

GENDER AND RACIAL DISPARITIES IN REKOGNITION

Angelina Kiman, Faculty Advisor Dr. Ryan Labrie, Winter 2021

Seattle Pacific University

1. INTRODUCTION

Algorithmic bias is a challenge in the AI (Artificial Intelligence) field. The rapid growth of computer vision has raised ethical concerns to society. A recent study shows that gender and racial bias are embedded in computer vision machine learning especially among people of color (Buolamwini & Gebru, 2018; Grother, 2019). Criado-Perez (2019) mentions in her book, *Invisible Women*, that sex-disaggregated data occurs when the data is collected and presented separately on women and men. The presence of sex-disaggregated data has resulted in an unequal decision-making process which impacts many people's lives. It is crucial to bridge the gender and racial data gap by gathering all the data that represents reality.

Tech giants such as Amazon, Microsoft, and IBM have been selling facial recognition software to law enforcement. Vagueness in AI law and ethics have put human autonomy at risk. For example, there were several cases of false imprisonment due to a flawed facial recognition system. Crockford (2020) emphasizes the importance of protecting civil liberties and expanding the United States' First and Fourth Amendment rights in the 21st century.

This project aims to investigate gender and racial disparities in Amazon's Rekognition service. Using the [FairFace dataset](#) by Karkkainen & Joo (2021), we (1) measure the model fairness based on precision and false negatives, (2) determine the model challenges with the dataset, and (3) identify key takeaways from commercial facial recognition software.

2. REKOGNITION

Amazon's Rekognition is a cloud-based computer vision software as a service (SaaS) that was launched at the end of 2016. The service offers image and video classification such as detection and recognition of objects, people, text, scene activities, and other custom categories. The core underlying machine learning technology of Rekognition is deep learning. The term deep learning refers to a subset of machine learning algorithms that use multiple layers of connected nodes (some equate this style of learning to that of the human brain; hence the term "neural networks"). In this paper, we will be using deep learning models provided by Rekognition to classify between gender and race and discover racial disparities in the models.

Boulamwini & Gebru (2018) evaluated three major facial recognition platforms including Microsoft, Face++, and IBM. Figure 1 shows that the model performs poorly on darker female in all three platforms. Microsoft and IBM perform best on lighter male, while Face++ performs best on darker male.






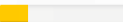





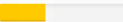





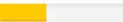
Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
 Microsoft	94.0% 	79.2% 	100% 	98.3% 	20.8% 
 FACE++	99.3% 	65.5% 	99.2% 	94.0% 	33.8% 
 IBM	88.0% 	65.3% 	99.7% 	92.9% 	34.4% 

Figure 1. Gender Classifier Model Performance (Boulamwini & Gebru, 2018)

Using Boulamwini & Gebru's studies as a comparison, we are investigating the gender and racial disparities in Rekognition as well as addressing the challenges that the model encounter.

3. FAIRFACE DATASET

The FairFace dataset contains 97,698 images. Figure 2 shows example images from the FairFace dataset. The dataset contains a wide range of genders, races, and ages photographed from different angles and lighting conditions. The images are a mix of both low and high resolution. Ground truth labels are provided for all images.

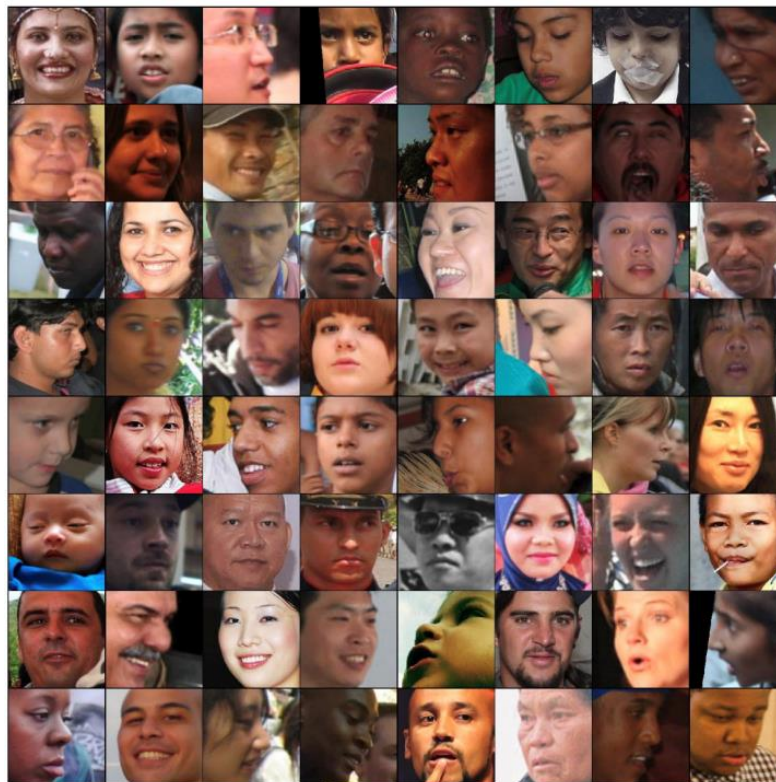


Figure 2. FairFace Dataset

Figure 3 shows the proportion of each gender in the dataset. The figure shows that there are 6% more male images than female in the dataset.

