Bias in Facial Classification ML Models

Patrick Connelly Grace Cooper Bhavana Jonnalagadda Carl Klein
Piya (Leo) Ngamkam Dhairya Veera

2023-12-13

Table of contents

Abstract

Bias in how facial classification machine learning (ML) models label faces is a burgeoning problem; as the use of such models becomes widespread, it is more important than ever to identify the weaknesses in the models and how they could potentially discriminate against various class, like race, gender, or age. In this study, we run two widely used facial classification models (FairFace and DeepFace) on a popular face dataset (the UTKFace Dataset) and perform two sample proportion hypothesis tests – as well as evaluating model output using common ML performance metrics – in order to highlight and identify potential bias in the aforementioned classes. We found that DeepFace had significant bias in age and race, with white males being classified more accurately than other factor categories; FairFace performed significantly better with less detected bias, affirming the intended goal of FairFace being built specifically to be more "fair" (less biased) on various categories. The implications lead us to recommend more work to be done on improving facial classification ML models, in order for them to be equitable and fair to all humans they are run on.

Report PDF and Code Location

A link to download the PDF version of this report, a link to the Github source code for this report, and the Youtube presention are available as icons in the top nav bar of this website.

1 Introduction

The issue of algorithmic bias, especially concerning sensitive and personal data, is an ongoing problem in today's use of Artificial Intelligence (AI). Facial recognition is one field that is struggling with mitigating and minimizing the issue. According to a report by the National Institute of Standards and Technology, the rates of false positives, or misidentification, of African and East Asian faces were 10 to 100 times higher than those for White or European faces (NIST 2020). Numerous studies have found that many facial recognition algorithms, having been based and created in white-dominated spaces, often lack accuracy with darker faces, especially compared to their identification of white faces. This issue has caused numerous problems throughout the development of facial recognition. For instance, a Georgetown study found that African Americans were significantly misidentified in law enforcement databases, due to being overrepresented in mugshots (Georgetown Law 2016). That sort of misinterpretation could lead to unlawful arrests, accusations, or sentencings. A facial recognition algorithm has two main areas where these sorts of biases occur: the actual coding/iteration, and the data used to train it. The databases used to teach an algorithm how to make decisions and identify faces matter, from the balance of different races, genders, and ages, to how well those databases use facial markers to identify anything. As facial recognition becomes more widespread, this becomes a key question of data ethics and misuse (Lohr 2018).

Thus, it is necessary to examine existing algorithms for their accuracy in identifying faces properly. Two easily accessible algorithms that claim to do just that are FairFace, created by UCLA researchers (Karkkainen and Joo 2021), and DeepFace (Serengil and Ozpinar 2021), created by a team of researchers at Facebook. Both claim to accurately identify the race, gender, and age of any given photo. FairFace claims to have reduced bias compared to other common facial recognition algorithms. FairFace was trained on a balanced dataset, eqully stratified across race, including Middle Eastern Faces. The creators point out in their work that the majority of training datasets overwhelmingly represent white and male subjects, lending to algorithmic biases in any models leveraging such data for training (Karkkainen and Joo 2021). The DeepFace algorithm was developed by a team at Facebook, now Meta, and also aims to be an accessible and accurate open-source facial recognition system. In their paper on research and development of DeepFace, the creators claim 97% accuracy on gender prediction, but only 68% accuracy on race and ethnicity. There is a more complex discussion of age prediction, and the creators further state that a previous study produced more accurate results when compared to the current model. Furthermore, the current model was claimed to be less accurate than human-provided predictions (Serengil and Ozpinar 2021).

Our goal in this research is to test the strength of the models' claims and compare the algorithms' ability to predict age, gender, and race against a source dataset. Both will be tested against the UTKFace dataset, which consists of over 24,000 labeled faces that can be used for research purposes ("UTKFace" 2021). We will identify potential biases in the modelsl using two-sample proportion hypothesis testing, and by inspect specific instances of such bias using performance metrics such as F1 score and accuracy.

2 Data

Pursuant to the study, the team sought out multiple datasets on which we could evaluate the performance of two selected recognition models (Karkkainen and Joo 2021; Serengil and Ozpinar 2021) to generate performance data and perform statistical analysis on their ability to accurately identify race, age, and gender of a subject in a photograph.

Collectively, we landed on the UTK dataset to perform our evaluation ("UTKFace" 2021). The dataset has three main sets available for download from the main page: A set of "in-the-wild" faces, which are the raw unprocessed images. The second set is the Aligned & Cropped Faces, which have been cut down to allow facial algorithms to read them more easily. The final file is the Landmarks (68 points) dataset, which contains the major facial landmark points that algorithms use and process to examine the images.

2.1 Data Selection

2.1.1 Motivation

Joy Buolamwini, a PhD candidate at MIT Media Lab, published a paper on gender and racial biases in facial recognition in algorithms (Buolamwini 2023). In her paper, she tested facial recongition softwares from multiple large technology companies such as Microsoft, IBM, and Amazon on its effectiveness for different demographic groups. Her research led to a surprising conclusion that most AI algorithms offer a substantially less accurate prediction for feminine/female faces, particularly those with dark skin color.

To determine the degree in which bias is still present in modern facial recognition models, a dataset which comprise of face images with high diversity in regards to eethnicity is required. Upon searching, UTKFace came out as one of the largest datasets which fit our preferred qualifications.

2.1.2 Data Collection Method

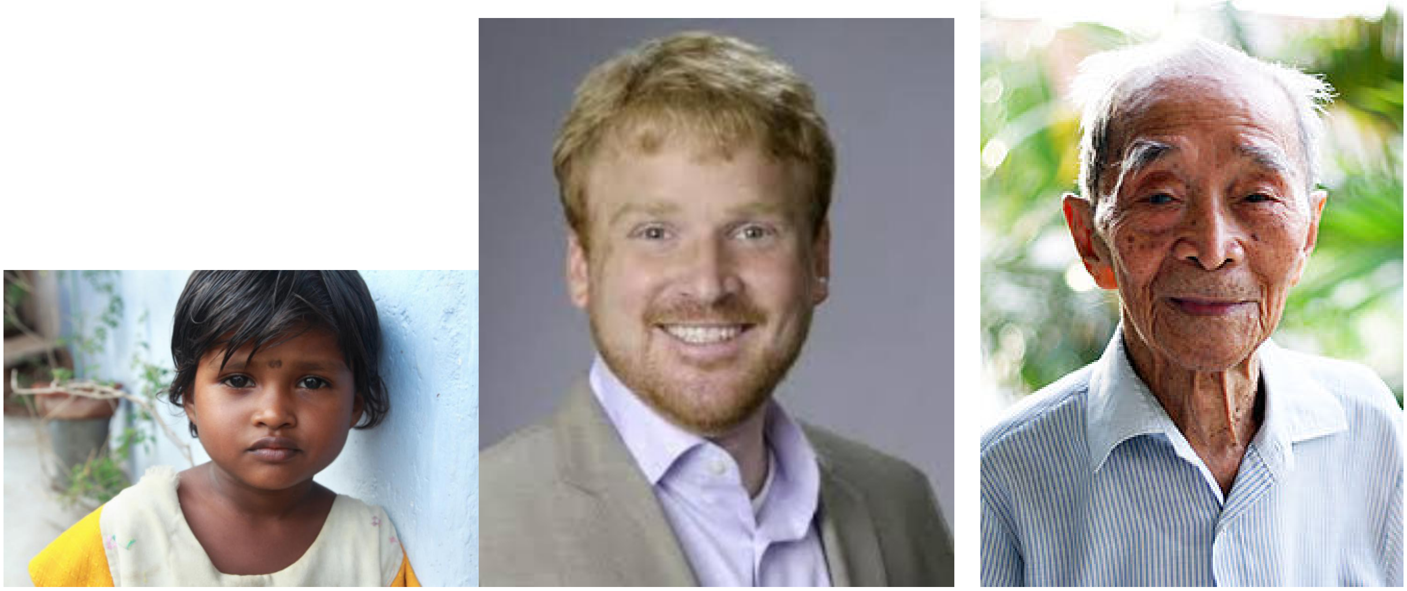
The dataset utilized for this research is UTKFace dataset. It is a publicly available large scale face dataset non-commercial on Github. The dataset was created by Yang Song and Zhifei Zhang, researchers at Adobe and PhD candidates at The University of Tennessee, Knoxville. On its Github page, it is specified that the images were collected from the internet. They appear to be obtained through the application of technique such as web scraping. The dataset contains more than 24,000 face images, representing a highly diversified demographics. However, face images vary in pose, facial expression, lighting, and resolution.

2.1.3 Dataset Features

The input dataset provided feature information natively in each filename without additional external data. The features contained therein include the following items for each image's subject. They are defined as follows:

- "race is an integer from 0 to 4, denoting White, Black, Asian, Indian, and Others (like Hispanic, Latino, Middle Eastern)."
- "gender is either 0 (male) or 1 (female)"
- "age is an integer from 0 to 116, indicating the age"

As our work is focused in potential biases in protected classes such as race, gender, and age, the features of UTKFace are sufficient to meet the needs for an input dataset for category prediction in our selected models. Examples of the source dataset images are in Figure 2.1.



(a) Age=6, Gender=F, Race=Indian (b) Age=38, Gender=M, Race=White (c) Age=80, Gender=M, Race=Asian

Figure 2.1: Example face images from the UTK dataset ("UTKFace" 2021) with their associated given labels.

2.1.4 Sources and Influences of Bias in the Dataset

Facial datasets can be extremely hard to categorize correctly, never mind reducing bias overall. Facial features that are androgynous or defer from the average features of the set can often be misrepresented or reported incorrectly. Those with features that make them look younger or older than their actual age may also be difficult for a computer to accurately guess.

