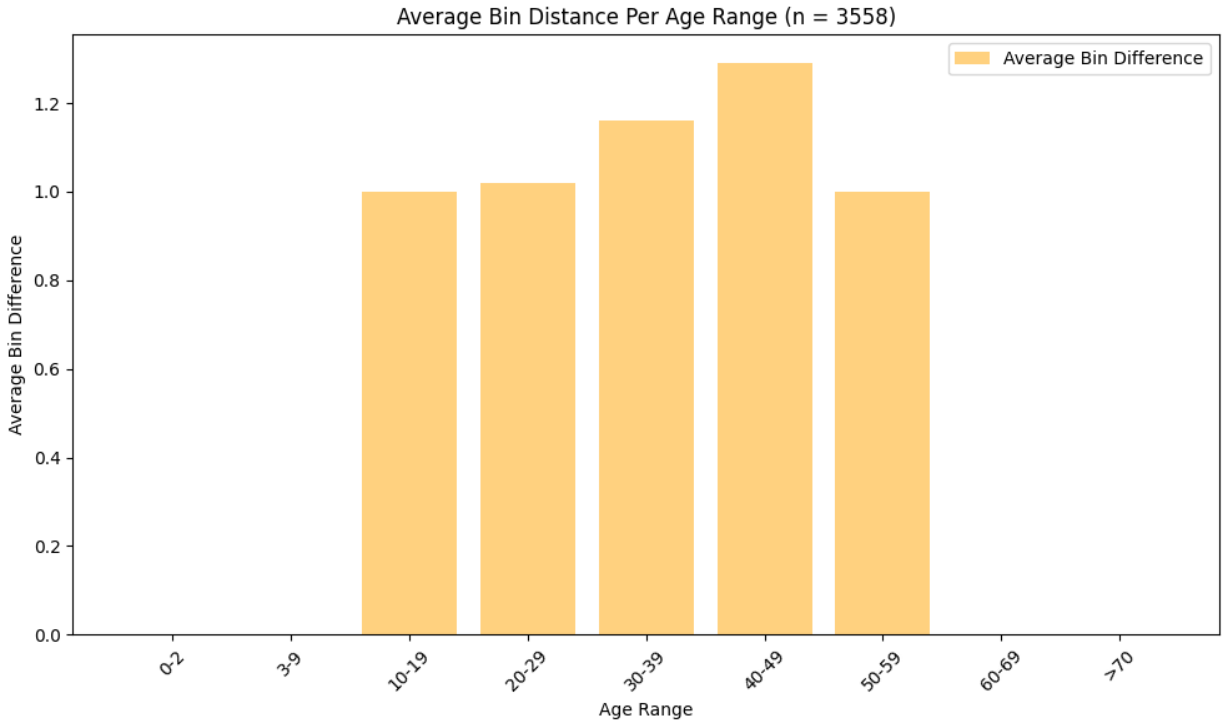
Andrew Hu

COSI-159 Assignment 4

1. Large public face datasets are heavily biased towards Caucasian faces and underrepresent other races, particularly darker faces. Most models are based on these datasets, so any systems involving these models will perform worse with non-Caucasian faces. This has negative implications for any field involving computer vision, but is particularly harmful in security and medicine, where recognizing a face or a feature has vital importance.
2. Model and cross-dataset performance were measured on a variety of datasets, testing the new FairFace against UTKFace, LFWA+, and CelebA. CelebA was only used for gender classification since it does not have a parameter for race. Each dataset was trained on a ResNet-34 model with ADAM optimization and learning rate of 0.0001. After training, the model accuracy was obtained on the testing sets, and compared to the cross-dataset performance by running each model on the testing set from each other dataset. To test generalization performance, the models were tested with novel data from geo-tagged tweets, media photographs, and protest datasets, which were all manually annotated.
   The researchers obtained several metrics from these tests. Accuracy is simply the proportion of correct classifications over total, measuring the model against the ground truth. The researchers also used standard deviation among different races and balanced accuracy, which weighs by class, to show how the models perform differently by race. Maximum accuracy disparity measures the greatest difference between classes on a logarithmic scale and shows how classes differ. Finally, pairwise distance analysis measures the distances between pairs of data points to show how well the model can separate faces.
3. The paper defines bias in face recognition as unequal outcomes among racial groups, with a less biased model having lower disparity. In this case, the researchers focus on how a biased dataset can cause a theoretically unbiased model to have worse performance for minority racial groups. They use the metrics listed above like maximum accuracy disparity to quantify the bias in each dataset and find that FairFace leads to better accuracy in all categories.