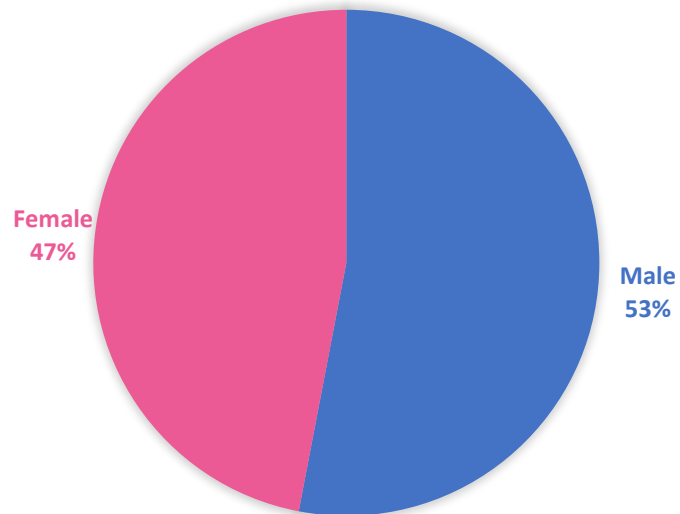


Figure 3. Gender

Figure 4 shows the breakdown of the seven races on the dataset: White, Latino (Hispanic), Indian, Black, East Asian, Southeast Asian, and Middle Eastern. From Figure 3, we can see that white faces have the most representation (19%) and Middle Eastern faces have the least representation (11%). The others share similar representation around 13-15%.

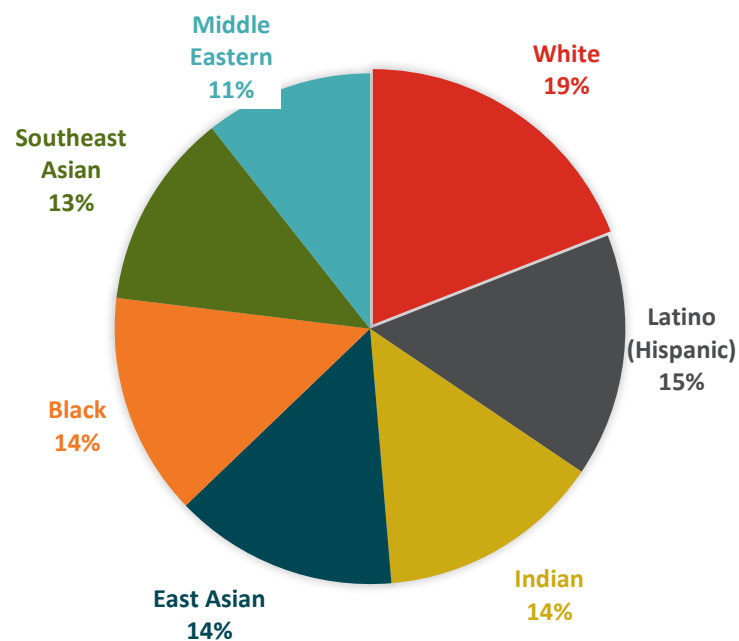


Figure 4. Race Dataset

4. EXPERIMENT SETUP

The FairFace dataset contains 86,744 training images and 10,954 testing images. We examine the model performance by training the model with different numbers of training images. For all experiments, we use the same training set of 10,954 labeled images. For training, we train nine models with an increasing number of training images for each model from 10,000 originally to almost 86,744 for the final model. Rekognition does not reveal the underlying deep learning model that was used for training and testing.

There are three metrics that the model uses to evaluate the test results. True positive is when the predicted label matches the ground truth label. False positive is when the predicted label does not match the ground truth label. False negative is when the ground truth label was not predicted by the model.

We use precision, recall, and F1-Score to measure the model performance. Precision is a metric that measures how many predictions were correct out of our total predictions. Recall is a metric that measures if all correct predictions were found (even at the expense of incorrect predictions). F1-Score is a metric to average both precision and recall. The formulas are shown below:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

5. RESULTS

The model performance shows average precision, overall recall, and F-1 score. Figure 5 shows that the number of training samples matters for the model performance. A small number of training samples with imbalanced data will have a higher risk of having *false positives* and *false negatives*. There is a 4.40% gain in average precision score from training 10,000 to 20,000 samples. When training 70,000 images, the average precision reaches its peak at 75.10%, but the overall recall drops to 77.40%. After training 80,000 images, the average precision rate starts to decline and the overall recall increases. This is known as the precision/recall trade off. Given that the overall F1 scores are above 75%, it is clear that Rekognition is able to learn how to distinguish between the races; however, we also see that the accuracy boost from more training data is flattening out. F1 scores around 75% may be good enough for many applications; however, I personally feel that for law enforcement applications, the F1 scores need to be near 100%.