The datasets used for analysis contain solely male/masculine and female/feminine faces. As stated above, the faces are labelled either 0, for male, or 1, for female. There are no gender non-conforming/non-binary/trans faces or people reported in the datasets, which could introduce potential bias. This absence of an entire category of facial features could also result in inaccurate guesses should these faces be added to the data later.

The datasets do not report nationality or ethnicity. This can introduce inaccuracy in the part of the identification, and it also may identify the face in a racial group that the person identified would consider inaccurate. This is as much a matter of potentially inaccurate data as it is social labels. There is also a level of erasure associated with simply creating a "multi-racial" category, given that it would bin all multiracial faces together with no further consideration. That is to say, there is no ideal solution to the issue at this time. However, it is always worth pointing out potential biases in data, research, and analysis.

The data given in the UTK dataset is composed purely of people who have their faces on the internet. This introduces a potential sampling bias. Given the topic, it is also likely to come from populations well-versed in technology. This can often exclude rural populations. Thus, the facial data present can be skewed towards urban residents or other characteristics, which can potentially create "lurking variables" that we aren't aware of within the data. This is a common problem that many Anthropological and Sociological studies face when collecting and analyzing data. Being aware of the possibility is often the first, and most crucial, step towards reducing it.

Our source dataset, and thus our results and conclusions, are dependent on the correctness of labeling of images within the UTK dataset. Given that the dataset was web-scraped, we do not know the degree of care placed on dataset labeling during web-scraping. Any incorrect labels present in the data can skew our results.

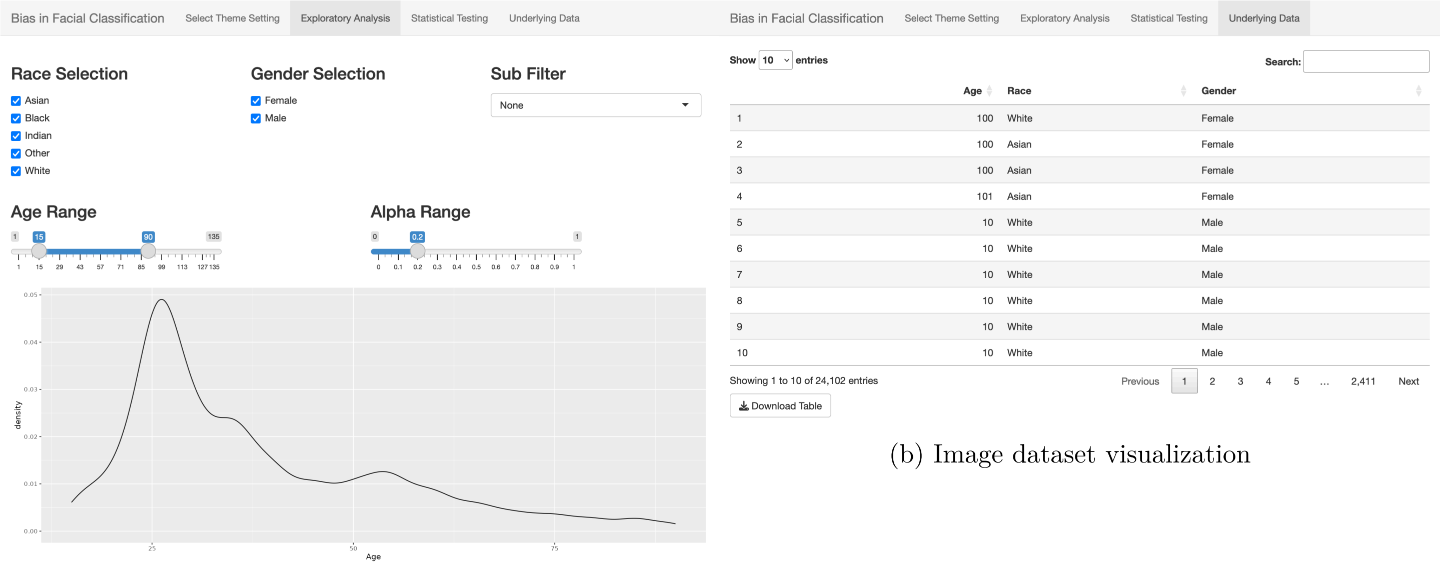
Overall, all the given potential biases listed above are simply the largest and most easily identified. It is possible that other sources of bias are present in the data that we haven't noticed. And identifying these biases does not mean that the data is not sound, or that any conclusions drawn from it are invalid. It simply indicates that further research should be done and that this data is far from the most complete picture of human facial features and identification. Examples of what is in the data, as well as a visualization of the bias present in the data, can be seen in **?@fig-data-eda**.

2.1.5 Exploration of Source Data

For initial exploration of the UTKFace dataset, we sought to determine the distribution of age, given other categorical variables. To support hypothesis testing, such as z-tests, t-tests, it is important for us to inspect our data for a normal distribution. In our case, we are only able to initially inspect age, as it is the only numerical variable from our data available.

Examining the data in **?@fig-data-eda**, we have a somewhat normal distribution of age with heavy tails, centered between the ages of 30 and 35. To examine distributions of race and gender, we will perform a bootstrapped sampling of proportions of these variables, and include them in our results section. Having such distributions will provide normal distributions and support us in evaluating our results.

A note on sample independence. For each of the selected facial recognition models, we assume that each model's training dataset is independent of the content of the UTKFace dataset. Independence between each model's output and the source data is a requirement for performing our testing. We have no means or methods to verify whether or not any UTKFace images were used in the training of either model, and must make this assumption before moving forward in our methods and results.



Race Selection

☑ Asian
☑ Black
☑ Indian
☑ Other
☑ White

Gender Selection

☑ Female
☑ Male

Sub Filter

None

Age Range

1 15 90 135

Alpha Range

0 0.2 1

Show 10 entries

Search:

	Age	Race	Gender
1	100	White	Female
2	100	Asian	Female
3	100	Asian	Female
4	101	Asian	Female
5	10	White	Male
6	10	White	Male
7	10	White	Male
8	10	White	Male
9	10	White	Male
10	10	White	Male

Showing 1 to 10 of 24,102 entries

Previous 1 2 3 4 5 … 2,411 Next

Download Table

(b) Image dataset visualization

(a) Image data EDA

Figure 2.2: Screenshots of the interactive figure showcasing the distributions of various data factors in the image dataset, and showcasing the underlying data. To see and interact with this figure, go to the website link

2.2 Selected Models

2.2.1 FairFace

Developed by researchers at University of California, Los Angeles, FairFace was specifically designed to mitigate gender and racial biases. The model (Karkkainen and Joo 2021) was trained on 100K+ face images of people of various ethnicities with approximately equal stratification across all groups. Beside facial recognition model, FairFace also provided the dataset (Karkkainen and Joo 2021) which it was trained on. The dataset is immensely popular among facial recognition algorithm developers. Owing to its reputation in bias mitigation, FairFace appears to be a valuable piece for the objective of this research.

2.2.2 DeepFace

DeepFace is a lightweight open-source model developed and used by Meta (Facebook). Being developed by one of the largest social media companies, it is widely known among developers. Therefore, its popularity prompts us to evaluate its performance. It should be noted that the DeepFace model we leverage in our evaluation is a free open source version (Serengil and Ozpinar 2021). It is highly unlikely that this version is as advanced as any model Meta uses internally for proprietary purposes. We should not view the resulting output of this model as being representative of algorithms internal to Meta.

2.2.3 FairFace Outputs

FairFace outputs provided predictions age and race, and two different predictions for race - one based upon their "Fair4" model, and the other based upon their "Fair7" model. In addition to these predictions, the output included scores for each category. With the nature of our planned analyses, the scores are of less importance to us in our evaluation.

To examine more in detail on "Fair" and "Fair4" models, the latter provided predictions of race in the following categories: [White, Black, Asian, Indian]. Of note, the "Fair4" model omitted "Other" categories as listed in the race category for the UTK dataset. However, the "Fair7" model provides predictions across [White, Black, Latino_Hispanic, East Asian, Southeast Asian, Indian, Middle Eastern]. We elected to use the Fair7 model, and to refactor the output categories to match those of the UTK dataset. Namely, we refactored instances of Middle Eastern and Latino_Hispanic as "Other" and instances of "East Asian" and "Southeast Asian" as "Asian" to match the categories explicitly listed in UTKFace.

Additionally, FairFace only provides a predicted age range as opposed to a specific, single, predicted age as a string. To enable comparison of actual values to the predicted values, we maintained this column as a categorical variable, and split it into a lower and upper bound of predicted age as an integer in the event we require it for our analyses.

With the above considerations in mind, the following output features are of import to the team:

Table 2.1: FairFace Output Format

Column Name	Data Type	Significance	Valid Values
name_face_align	String	The name and path of the file upon which FairFace made predictions	[filepath]
race_preds_fair7	String	The predicted race of the image subject	[White|Black|Latino_Hispanic|East Asian|Southeast Asian|Middle Eastern|Indian]
gender_preds_fair	String	The predicted gender of the image subject	[Male|Female]

Table 2.1: FairFace Output Format

Column Name	Data Type	Significance	Valid Values
age_preds_fair	String	The predicted age range of the image subject	[‘0-2’|‘3-9’|‘10-19’|‘20-29’|‘30-39’|‘40-49’|‘50-59’|‘60-69’|‘70+’]

2.2.4 DeepFace Outputs

Default outputs provide a wide range of information for the user. In addition to providing its predictions, DeepFace also provides scores associated with each evaluation on a per-class basis (i.e. 92% for Race #1, 3% Race #2, 1% Race #3, and 4% Race #4). For our planned analyses, the score features are of less concern to us.

