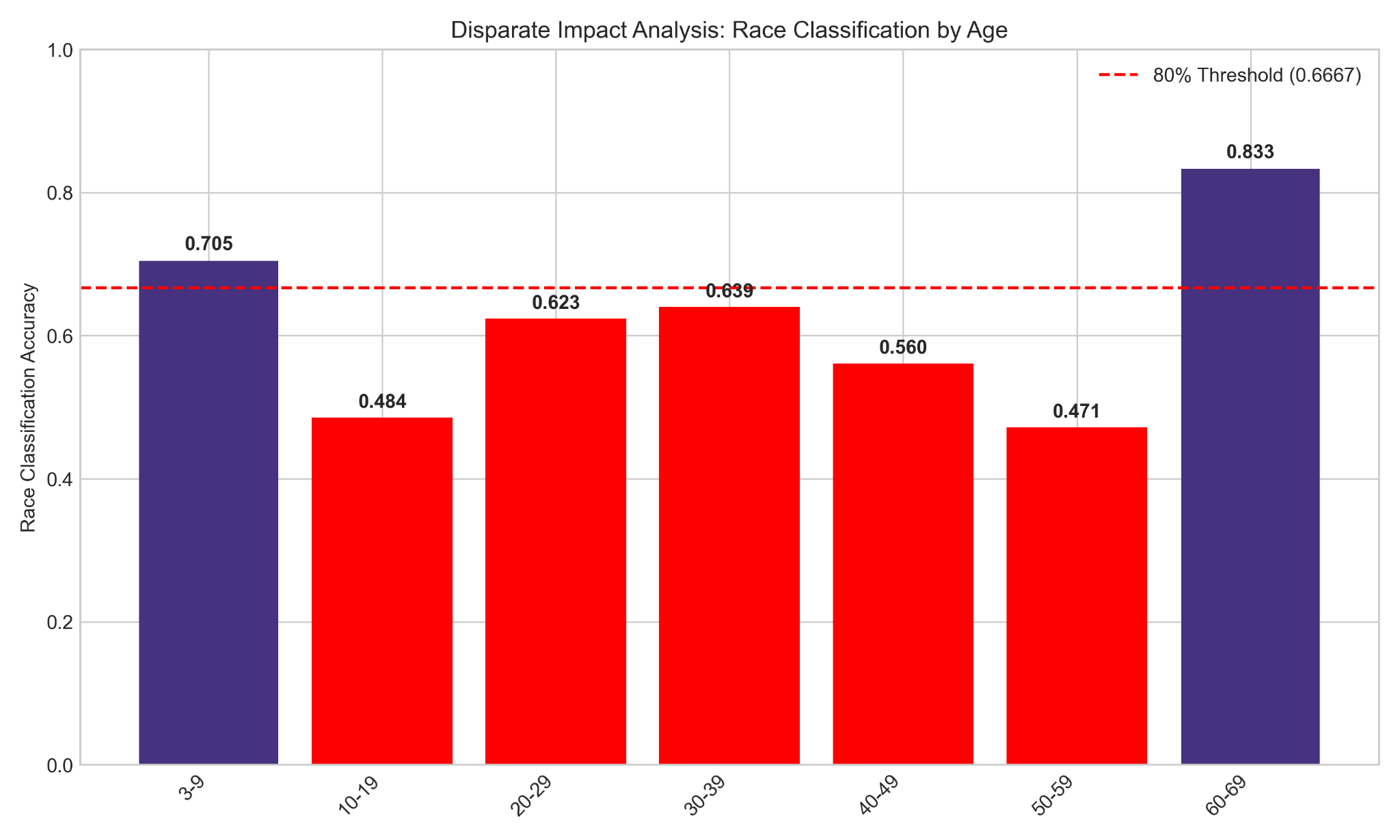
**Model Results**

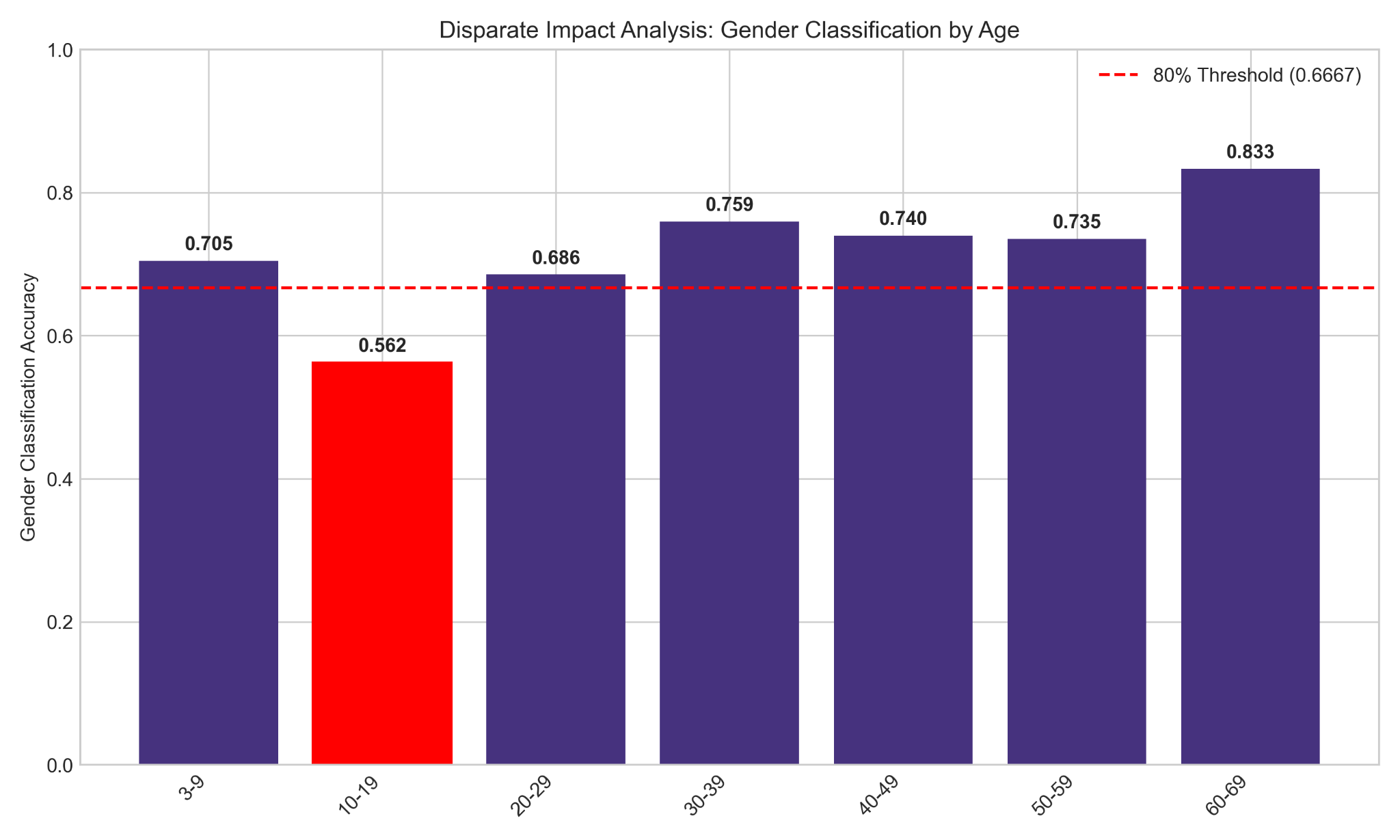
* Looked at gender, race and age
* Measured disparate impact
* Motivation
  + Detecting racial bias still existing in these models
  + Fairface is still more balanced than existing datasets
    - Benchmarks mentioned
* 
* The age-stratified gender classification data reveals notably degraded performance for cohorts aged 10-19 and 20-29. While this observation likely reflects biological realities, such as the less pronounced sexuality during adolescence and early adulthood, it does raise red flags about long term model quality. The performance degradation in younger cohorts may signal a fundamental limitation in the model's ability to adapt to evolving gender expressions and aesthetic presentations characteristic of younger generations. This raises a critical concern: as fashion trends and cultural norms surrounding gender presentation continue to evolve, will current classification systems trained on historical data maintain its relevance? The convergence of biological factors with rapidly shifting sociocultural expressions of gender among younger populations suggests that these models may face accelerating performance degradation over time, as the stylistic and presentational markers they rely upon become increasingly disconnected from contemporary gender expressions. This temporal dimension of algorithmic bias represents an understudied vulnerability in classification systems, wherein models optimized for current or historical populations may systematically fail to accommodate the dynamic nature of human self-presentation across generational cohorts.
* 