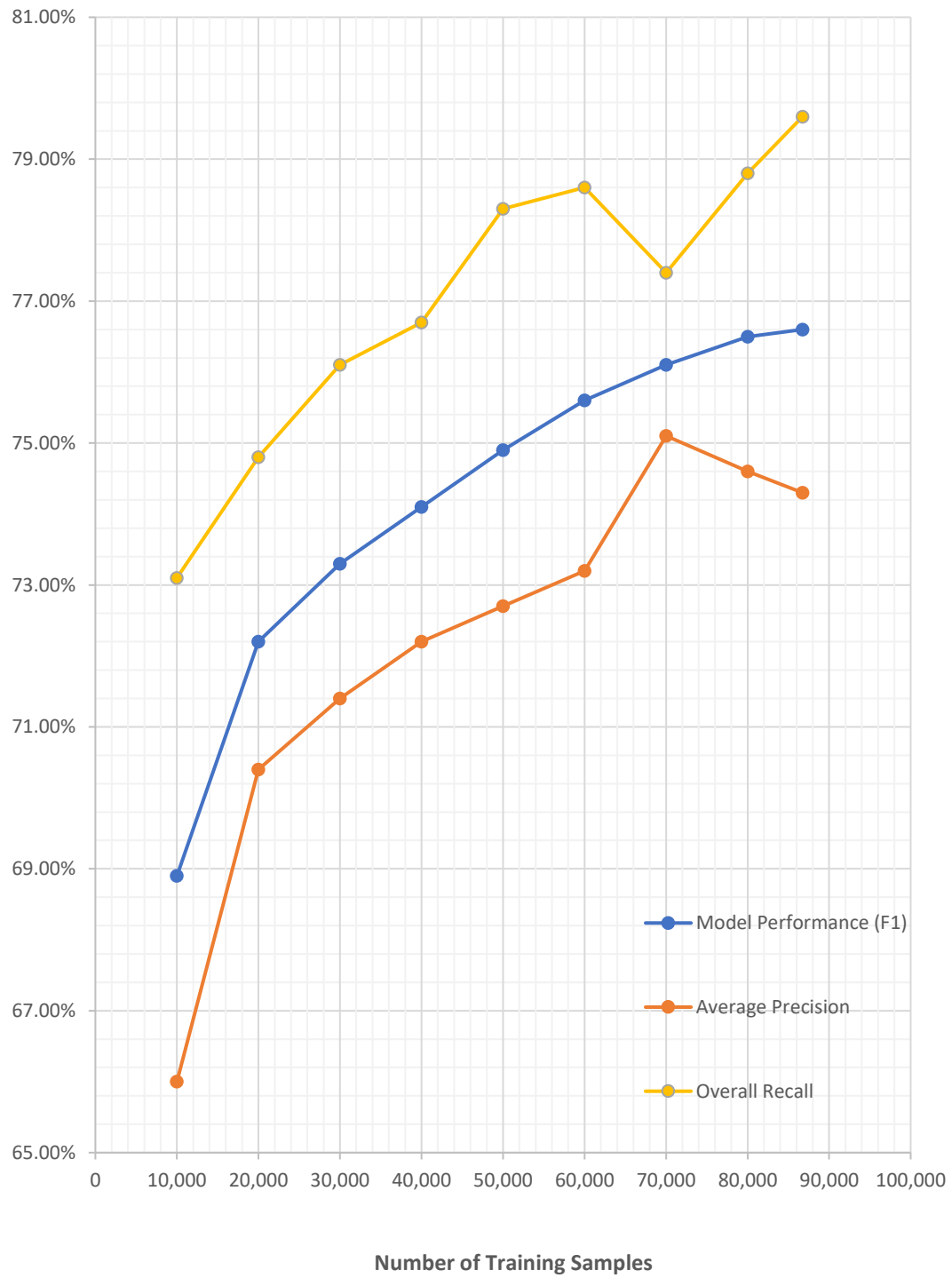


Figure 5. Rekognition Model Performance

Figure 6 shows gender classification performance. The F1-score between male and female are the same (around 93%). Female has higher recall by 2.60% compared to male. Male has higher precision rate by 3.80%.