We focus on the following select features from DeepFace outputs to have the ability to cross-compare between UTKFace, FairFace, and DeepFace:

Table 2.2: DeepFace Output Format

Column Name	Data Type	Significance	Valid Values
Age	Integer	The predicted age of the image subject	Any Integer
Dominant Gender	String	The predicted gender of the iamge subject	[Man|Woman]
Dominant Race	String	The predicted race of the image subject	[middle eastern|asian|white|latino hispanic|black|indian]

2.3 Evaluating Permutations of Inputs and Models for Equitable Evaluation

Aside from the differences in the outputs of each model in terms of age, race, and gender, there are also substantial differences between FairFace and DeepFace in terms of their available settings when attempting to categorize and predict the features associated with an image.

The need for this permutation evaluation rose from some initial scripting and testing of these models on a small sample of images from another facial dataset. We immediately grew concerned with DeepFace’s performance using default settings (namely, enforcing requirement to detect a face prior to categorization/prediction, and using OpenCV as the default detection backend). Running these initial scripting tests, we encountered a face detection failure rate, and thus a prediction failure rate, in DeepFace of approximately 70%.

We performed further exploratory analysis on both models in light of these facts, and sought some specific permutations of settings to determine which may provide the most fair and equitable comparison of the models prior to proceeding to analysis.

The goal for us in performing this exploration was to identify the settings for each model that might best increase the likelihood that the model’s output would result in a failure to reject our null hypotheses; our tests sought out the combination of settings that give each model the benefit of the doubt, and for each to deliver the greatest accuracy in their predictions. For simplicity’s sake, we leaned solely on the proportion of true positives across each category when compared with the source information to decide which settings to use.

2.3.1 DeepFace Analysis Options

DeepFace has a robust degree of available settings when performing facial categorization and recognition. These include enforcing facial detection prior to classification of an image, as well as 8 different facial detection models to detect a face prior to categorization. The default of these settings is OpenCV detection with detection enabled. Other detection backends include ssd, dlib, mtcnn, retinaface, mediapipe, yolov8, yunet, and fastmtcnn.

In a Python 3.8 environment, attempting to run detections using dlib, fastmtcnn, retinaface, mediapipe, yolov8, and yunet failed to run, or failed to install the appropriate models directly from source during execution. Repairing any challenges or issues with the core functionality of DeepFace and FairFace's code is outside the scope of our work, and as such, we have excluded any of these non-functioning models from our settings permutation evaluation.

2.3.2 FairFace Analysis Options

The default script from FairFace provided no options via its command line script to change runtime settings. It uses dlib/resnet34 models for facial detection and image preprocessing, and uses its own Fair4 and Fair7 models for categorization. There are no other options or flags that can be set by a user when processing a batch of images.

We converted the simple script to a class in Python without addressing any feature bugs or errors in the underlying code. This change provided us some additional options when performing the analysis of an input image using FairFace - namely, the ability to analyze and categorize an image with or without facial detection, like the functionality of DeepFace. FairFace remains limited in the fact that is only detection model backend is built in dlib, but this change from a script to a class object gave us more options when considering what type of images to use and what settings to use on both models before generating our final dataset for analysis.

2.3.3 Specific Permutations

With the above options in mind, we designed the following permutations for evaluation on a subset of the UTK dataset:

Table 2.3: List of Permutation Evaluations

Detection	Detection Model	Image Source
Enabled	FairFace=Dlib; DeepFace=OpenCV	Pre-cropped
Enabled	FairFace=Dlib; DeepFace=OpenCV	In-The-Wild
Enabled	FairFace=Dlib; DeepFace=mtcnn	Pre-cropped
Enabled	FairFace=Dlib; DeepFace=mtcnn	In-The-Wild
Disabled	FairFace,DeepFace=None	Pre-cropped
Disabled	FairFace,DeepFace=None	In-The-Wild

We processed each of the above setting permutations against approximately 9800 images, consisting of images from part 1 of 3 from the UTK dataset. Each of the cropped images (cropped_UTK_dataset.csv) and uncropped images (uncropped_UTK_dataset.csv) came from the same underlying subject in each image; the only difference between each image was whether or not it was pre-processed before evaluation by each model. Having the same underlying source subject enables us to perform a direct comparison of results between cropped vs. in-the-wild images, and better support a conclusion of which settings to use.

Table 2.4: Results of Permutation Evaluation

pred_model	detection_enabled	detection_model	image_type	all_rate	age_grp_rate	gender_rate	race_
DeepFace	False	None	cropped	0.0724949	0.1601227	0.6667689	0.6951
DeepFace	False	None	uncropped	0.0834356	0.1522495	0.7326176	0.6457
DeepFace	True	mtcnn	cropped	0.0889571	0.1534765	0.7249489	0.6807
DeepFace	True	mtcnn	uncropped	0.1023517	0.1615542	0.7834356	0.6665
DeepFace	True	opencv	cropped	0.0267894	0.0765849	0.1887526	0.1983
DeepFace	True	opencv	uncropped	0.0806748	0.1455010	0.6619632	0.5855
FairFace	False	None	cropped	0.4015337	0.6101227	0.8921268	0.7689
FairFace	False	None	uncropped	0.1031697	0.2671779	0.7599182	0.4477
FairFace	True	dlib	cropped	0.4015337	0.6101227	0.8921268	0.7689
FairFace	True	dlib	uncropped	0.4353783	0.6230061	0.9155419	0.7914

Examining the true positive ratios for each case, our team concluded that the settings that gave both models the best chance for success in correctly predicting the age, gender, and race of subject images are as follows:

- FairFace: enforce facial detection with dlib, and use uncropped images for evaluation
- DeepFace: enforce facial detection with MTCNN detection backend and use uncropped images for evaluation.

These settings are equitable and make a degree of sense. Using facial detection, specifically coded for each model, should give each model the ability to isolate the portions of a face necessary for them to make a prediction, as opposed to using a pre-cropped image that could include unneeded information, or exclude needed information.

Having decided on these settings, our team proceeded to run the entirety of the UTK dataset through both DeepFace and FairFace models using a custom coded script that allowed us to apply multiprocessing across the list of images and evaluate all items in a reasonable amount of time.

Due to the resource-intensive design of FairFace, our script enables multiprocessing of FairFace to allow for multiple simultaneous instances of the FairFace class as a pool of worker threads to iterate over the source data.

We attempted the same multiprocessing methodology for DeepFace, but encountered issues with silent errors and halting program execution when iterating over all images using DeepFace. To alleviate this challenge, we processed DeepFace in a single-threaded manner, and with smaller portions of the dataset vs. pursuing an all-in-one go execution. We proceeded to store the data for each of these smaller runs in multiple output files to combine once we completed all processing requirements.

2.4 Model Evaluation Data Format

The final listing of all inputs and outputs from each model, with standardization methods discussed in this section applied, are summarized in Table 2.5.

Table 2.5: Data Format for All Inputs and Outputs

Column Name	Definition
img_path	Relative path location of the file within the UTK dataset
file	The filename of each file within the UTK dataset
src_age	The age of the subject in each image from the UTK dataset
src_gender	The gender of the subject in each image from the UTK dataset
src_race	The race of the subject in each image from the UTK datset
src_timestamp	The time at which the image was submitted to the UTK dataset
src_age_grp	The age group (matching the predicted age ranges from the FairFace outputs) for each image in
pred_model	The model used to produce the predicted output (FairFace or DeepFace)
pred_race	The race of the subject in the image, predicted by the given prediction model under the pred_m
pred_gender	The gender of the subject in the image, predicted by the given prediction model under the pred_
pred_age_DF_only	The integer-predicted age by DeepFace of the subject in the image
pred_age_grp	The age group of the subject in the image, predicted by the given prediction model under the pr
pred_age_lower	The integer lower bound of the predicted age group
pred_age_upper	The integer upper bound of the predicted age group

3 Methods

As described in the previous section, the two selected models (DeepFace and FairFace) are run on the UTK face dataset in order to generate output of classification across 3 categories (age, race, and gender). We evaluate the performance of this classification, and perform hypothesis testing in order to answer the key research questions.

3.1 Data Cleaning: Standardizing Model Outputs

As can be seen in Chapter 2, there are some key differences between the outputs of both models as well as the source data that we needed to resolve to enable comparison of each dataset to one another. We'll focus on the primary features of age, gender, and race from each dataset.

3.1.1 FairFace Output Modifications

We'll discuss FairFace first, as it introduces a requirement for modification to both our input information as well as the outputs for DeepFace.

- **Age**: FairFace only provides a categorical predicted age range as opposed to a specific numeric age. We retain this age format and modify the last category of "70+" to "70-130" to ensure we can capture the gamut of all input and output ages in all datasets.
- **Gender**: No changes to predicted values; use "Male" and "Female"
- **Race**: the source data from UTKFace has 5 categories "White" "Black" "Asian" "Indian" and "Other". Using the definitions from UTKFace, we collapse the output categories of FairFace's Fair7 model as follows:

["Southeast Asian","East Asian"] => "Asian" ["Middle Eastern" , "Latino_Hispanic] =>"Other"

3.1.2 DeepFace Output Modifications

- **Age**: Cut the predicted age into bins based upon the same prediction ranges provided by FairFace. If the DeepFace predicted age falls into a range provided by FairFace, provide that as the predicted age range for DeepFace.
- **Gender**: we adjust the DeepFace gender prediction outputs to match that of the source and FairFace data with the following refactoring: "Man" => "Male" "Woman" => "Female"
- **Race**: we adjust the DeepFace race prediction outputs to match that of the source dataset with the following refactoring:
- "white" => "White"
- "black" => "Black"
- "indian" => "Indian"
- "asian" => "Asian"
- ["middle eastern", "latino hispanic"] => "Other"

3.1.3 Source Data Modifications

- **Age:** We cut the predicted age into bins based upon the same prediction ranges provided by FairFace. If the input / source data age falls into a range provided by FairFace, provide that is the source age range for the image subject.
- **Gender:** No changes.
- **Race:** No changes.