4. I calculated two main statistics with the DeepFace API running on the FairFace dataset: distance from the true age group among each age group, and percent mismatch of age groups over each racial group. The first was to ensure that the data was even when weighted by age, and the latter was to detect bias in either the model or the dataset. The distance from true age group was measured by placement of bin: for example, a prediction of 30-39 for a 10-19 person would be a value of 2. The DeepFace API was not configured to distinguish between Eastern and Southern Asians, so they were merged into a single group.

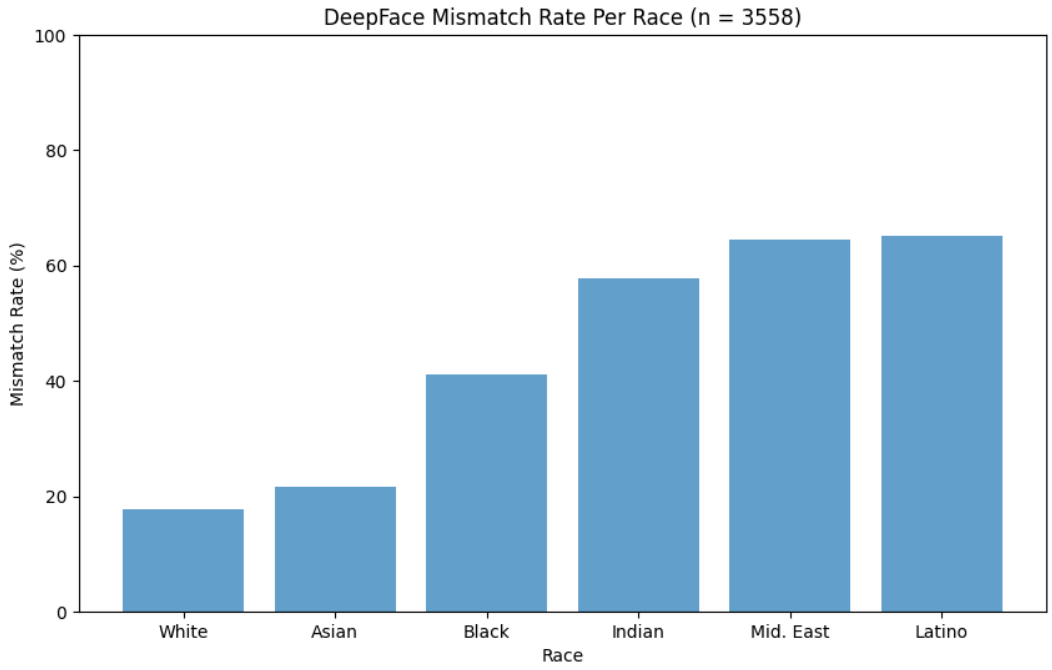
	Age Range								
	0-2	3-9	10-19	20-29	30-39	40-49	50-59	60-69	70+
Total Images	0	0	1	1921	1529	105	2	0	0
Avg. Bin Distance	-	-	1.00	1.02	1.16	1.29	1.00	-	-



This metric is not particularly helpful, especially because DeepFace erroneously guessed entirely in the center range, but it does show that this algorithm is not consistently reliable with the parameters and setup that I used. This visualization does not fully capture the fact that the outer two bars are made up of one data point

while the inner bars are hundreds or up to one thousand, but it does show again that this method is unreliable at best.

	Mismatch Rate Per Race					
	White	Black	Asian	Latino/ Hispanic	Middle Eastern	Indian
Mismatches	115	124	234	399	240	314
Total Images	647	302	1080	613	372	544
Mismatch Rate	**17.8%**	41.1%	21.7%	65.1%	64.5%	57.7%



I calculated the rate of age mismatch per age range and age mismatch per race, which showed that the ages of white people were substantially less likely to be misidentified than other races. The initial count of faces was 10954, but only 3558 were recognized, so it is possible that the model failed to recognize darker faces entirely, leading to fewer numbers and skewed results. White and especially Asian faces were significantly more prevalent in the result than Black and Middle Eastern faces, although Latino and Indian faces had highly unreliable results even with their higher sample sizes. The fact that three groups were worse than guessing, and that Black people were misrepresented 40% of the time, makes this combination of model and database entirely unviable in any real scenarios. It is, however, difficult to

make a definitive statement from this data because the face recognition was itself flawed.