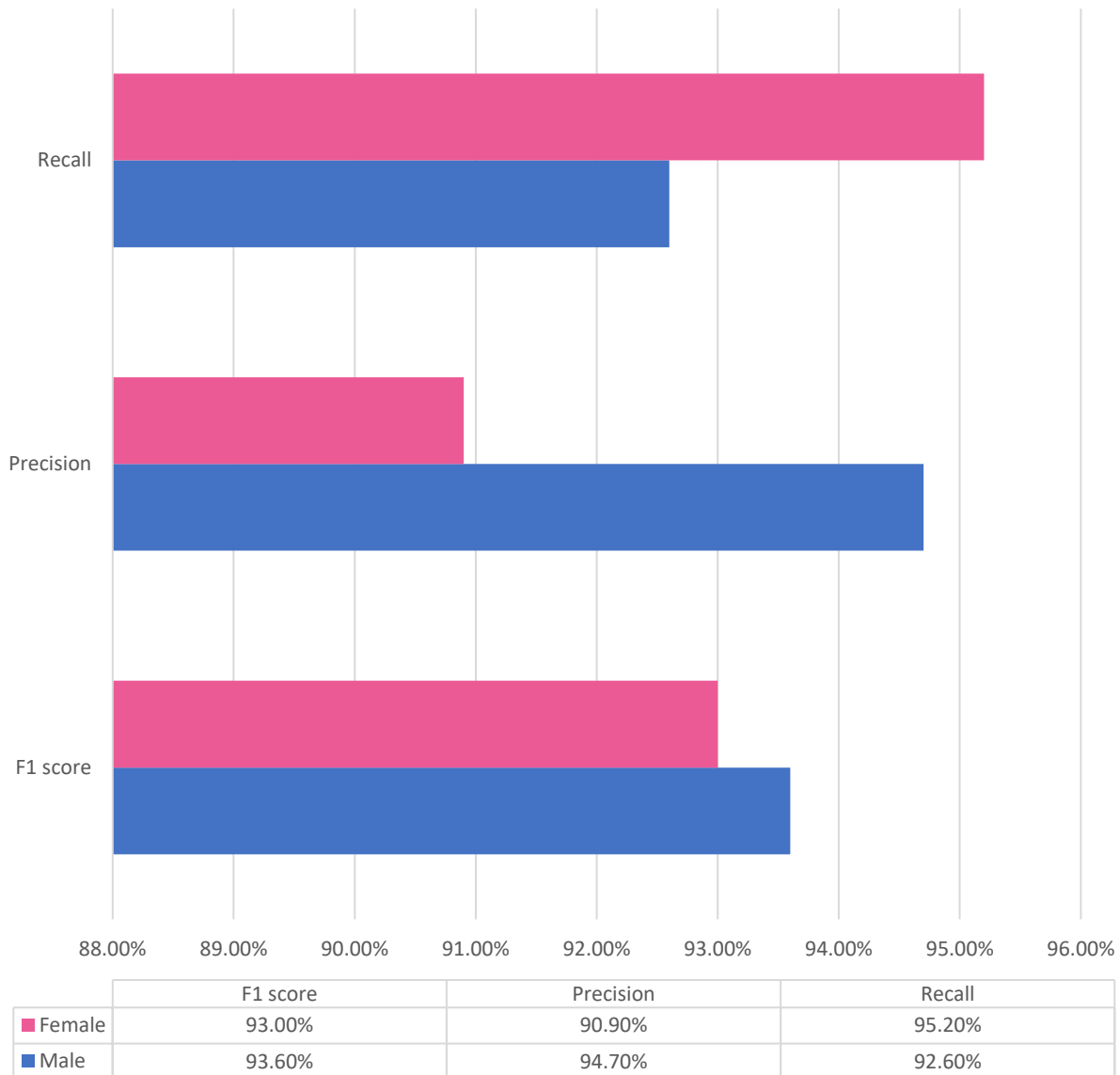


Figure 6. Gender Performance Metrics

Figure 7 shows the model performance based on race. The model performs worst on Latino (Hispanic) with a F1 score of 56.40% even though it has the second largest dataset. The model performs best on Black with a F1 score of 86.30%, while White F1 score is 10% less than the Black. Indian, East Asian, and Southeast Asian have different performance scores even though they share similar representation. This shows that the number of samples in an attribute cannot guarantee a satisfactory model performance. Disproportionate samples may exacerbate performance metrics in certain races.