3.2 Exploratory Data Analysis (EDA)

Our EDA performed on the source UTK dataset can be seen in the previous section in Figure 2.2. The EDA performed on the output from the models can be summarized as follows, and is presented in the Results section:

- Visualization of the histograms of distributions of predictions, per each category, per each model

We also perform some meta-analysis on the statistics and performance metrics calculated from the model outputs:

- Visualization of the p-values vs F1-score across all hypothesis tests across both models
- Confusion matrix of whether we reject or fail to reject the null hypothesis based on power and F1 score

3.3 Research Questions

We evaluate the output of the 2 models in order to answer the following questions:

- Is it possible to identify, detect, and classify biases in machine learning models using proportionality testing?
- Is bias prevalent in facial recognition machine learning models?
- Can one model be shown to have a greater quantity of statistically significant biases than the other?
  - in all aspects?
  - In specific/certain aspects?
- Are there disparate outcomes (i.e. higher chance of incorrect predictions) for one racial group vs. another?

3.4 Hypothesis Testing

Our data consists of three main sets: the source input data, the Fairface output data, and the Deepface output data.

We'll be creating our hypothesis tests by running as two-sample proportion tests. The population is the set of all labels (of race, age, and gender as defined below) for a given image, for all face images. The first sample will be the source dataset "correct" labels of the images, and the 2nd sample will be the output of a given model between FairFace and DeepFace, respectively. The base null hypothesis will produce no difference in sample proportions. Gaining a statistically significant result would allow us to reject our *null hypothesis* in favor of the *alternative hypothesis*. In other words, rejecting the original assumption means there is a statistically large enough difference between the source data and output data, and could indicate that the source and predicted information originate from differing populations, which is a potential indicator of bias for or against the protected classes in question. We use a significance level of 99.7% to mitigate the risk of rejecting the the null hypothesis when it is true.

We'll be testing across different subsets contained within the data, as listed below:

3.4.1 Demographics

- Age Group
- Gender
- Race

3.4.2 Demographics' Subgroups

- Age Group (9 groups)
  - 0-2
  - 3-9
  - 10-19
  - 20-29
  - 30-39
  - 40-49
  - 50-59
  - 60-69
  - 70-130
- Gender (2 groups)
  - Female
  - Male
- Race (5 groups)
  - Asian
  - Black
  - Indian
  - Other
  - White

3.4.3 The General Proportion Tests

Our hypothesis tests will be testing different proportions within these subgroups between the source data and the output data.

The general format of our hypothesis tests will be:

$H_0 : p_1 = p_2$

$H_A : p_1 \neq p_2$

With the following test statistic:

$$\frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_{p_1}} + \frac{1}{n_{p_2}})}}$$

With the p-value being calculated by:

$P(|Z| > z|H_0)$

$$= P(|Z| > \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_{p_1}} + \frac{1}{n_{p_2}})}},$$

Where:

- p_1 = the source dataset categories labels given and p_2 = the chosen model's labels given.
- $\hat{p}$ = the pooled proportion.
- n_{p_1}, n_{p_2} = the size of each sample.

We also calculate the power of each test performed, and use a power level threshold of 0.8 in order to assess the strength of the p-value calculated.

We believe that using two-sample proportion testing is an appropriate means by which we can evaluate the outputs of the two facial recognition models in comparison to the source data. In leveraging two-sample proportion tests, we can infer whether or not the proportions of age, gender, or race (or some combination thereof) from the UTKFace dataset (i.e. 1st sample) originate from the same population as the outputs from each facial recognition model (i.e. 2nd dataset).

In theory, similar proportions of protected classes between the two datasets could suggest that the source data and predicted data originate from the same population (pictures of people), and would thus indicate an absence of bias against the protected class in question. Vastly different proportions, however, could indicate that the source data and predicted data are from differing populations and indicate a bias against the protected classes in question.

Leveraging p-values and powers calculated on our samples for our protected classes of age, gender, and race, should enable us to provide a clear picture of any biases that may manifest from one or both models. Leveraging F1 scores (as described below) will help us identify the specific cases of bias, and whether it is in favor of or against a specific group.

3.4.4 Notation

We introduce notation for the specific tests we perform:

Let R be race, then $R \in \{Asian, Black, Indian, Other, White\} = \{A, B, I, O, W\}$

Let G be gender, then $G \in \{Female, Male\} = \{F, M\}$

Let A be age, then $A \in \{[0,2], [3,9], [10,19], [20,29], [30,39], [40,49], [50,59], [60,69], [70,130]\}$; or $A = \{1,2,3,4,5,6,7,8,9\}$

Let D be the dataset, then $D \in \{Source, Fairface, Deepface\} = \{D_0, D_f, D_d\}$

3.4.5 Proportion Testing of Subsets

Using this notation, we can simplify our nomenclature for testing a certain proportion of an overall demographic.

For example, we can test if the proportion of *Female* in the Fairface output is statistically different than the proportion of *Female* from the source.

Hypothesis Test:

$H_0 : p_{F,D_f} = p_{F,D_0}$

$H_A : p_{F,D_f} \neq p_{F,D_0}$

P-value Calculation:

$$P(|Z| > \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_{p_1}} + \frac{1}{n_{p_2}})}}),$$

where

- $\hat{p}_1 = p_{F,D_0}$: proportion of females from the source data
- $\widehat{p_2} = p_{F,D_f}$: proportion of females from the FairFace output

Additionally, we could test for different combinations of subsets within demographics. For instance, if we wanted to test for a statistically significant difference between the proportion of those who *Female*, given that they were *Black*, as predicted by DeepFace, then we could write a hypothesis test like:

$H_0 : p_{D_d,F|B} = p_{D_0,F|B}$

$H_A : p_{D_d,F|B} \neq p_{D_0,F|B}$

These were two specific hypothesis tests, however, we'll be testing all combinations of these parameters and reporting back on any significant findings.

In the above, we've outlined our methods for examining a total of 432 hypothesis tests per recognition model on the totality of, and smaller samples of, our overall dataset. We have elected to sub-divide our source and predicted samples by these protected classes to inspect and investigate whether or not there may be bias against groupings of protected classes.

For instance, in the performance of our hypothesis tests, we may find an absence of bias when only examining proportions of gender between samples. However, by examining a subset of our samples, such as subject gender given the subject's membership in a specific racial category, we may find biases in predictions of subject gender given their membership in a specific racial group.

This could help us answer questions and draw conclusions about such groups. For example:

"Model X demonstrates bias in predicting the race of older subjects." Such a statement is not one of bias for or against the target group, but that a bias exists. A bias in either direction, if used in a decision-making process, could result in age discrimination.

"Model Y demonstrates bias in predicting gender, given the subject is Black, Asian, or Other." Such a statement is not one of bias for or against the target groups, but a statement that a bias exists. Such a bias, if used in a decision-making process, could result in gender or racial discrimination.

Structuring our tests in this manner will enable us to quickly analyze and report on the results of our tests.

3.5 Performance Measurement

We evaluate the performance of the models in order to choose which models to use (as described in the Data section), to ensure data integrity, and to evaluate the hypothesis testing in context of performance. These measures are not used in the calculation of the statistical/hypothesis testing.

There are four main measures of performance when evaluating a model:

- **Accuracy**
- **Precision**
- **Recall**
- **F1-Score**

Each of these performance measures has their own place in evaluating models; in order to explain the differences between these metrics, we start with concepts of positive and negative outcomes.

- **True Positive:** predicted positive, was actually positive (correct)
- **False Positive:** predicted positive, was actually negative (incorrect)
- **True Negative:** predicted negative, was actually negative (correct)
- **False Negative:** predicted negative, was actually positive (incorrect)

These outcomes can be visualized in a confusion matrix. In Figure 3.1, green are correct predictions while red are incorrect predictions.