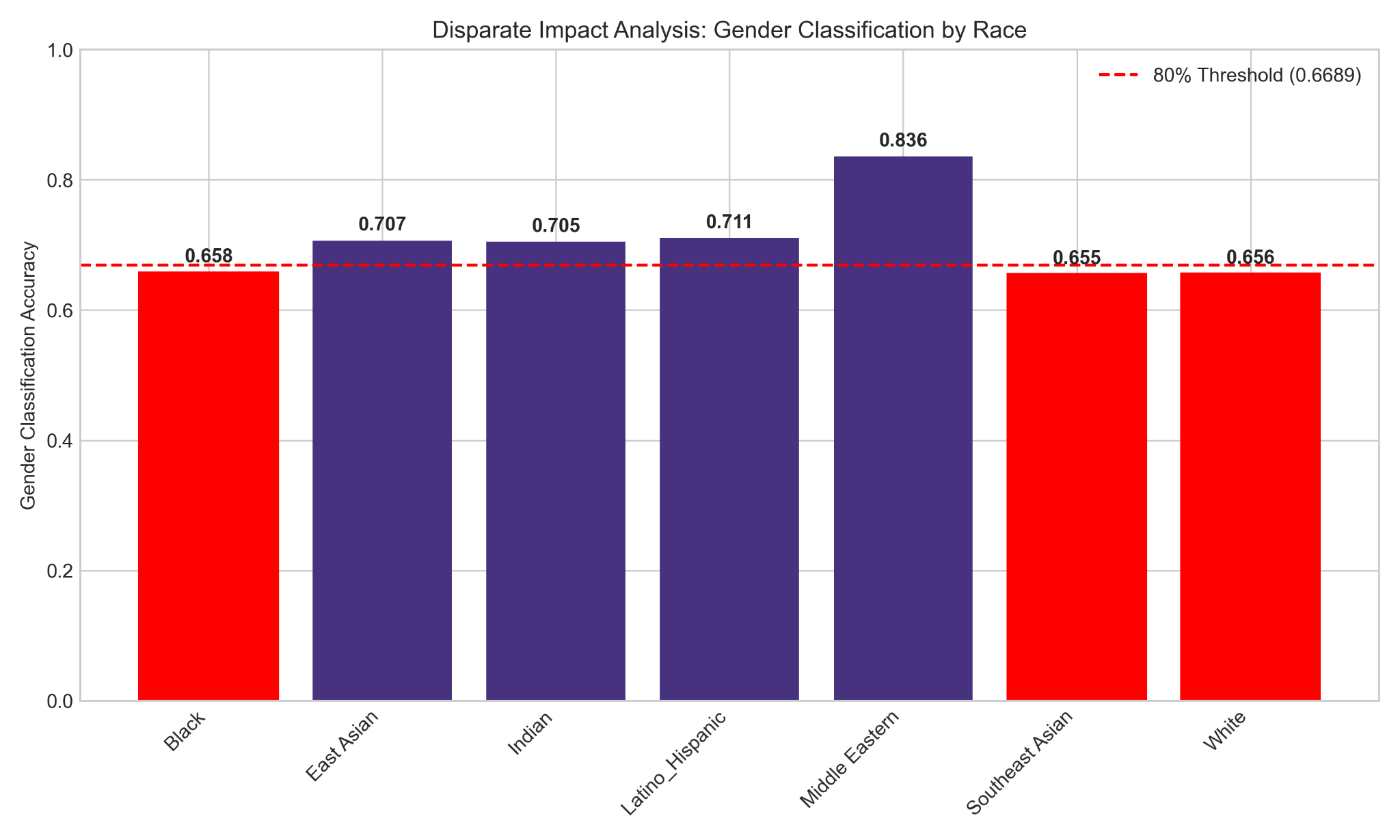
The classification performance matrix reveals a stark racial hierarchy in algorithmic accuracy, with rows containing green boxes indicating achievement of the 70% correct classification threshold. Notably, the system demonstrates robust performance for Black, East/Southeast Asian, and White racial categories, consistently meeting or exceeding acceptable accuracy standards. In contrast, Indian, Latino, and Middle Eastern populations experience significantly degraded classification performance, falling well below the 70% threshold. This differential performance has profound implications for data collection and analysis, as systems relying on these classifications may systematically gather skewed data for underperforming racial categories or, more troublingly, may fail to capture data from these populations entirely. Further compounding these disparities, the analysis reveals a specific pattern of misclassification wherein Latino and Middle Eastern individuals are erroneously classified as White at disproportionately high rates, suggesting not merely random error but systematic bias toward dominant racial categories that effectively erases the distinct identities of these populations from algorithmic datasets.

* 

The gender classification analysis reveals a troubling asymmetry in algorithmic performance, demonstrating poor classification accuracy when the ground truth gender is female while achieving notably superior performance for male subjects. This systematic disparity raises critical questions about the potential for AI systems to reinforce and perpetuate existing gender stereotypes through their operational biases. The implications of this problem extends beyond mere performance; females whose appearances are closer to males will not have their data collected in the female bucket. Thus, systems’ data on females will only reflect gender-typical appearing people.







* Issues for most ages in race classification but particularly 10-19 and 50-59
* Issues for ages 10-19 in gender classification
  + Reflects issue brought up earlier
* Certain Races have disparate impact issues when assigning gender
  + Cross-sectional bias
* Overall: many algorithmic biases found that affect certain racial groups disproportionately

**Pitfalls and Limitations**

* Model doesn’t classify between southeast asian and asian
* For age there aren’t proportionate samples in each age group
* Balanced dataset - FairFace is more balanced than past benchmarks, but still any imbalance points towards less validity
* Had originally planned to do 3 different FRTs, but they do different things
  + All main results from deepface but OpenCV still had bias

What are the limitations of your work?

* Many other ethical concerns with facial recognition technology
* Algorithmic biases past racial biases
* Didn’t do fine tuning

How could the limitations be addressed through a future iteration of your work.

* Fine tuning analysis could have been performed if …
* As more advanced technologies (that may be currently used in private sectors) are released and built up by the open source community, the project could provide even more insight into FRT systems currently in use.
  + Systems that are more relevant to actual use cases.

1. https://mit-serc.pubpub.org/pub/bias-in-machine/release/1 [↑](#footnote-ref-0)
2. https://www.nytimes.com/2022/12/22/nyregion/madison-square-garden-facial-recognition.html [↑](#footnote-ref-1)