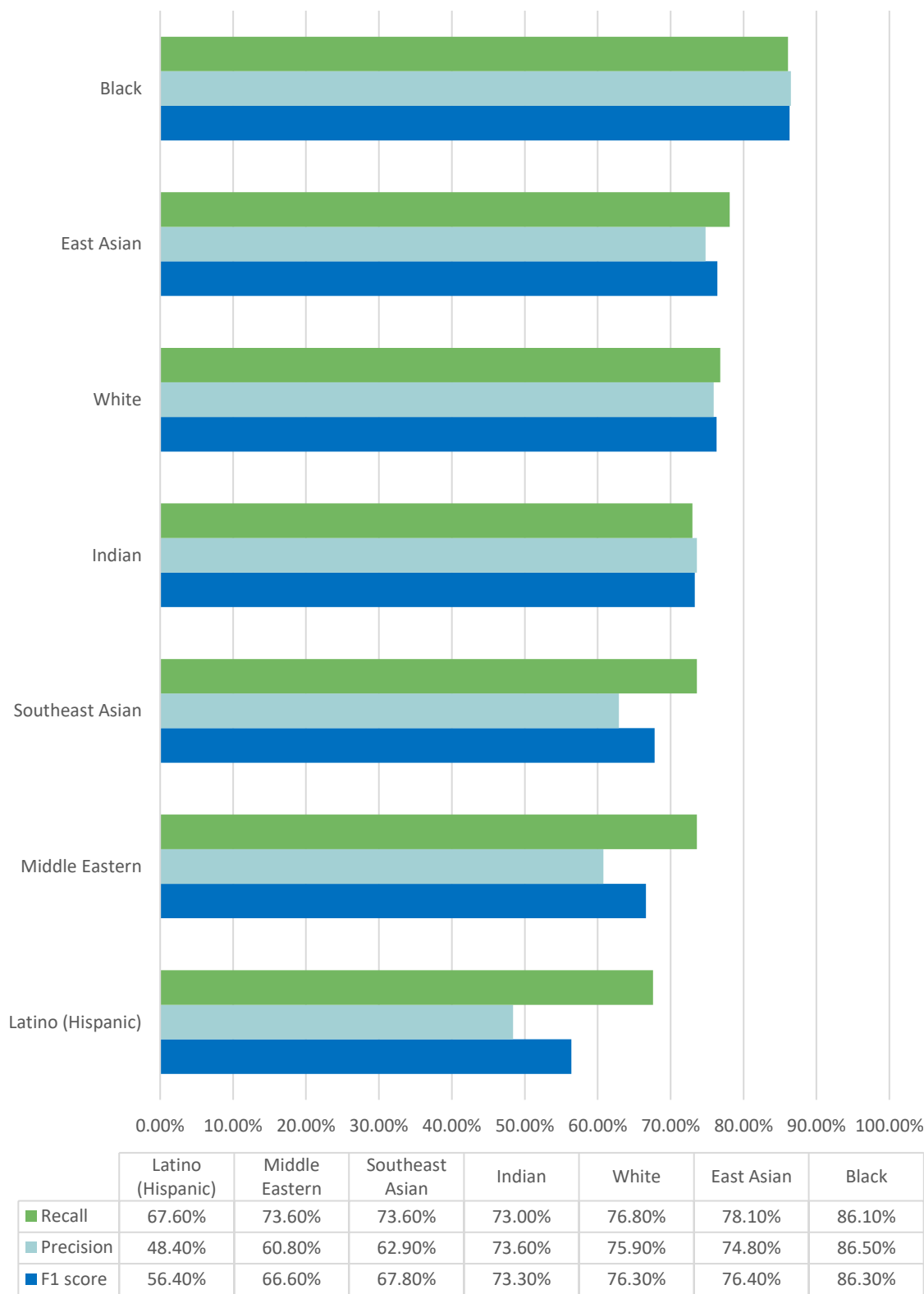


Figure 7. Race Performance Metrics

Figure 8 shows a confusion matrix for our best model across all race attributes. The rows represent the predicted values, and the columns represent the actual values. From the table, we learn that of all the races, East Asian is confused with Southeast Asian most often. The model also confuses Middle Eastern as Latino (Hispanic) and White for 13.98% and 13.65% respectively. Similarly, the model confuses Latino (Hispanic) and Middle Eastern when the actual label is the White race. The model mistakenly predicts 14.05% Latino (Hispanic), 5.67% Black, and 5.15% Middle Eastern as Indian. Black race has the best performance among the seven attributes in which 6.23% is predicted as Latino (Hispanic) and 4.18% is predicted as Indian.

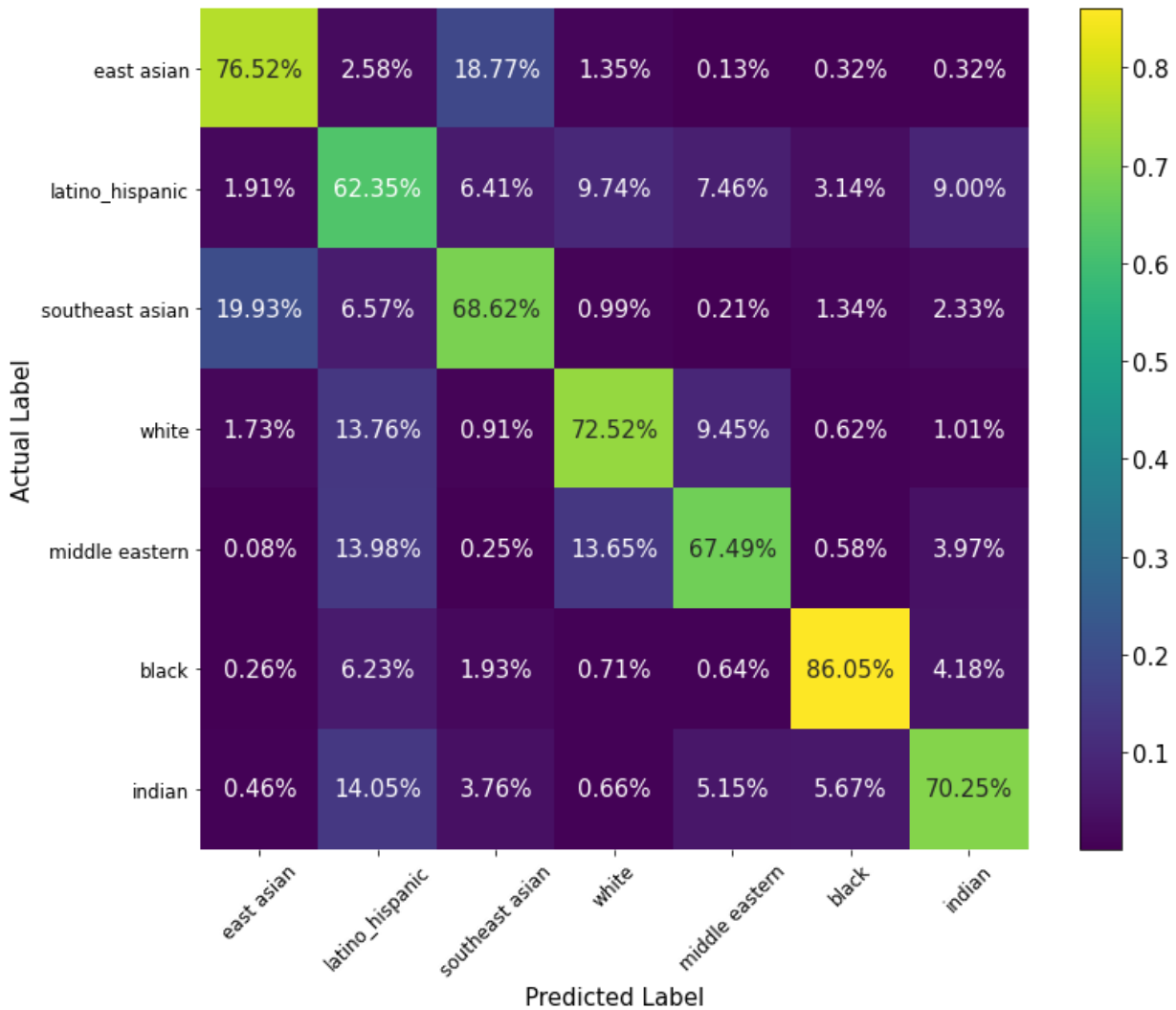


Figure 8. Race Confusion Matrix

Figure 9 shows examples of a male and a female in which the model does not predict Middle Eastern as the ground truth. The model instead predicts all genders as Latino (Hispanic) with 82.4% and 65.8% confidence.