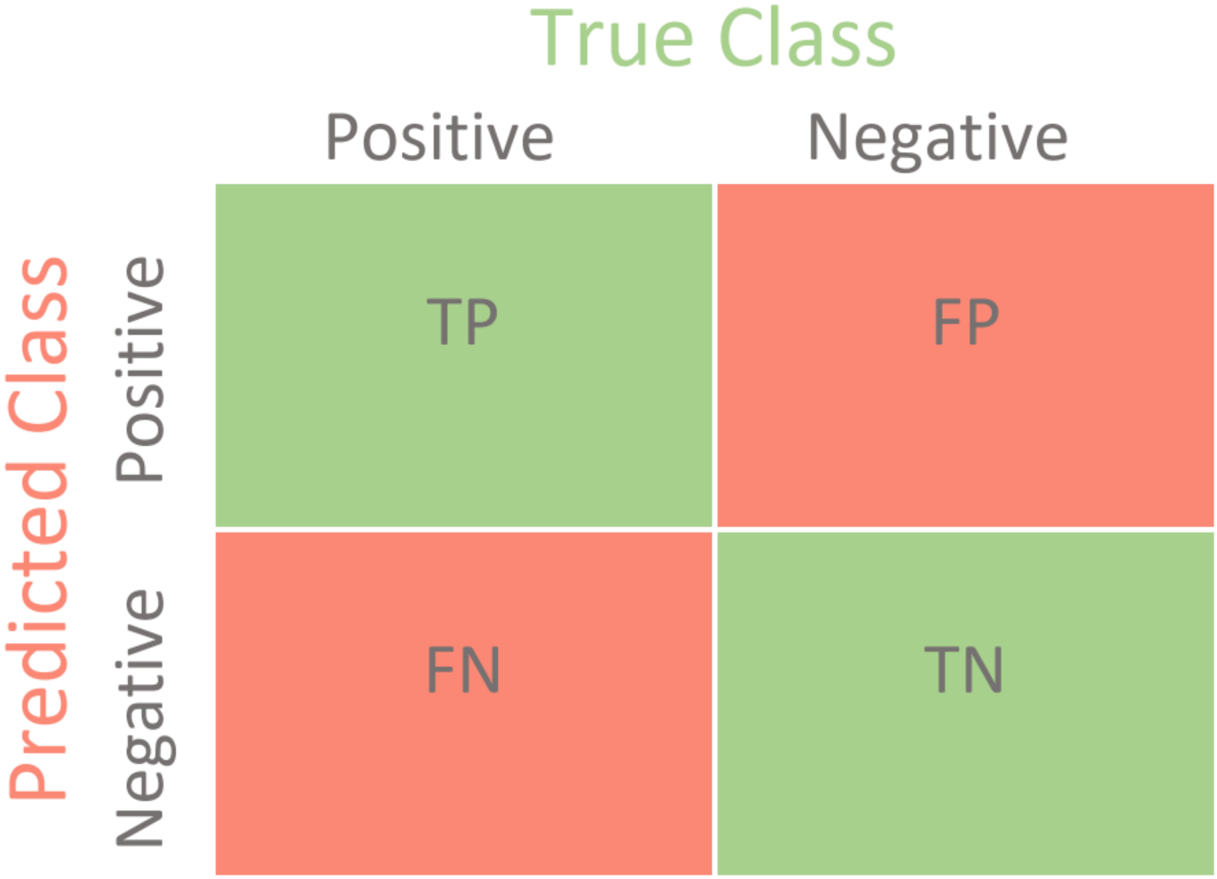


Figure 3.1: Confusion_matrix

3.5.1 Accuracy

Accuracy is the ratio of correct predictions to all predictions. In other words, the total of the green squares divided by the entire matrix. This is arguably the most common concept of measuring performance. It ranges from 0-1 with 1 being the best performance.

$Acccuracy = \frac{TP+TN}{TP+TN+FP+FN}$

3.5.2 Precision

Precision is the ratio of true positives to the total number of positives (true positive + true negative).

3.5.3 Recall

Recall is the ratio of true positives to the number of total correct predictions (true positive + false negative).

3.5.4 F1-Score

F1-Score* is known as the harmonic mean between precision and recall. **Precision** and **Recall** are useful in their own rights, but the F1-Score is useful in the fact it's a balanced combination of both precision and recall. It ranges from 0-1 with 1 being the best performance.

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

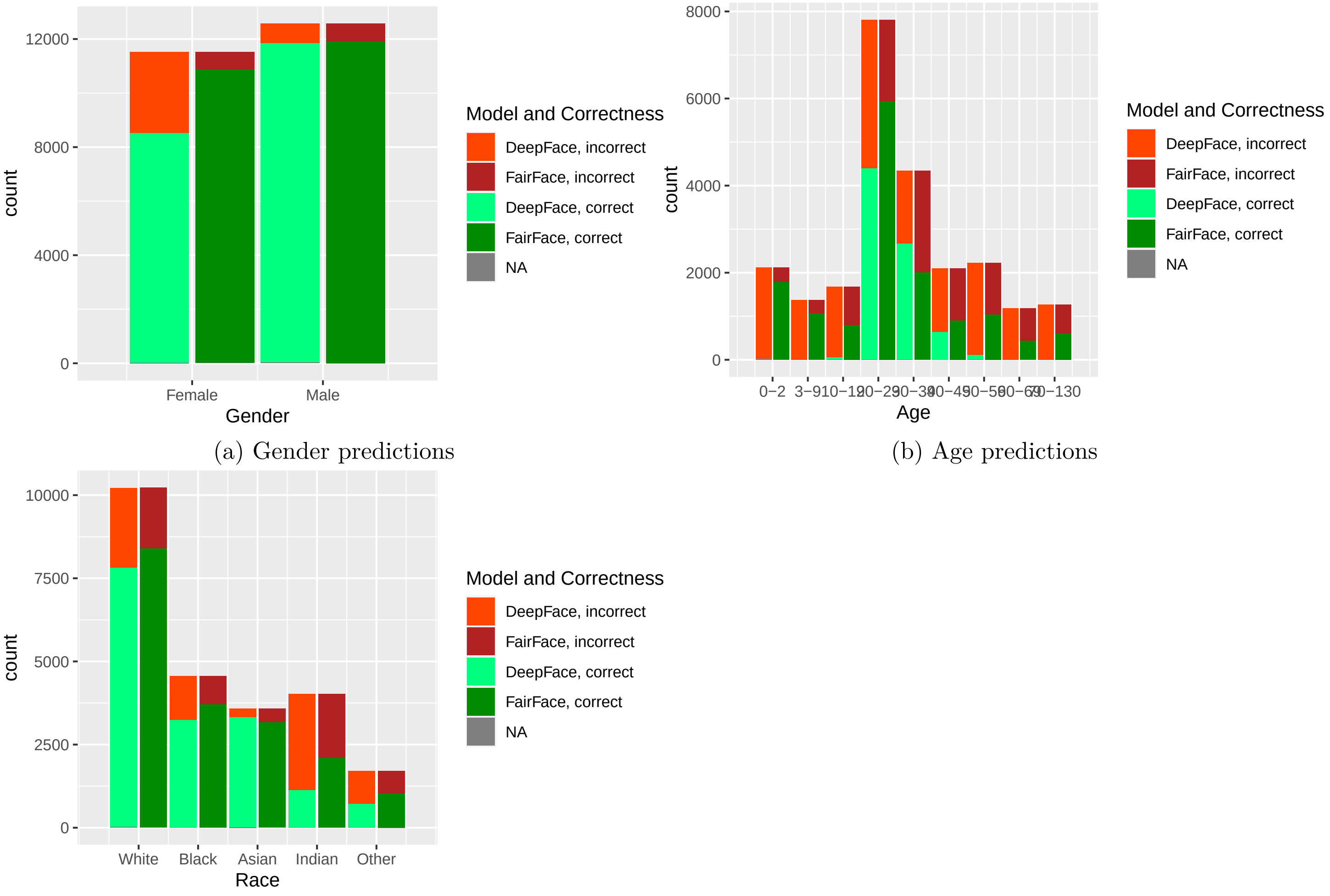
When considering the classification of a subject by protected classes of age, gender, and race, we believe that stronger penalties should be assigned in making an improper classification decision. Due to F1 being the harmonic mean of precision and recall, incorrect classification will more directly impact the score of each model in its prediction of protected classes, and do so more strongly than an accuracy calculation (Huilgol 2021).

We calculate F1 score as a measure of performance of our selected machine learning models. This was not used in the calculation or results of the hypothesis tests, but will be used for when we draw conclusions of our tests based upon p-value and statistical power. Namely, we do not plan to control for statistical power / Type-II error when running our 432 hypothesis tests, so statistical power may vary from test to test. Using F1 and Accuracy scores can support us in identifying specific cases of bias (and whether it is for or against) one or more protected classes. We elect to use an F1 score threshold of 0.9 to make this determination.

4 Results

4.1 Model Output

The two models, DeepFace and FairFace, were run on the dataset described previously. In Figure 4.1, one can see the results of the predictions done by each model, by each factor that was considered: age, gender, and race. Note that the total (across correct and incorrect) histogram distributions match the correct (source dataset) distributions of values in each category, so we can see exactly the difference between what was provided and what was predicted, along with how well each model did on each category within each factor.



(a) Gender predictions

(b) Age predictions

(c) Race predictions

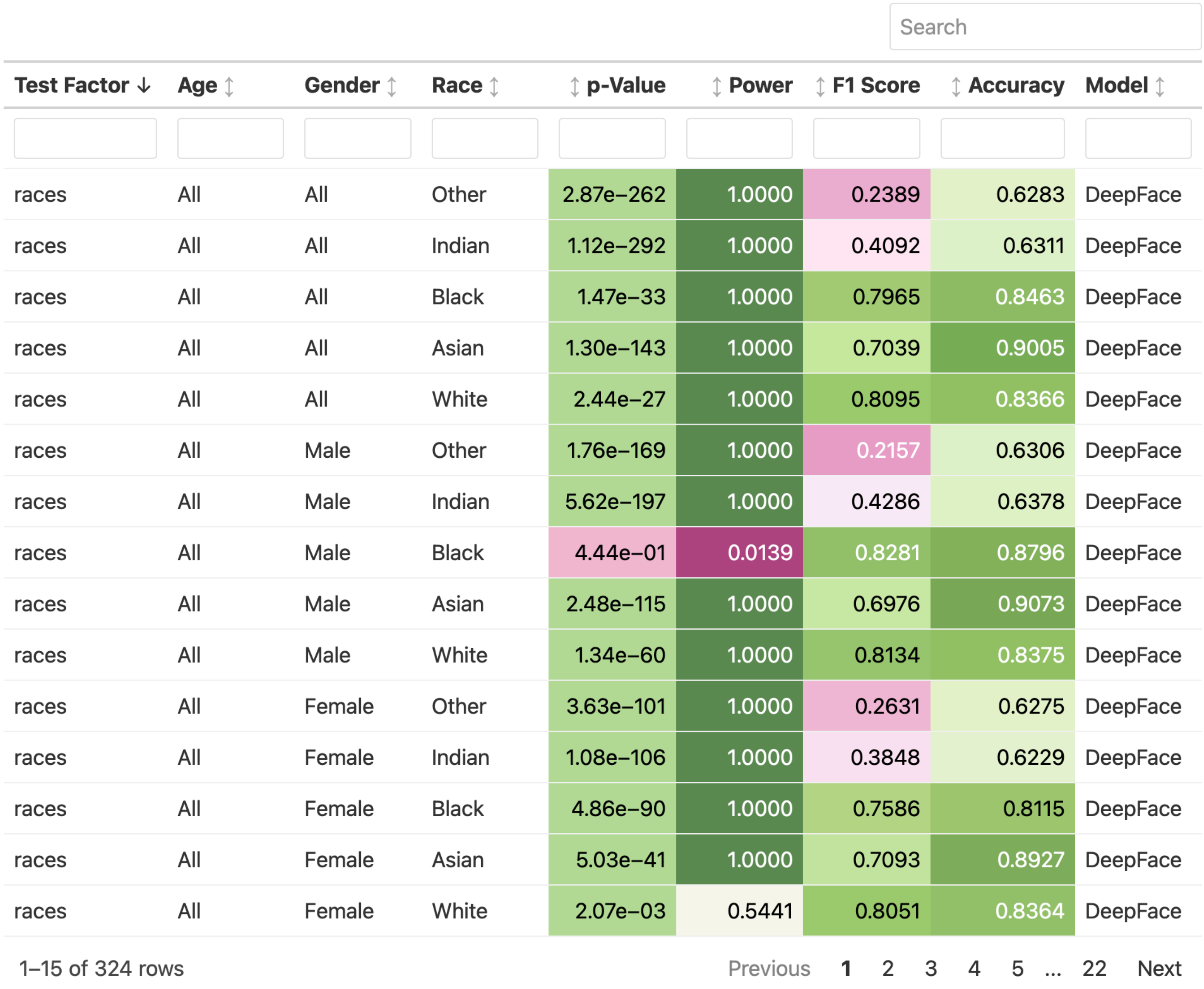
Figure 4.1: Histograms of the output from DeepFace and FairFace, with correct vs incorrect values colored. Note that the distributions match the correct (source dataset) distributions.

4.2 Model Performance, Hypothesis Testing

For each factor category and model, we calculate the F1 score, accuracy, p-value, and power, as described in section 3. Cell values are colored according to the strength of the metric; p-value is colored as to whether it crosses the significance value threshold of 0.003. We calculate these metrics and hypothesis tests across all categories of each factor, but also with conditional filtering on other factors; the value “All” indicates we did not filter/condition on that factor. The column `Test Factor` indicates which factor we are calculating the proportion for that hypothesis test. For example, the following column value subsets would indicate the given hypothesis test:

Test Factor	Age	Gender	Race	Model	Null Hypothesis	Description
gender	0-2	Female	All	FairFace	$p_{F,D_f\|A_1} = p_{F,D_0\|A_1}$	H_0 : The proportions of Female labels, given that the source age label is 0-2, are equal.
race	All	All	Black	DeepFace	$p_{R_B,D_d} = p_{R_B,D_0}$	H_0: The proportions of Black labels are equal.

The results are summarized in Figure 4.2.



Test Factor ↓	Age ↕	Gender ↕	Race ↕	↕ p-Value	↕ Power	↕ F1 Score	↕ Accuracy	Model ↕
races	All	All	Other	2.87e−262	1.0000	0.2389	0.6283	DeepFace
races	All	All	Indian	1.12e−292	1.0000	0.4092	0.6311	DeepFace
races	All	All	Black	1.47e−33	1.0000	0.7965	0.8463	DeepFace
races	All	All	Asian	1.30e−143	1.0000	0.7039	0.9005	DeepFace
races	All	All	White	2.44e−27	1.0000	0.8095	0.8366	DeepFace
races	All	Male	Other	1.76e−169	1.0000	0.2157	0.6306	DeepFace
races	All	Male	Indian	5.62e−197	1.0000	0.4286	0.6378	DeepFace
races	All	Male	Black	4.44e−01	0.0139	0.8281	0.8796	DeepFace
races	All	Male	Asian	2.48e−115	1.0000	0.6976	0.9073	DeepFace
races	All	Male	White	1.34e−60	1.0000	0.8134	0.8375	DeepFace
races	All	Female	Other	3.63e−101	1.0000	0.2631	0.6275	DeepFace
races	All	Female	Indian	1.08e−106	1.0000	0.3848	0.6229	DeepFace
races	All	Female	Black	4.86e−90	1.0000	0.7586	0.8115	DeepFace
races	All	Female	Asian	5.03e−41	1.0000	0.7093	0.8927	DeepFace
races	All	Female	White	2.07e−03	0.5441	0.8051	0.8364	DeepFace

1–15 of 324 rows

Figure 4.2: Screenshot of the interactive table showing F1 score, accuracy, p-value, and power, by each factor and category evaluated by the models, with a potential filtering condition. To see and interact with this table, go to the website link

4.2.1 p-value Critical Values

From the previous table, we extract and highlight key values; namely, where we reject the null hypothesis and where we do not, based on our criteria:

- Significance level of 99.7%
- Power threshold of 0.8
- F1-Score of 0.9

Which come from the rationale described in Chapter 3. We show the test values where there is no sub-filtering/conditions by another category; then, we also highlight the reverse null hypothesis decisions made with filtering for a sub-condition and for the specific rows as described in the table captions. The values are displayed in Table 4.2. There is only a Fairface table for not rejecting the null hypothesis (with no condition subfiltering) because no DeepFace values passed our given thresholds for not rejecting; the same reasoning is why there is no table for FairFace rejecting the null hypothesis with condition subfiltering.

Table 4.2: Highlighted statistics/metrics for DeepFace and FairFace, that pass the given significance level/power/F1-score thresholding.

	Category	p-Value	Power	F1 Score
age	70-130	$2.83e-43$	1.0000	0.6271
	3-9	$1.37e-05$	0.9198	0.7176
	10-19	$5.22e-05$	0.8640	0.5052
	0-2	$3.11e-06$	0.9568	0.8960
	20-29	$2.14e-08$	0.9959	0.7333
	40-49	$1.65e-08$	0.9965	0.3944
race	White	$5.83e-18$	1.0000	0.8610
	Black	$7.46e-12$	1.0000	0.8685
	Indian	$8.84e-94$	1.0000	0.6402
	Other	$0.00e00$	1.0000	0.3087

	Age	Gender	Race	p-Value	Power	F1 Score
age	0-2	Male	All	$4.94e-01$	0.0120	0.9190

	Category	p-Value	Power	F1 Score
age	70-130	$1.08e-283$	1.0000	NA
	3-9	$9.20e-293$	1.0000	NA
	10-19	$2.52e-148$	1.0000	0.0479
	0-2	$0.00e00$	1.0000	NA
	20-29	$2.00e-65$	1.0000	0.5054
	30-39	$0.00e00$	1.0000	0.3786
	40-49	$1.65e-91$	1.0000	0.2276
	50-59	$3.66e-202$	1.0000	0.0802
	60-69	$9.81e-229$	1.0000	0.0016
gender	Female	$1.18e-97$	1.0000	0.8198
	Male	$1.18e-97$	1.0000	0.8637
race	White	$2.70e-27$	1.0000	0.8095
	Asian	$1.75e-143$	1.0000	0.7039
	Black	$1.71e-33$	1.0000	0.7965
	Indian	$1.90e-292$	1.0000	0.4092
	Other	$4.64e-262$	1.0000	0.2389

	Age	Gender	Race	p-Value	Power	F1 Scor
gender	30-39	Male	All	$7.70e-02$	0.1185	0.922

	Category	p-Value	Power	F1 Score
gender	Female	$7.07e-01$	0.0053	0.9429
	Male	$7.07e-01$	0.0053	0.9476

4.3 Meta-Analysis Plots

In Figure 4.3, we show F1-score vs accuracy for all hypothesis tests that were performed. Note the relationship is not perfectly linear.

In Figure 4.4 and Figure 4.5 we explore our research question of whether or not two-sample proportion tests can approximate or predict the performance of a machine learning model. In each plot, we transform the p-value to 0 in cases where we would reject the null hypothesis, and 1 in cases for which we would fail to reject.