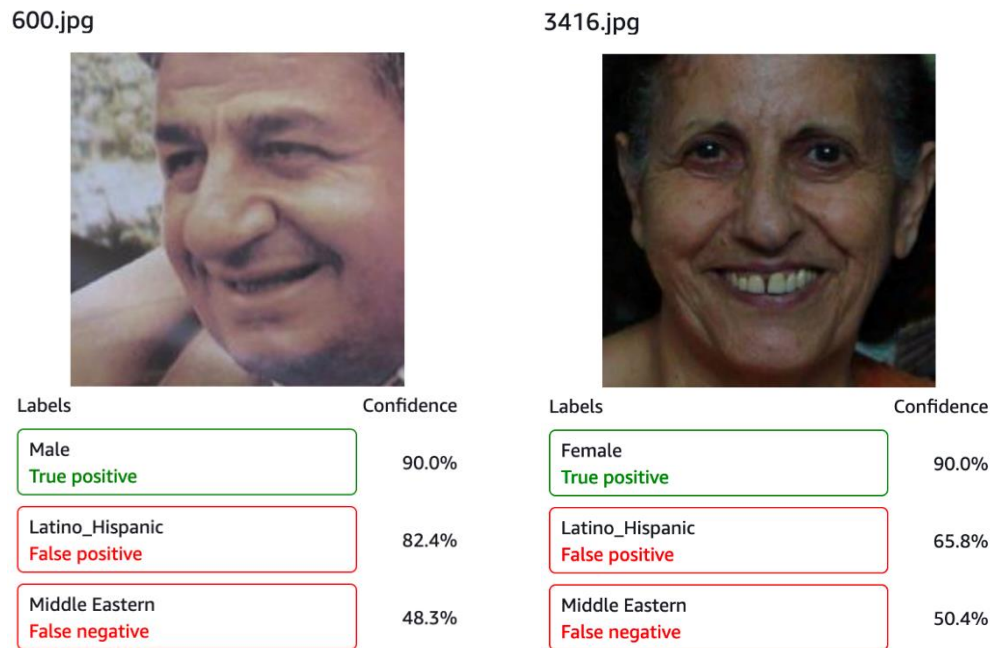


Figure 9. False positive and false negative results between Latino (Hispanic) and Middle Eastern

Figure 10 shows examples of White male and female with tan skin tones where the model predicted them as either Middle Eastern or Latino (Hispanic). There is a possibility that the model gets confused because of the lighting effects.

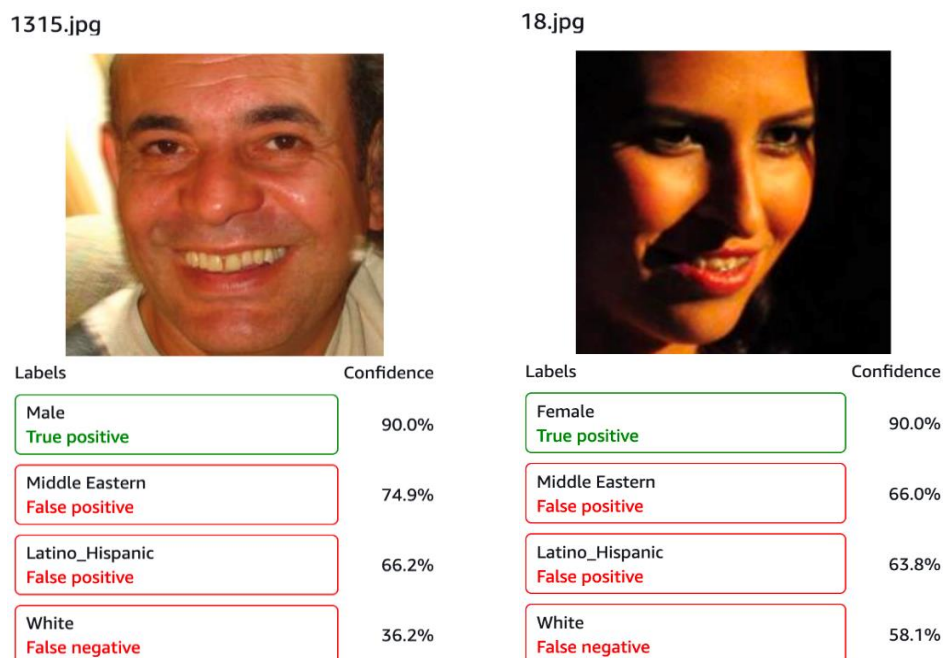


Figure 10. False positive and false negative results between Latino (Hispanic), Middle Eastern, and White

Given a situation when the model is given close-up and blurry images like in Figure 11 and 12, the model cannot differentiate between Indian and Latino (Hispanic).