In Figure 4.6, we display confusion matrices of our null hypothesis rejections. We define the true/false positive/negative as follows:

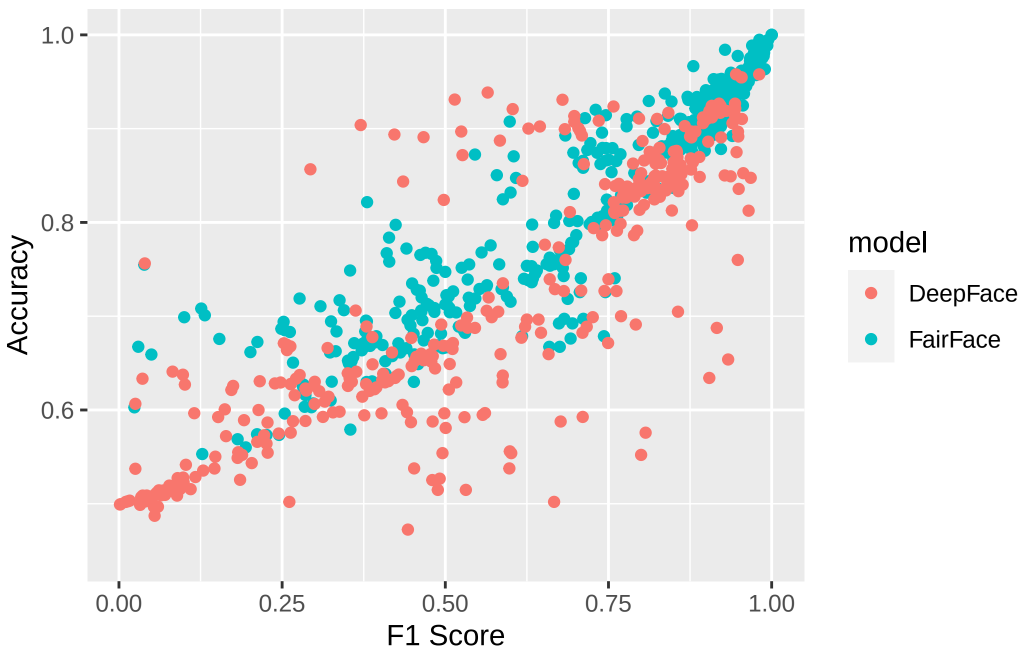
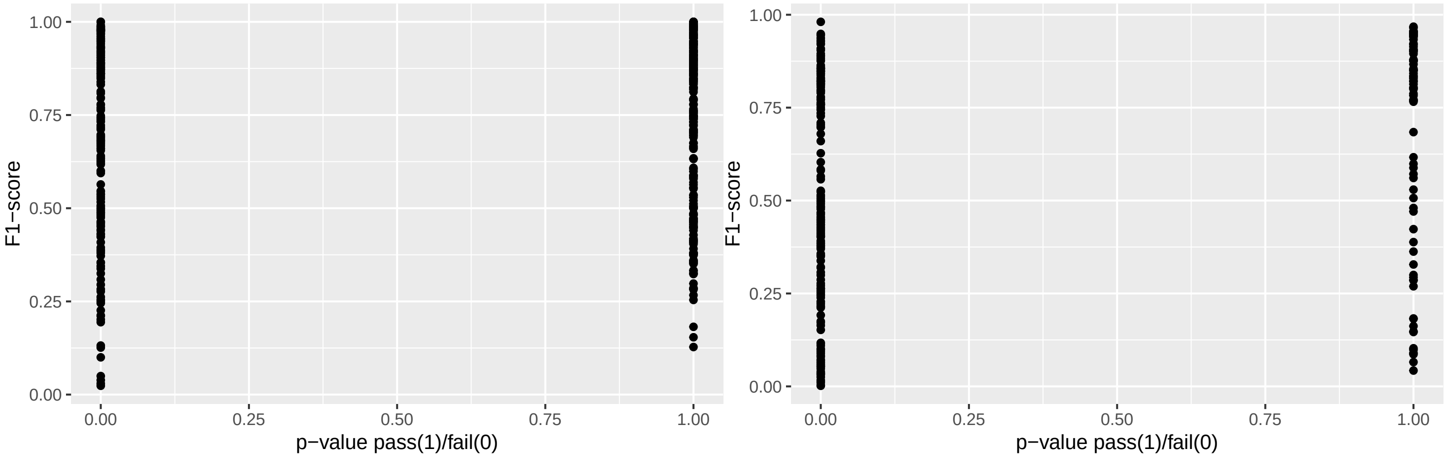


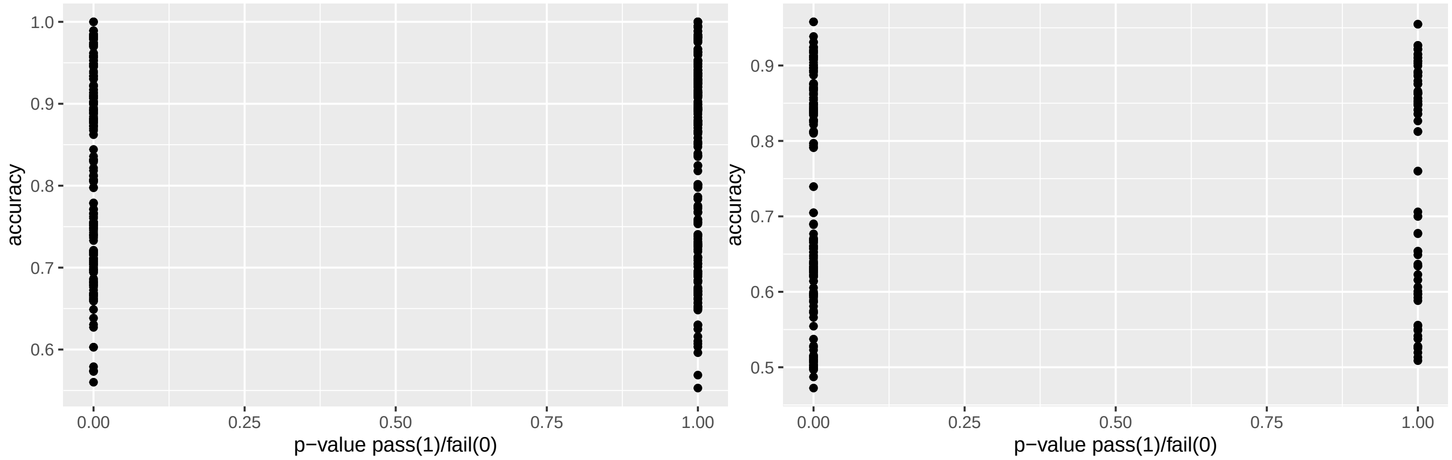
Figure 4.3: F1-Score vs Accuracy for all hypothesis tests performed.



(a) FairFace: two-sample proportion p-value vs F1

(b) DeepFace: two-sample proportion p-value vs F1

Figure 4.4: p-value vs F1 score for all hypothesis tests performed.

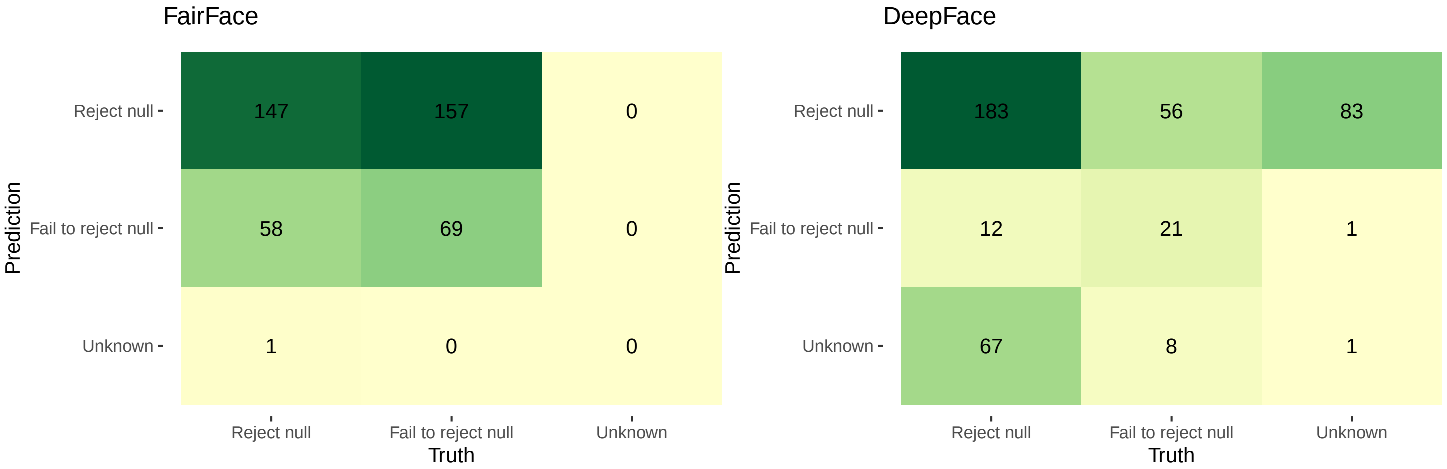


(a) FairFace: two-sample proportion p-value vs accuracy

(b) DeepFace: two-sample proportion p-value vs accuracy

Figure 4.5: p-value vs accuracy score for all hypothesis tests performed.

- Reject null when we should reject null: p-value < 0.003, F1 < 0.9, power >= 0.8
- Reject null when we should fail to reject null: p-value < 0.003, F1 > 0.9, power >= 0.8
- Fail to reject null, when we should reject null: p-value >= 0.003, F1 < 0.9, power < 0.8
- Fail to reject null, when we should fail to reject null: p-value >=0.003, F1 >= 0.9, power < 0.8
- Uknown: One of the values was NaN.