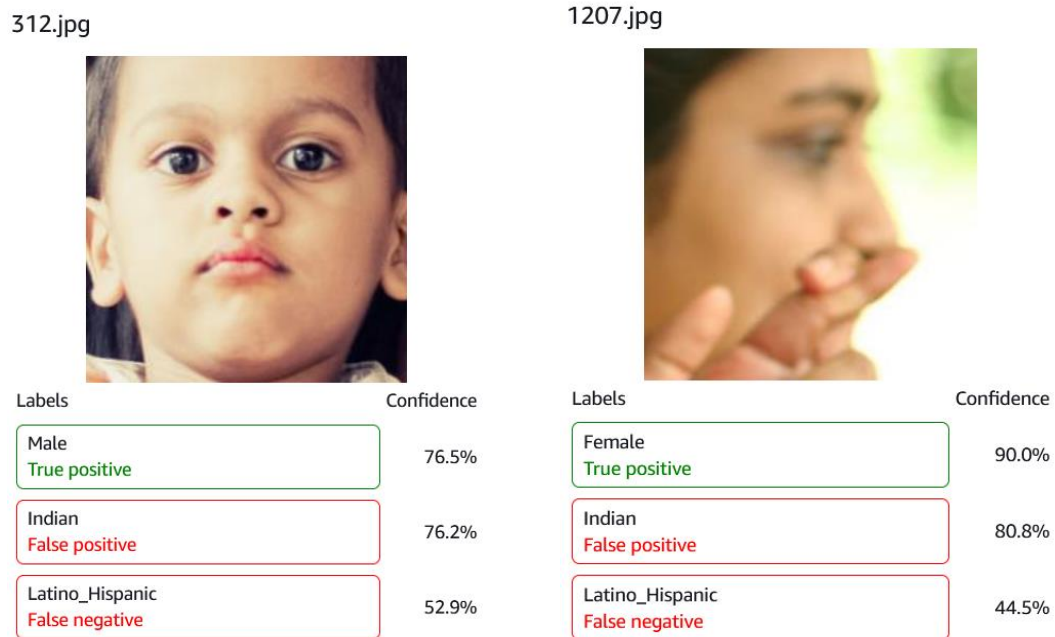


Figure 11. False positive and false negative results between Latino (Hispanic) and Indian

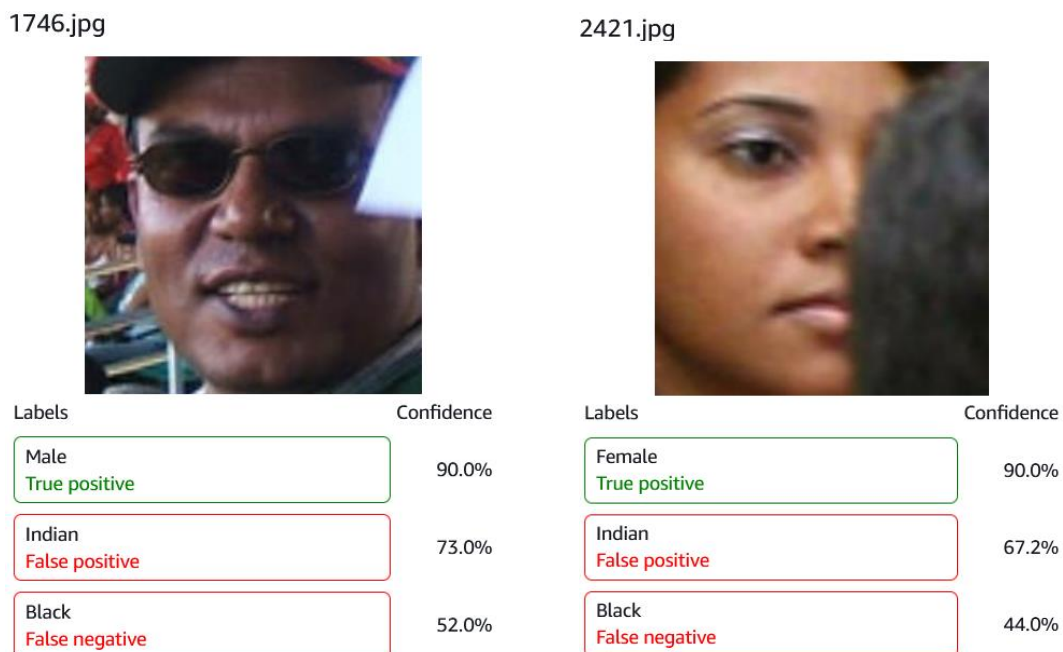


Figure 12. False positive and false negative results between Indian and Black

Figure 13 shows an example of the model struggling to choose between East Asian and Southeast Asian. The model predicts both as East Asian for 90% and 83.1% instead of Southeast Asian for 44.4% and 46.1%.

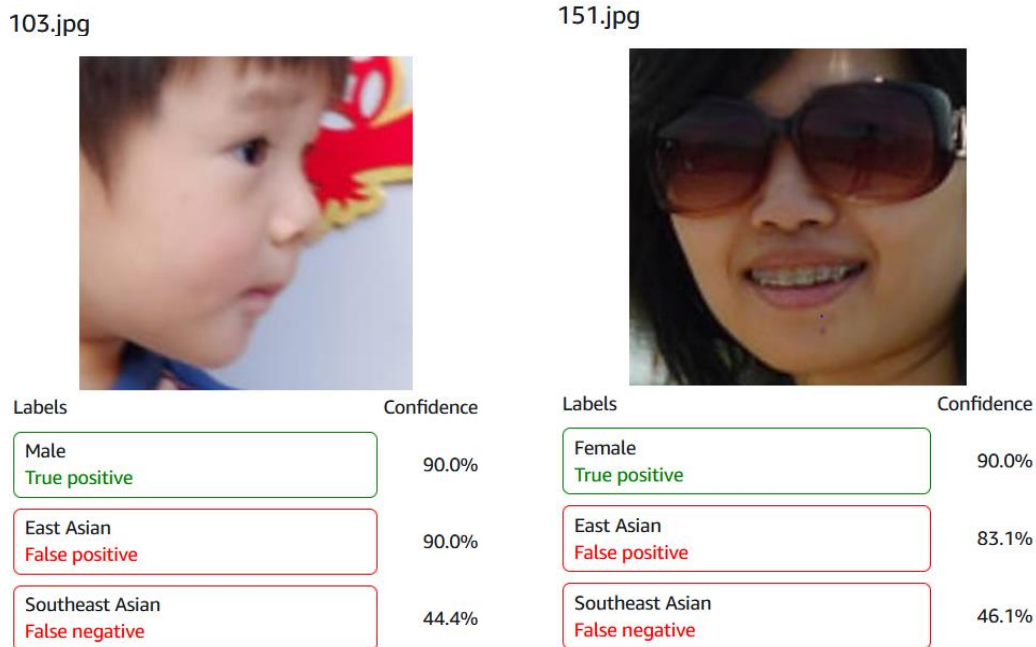


Figure 13. False positive and false negative results between Southeast Asian and East Asian

The model cannot predict any attributes correctly in Figure 14 and 15. Two possible reasons that the model gets confused are the variations in camera angles and lighting. We must bear in mind that high quality images may still not be accessible in some developing countries. Hence, image qualities may become the major obstacle in facial and image APIs (Application Programming Interface).



Figure 14. False positive and false negative results gender and between Southeast/East Asian and White

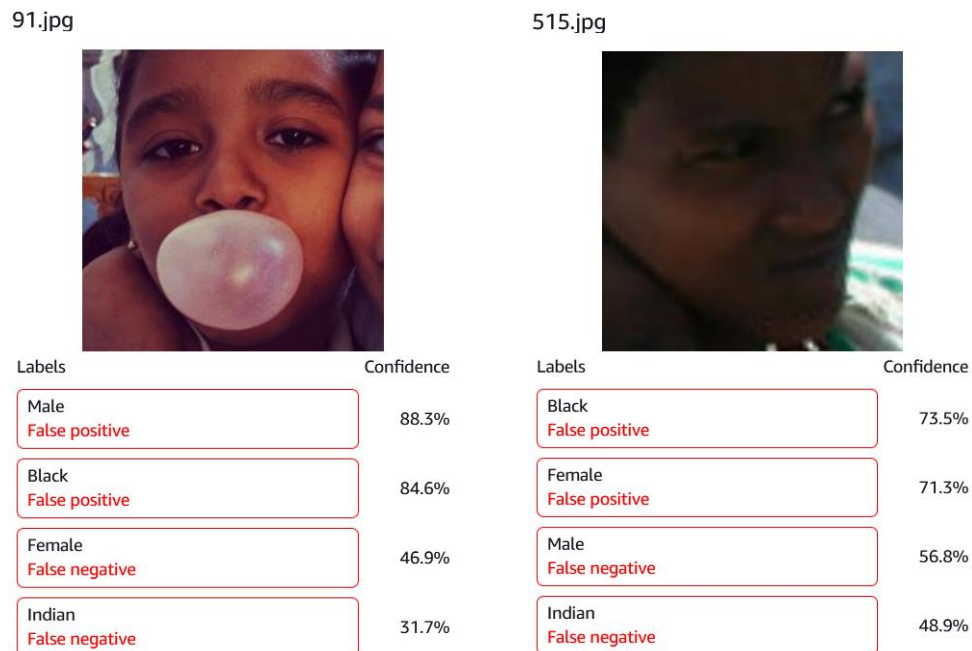


Figure 15. False positive and false negative results gender and between Black and Indian

Based on the results of my experiments, there are several interesting findings worth summarizing. First, we can conclude that the model performance (F1-Score) of male and female is about the same (around 93%). Second, the model performs the best for the Black race. Third, despite having the most training data, the model performs worst for the Latino (Hispanic) race. This is a particularly interesting and important finding because it means that we cannot solve race classification with data alone. Lastly, we must consider possible reasons why the model does not perform perfectly. Based on my observations of the experiments, I think that model performance was affected by many of the images being blurry, close-ups, having varied lighting, and/or being taken from different angles.

6. CONCLUSION

Racial disparity is a critical issue for facial recognition software. The FairFace dataset contains a disproportionate amount of gender and race images. The images are taken from different angles, lightings, and from a mixture of low-resolution and high-resolution qualities. The gender shades study by Boulamwini & Gebru (2018) has brought social awareness that gender and racial biases are embedded in commercial recognition applications. Although their study did not encompass Rekognition, this project uses Boulamwini & Gebru's study as a guideline to measure fairness based on the prediction results. Rekognition's model performance shows a significant gain of 4.4% gain from training 10,000 to 20,000 images and the accuracy boost is flattening out from more training data. The model performs worst on Latino (Hispanic) even though they have the second largest dataset while the Black race achieves the highest model performance. We can conclude attributes who have the most images do not guarantee a higher model performance score. The model often gets confused between the East Asian and Southeast Asian race, Indian/Black/Latino

(Hispanic), and White/Middle Eastern/Latino (Hispanic). The model prediction gets influenced by the lighting effects and cannot predict images correctly on blurry, low-resolution, side angle, tilted angle, and extreme close-up angle. Given the model performance is between 65% and 80%, it is clear that Rekognition can differentiate between the seven races. However, if facial recognition system wants to be used in law enforcement, I feel the model needs to have nearly 100% accuracy.

REFERENCES

- Buolamwini, J. & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, in PMLR 81:77-91.
- Criado-Perez, C. (2019). *Invisible Women: Data Bias in a World Designed for Men*. Abrams Press.
- Crockford, K. (2019 November). What you need to know about face surveillance. Retrieved from https://www.ted.com/talks/kade_crockford_what_you_need_to_know_about_face_surveillance/up-next#t-226675
- Drew, H. (2018, October 23). Amazon met with ICE officials over facial-recognition system that could identify immigrants. Retrieved from <https://www.washingtonpost.com/technology/2018/10/23/amazon-met-with-ice-officials-over-facial-recognition-system-that-could-identify-immigrants/>
- Drew, H. (2019, April 30). Oregon became a testing ground for Amazon's facial-recognition policing. But what if Rekognition gets it wrong? Retrieved from <https://www.washingtonpost.com/technology/2019/04/30/amazons-facial-recognition-technology-is-supercharging-local-police/>
- Karkkainen, K., & Joo, J. (2021). FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 1548-1558).