FairFace

Prediction \ Truth	Reject null	Fail to reject null	Unknown
Reject null	147	157	0
Fail to reject null	58	69	0
Unknown	1	0	0

(a) Matrix for FairFace

DeepFace

Prediction \ Truth	Reject null	Fail to reject null	Unknown
Reject null	183	56	83
Fail to reject null	12	21	1
Unknown	67	8	1

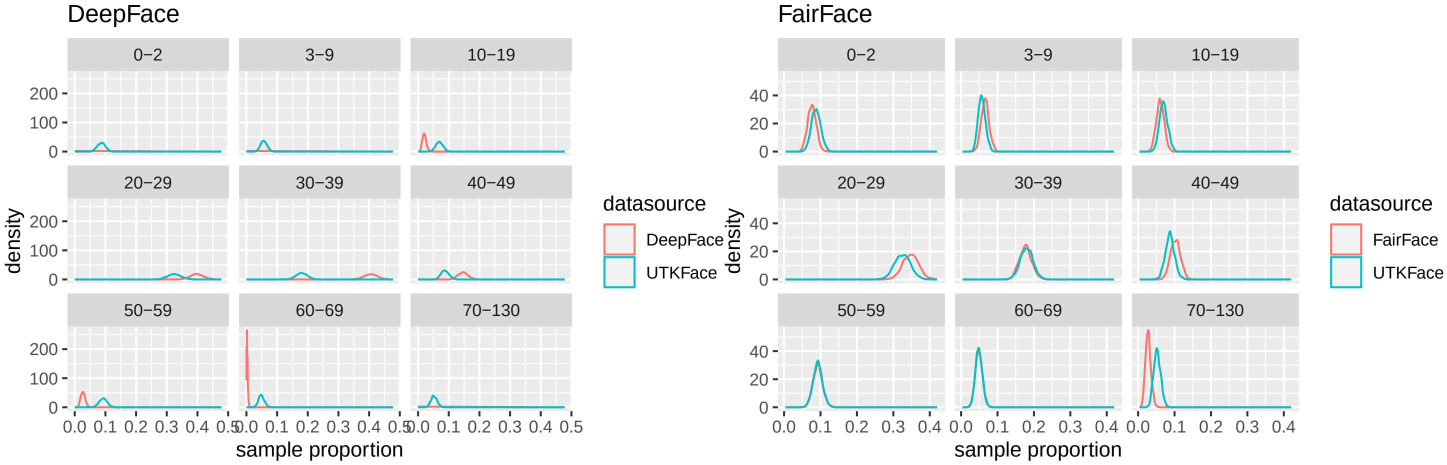
(b) Matrix for DeepFace

Figure 4.6: Confusion matrices of null rejection decisions.

4.4 Population Estmate Plots - UTK Face vs. Model

We used a resampling technique to produce estimated population proportion distributions for each sample. Each resampling included 2000 samples of 500 subjects under their respective test conditions.

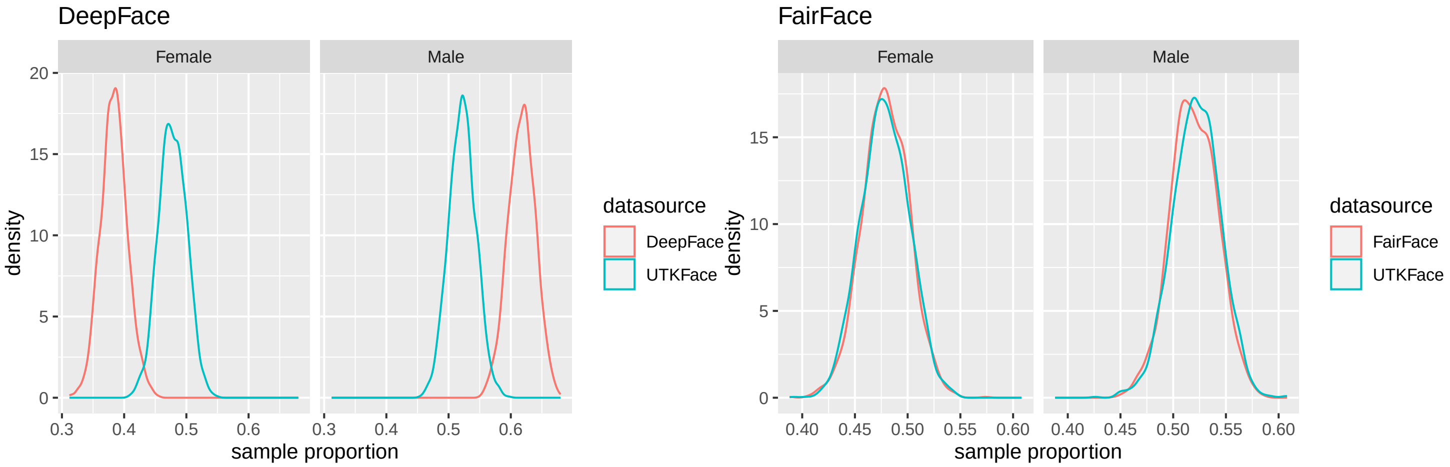
To support our analysis and conclusions, we leveraged a resampling technique (bootstrap sampling) to build approximations of each sample's parent population. The resampling took 2000 samples of 500 random subjects, with replacement, to build the estimated distribution of proportions in the population under specified test conditions. The plots can be seen in Figure 4.7 to Figure 4.9. We find that these plots coincide with our hypothesis testing results – namely, that higher p-values result in greater overlap between the predicted and actual distributions, and lower p-values result in less overlap between the distributions. As such, these distributions will support us in drawing our conclusions.



(a) DeepFace vs. UTKFace Age Distributions

(b) FairFace vs. UTKFace Age Distributions

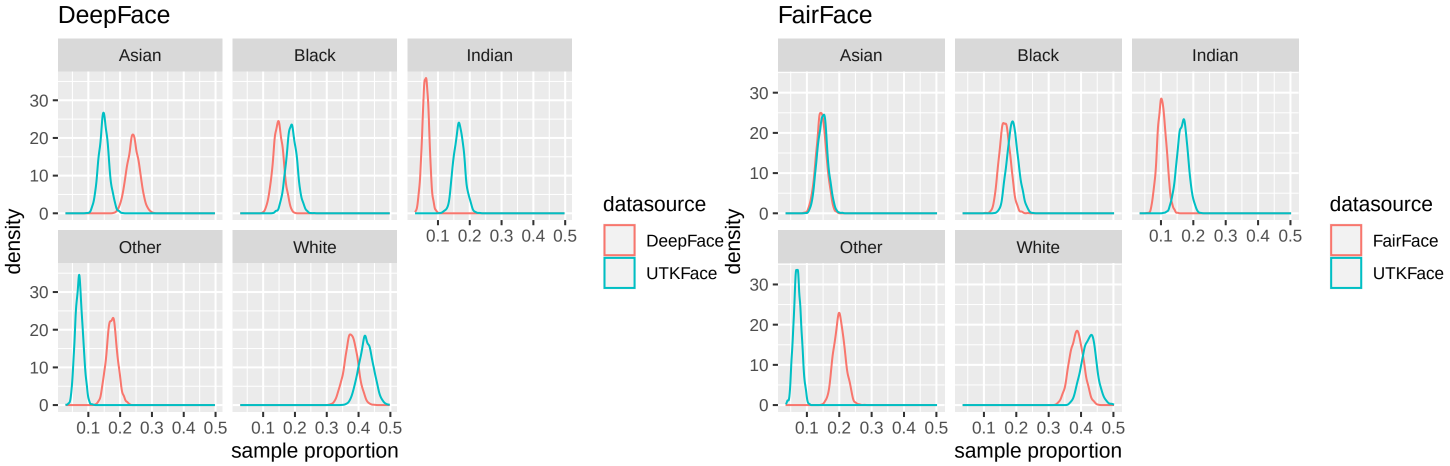
Figure 4.7: Distribution Plots of Age



(a) DeepFace vs. UTKFace Gender Distributions

(b) FairFace vs. UTKFace Gender Distributions

Figure 4.8: Distribution Plots of Gender



(a) DeepFace vs. UTKFace Race Distributions

(b) FairFace vs. UTKFace Race Distributions

Figure 4.9: Distribution Plots of Race

5 Conclusions

5.1 Evaluation of Test Results

To evaluate our tests, we will first examine our hypothesis tests, and then move on to evaluate F1 and Accuracy scores. We theorize that our hypothesis testing, specifically in cases in which we reject the null, may tell us where bias may exist in our data. Separately, F1 and Accuracy scores may tell us specific instances where bias exists in favor of, or against, specific protected classes. Throughout this section, we will use the language "potential bias" for any scenario in which we reject the null hypothesis.

5.2 Hypothesis Testing Results

The design of our hypothesis testing provides us with cases in which datum from the source population differs from that of the predicted population of each model. This may show where biases exist. When the hypothesis test result produces a value less than 0.003 and with test power greater than or equal to 0.8, the test could be indicative of bias. Inversely, a p-value greater than or equal to 0.003 will not provide sufficient evidence to indicate bias in the given test case. The p-value alone, however, cannot tell us whether the indicated bias is in favor of, or against, the protected class group(s) in question. This is because the the hypothesis tests only tell us the probability that the source and predicted results come from the same population.

5.2.1 Age Prediction

5.2.1.1 FairFace

FairFace displays potential for bias in age prediciton with subjects in the age ranges 0-29, 40-49, and 70-130 For age proportions, given gender, potential biases manifest solely for Females in age ranges 0-19, all subjects 20-29, solely for Males 40-49, and all subjects 70+. For age proportions, given race, potential biases are most prominent against the Other, Asian, and White categories, between them spanning age groups from 0-69.

5.2.1.2 DeepFace

DeepFace displays extreme potential for bias in age prediction for subjects in all age groups, with incredibly small p-values. This holds true for all age groups, given any gender. Age given race, however, potential bias does not manifest for 20-29 and 40-49, given they are Black or Indian, and for 50-59 year olds, given they are Asian or Other.

5.2.2 Race Prediction

5.2.2.1 FairFace

FairFace demonstrates potential biases in predicting race for Black, Indian, White, and Other categories. Given either gender, this holds true only for Black, Indian, and Other categories, with more specific potential biases for White, given they are male. Interestingly, given age, potential biases manifest for Asians, given they are 30-39, with little to no changes for the other races across every age category.

5.2.2.2 DeepFace

DeepFace demonstrates potential biases in predicting race for all racial groups, absent other test conditions. Given gender, the potential remains for all categories except for Black subjects, given they are male. Given age, the potential remains for all racial categories, given they are between the ages of 10 and 69.

5.2.3 Gender Prediction

5.2.3.1 FairFace

We lack sufficient information in our hypothesis tests to conclude that FairFace, absent other test conditions, presents potential bias in correct prediction of subject gender. Given race, potential bias manifests for any gender, given they are White or Other. Given age, they manifest for every gender, given age ranges of 0-2 and 20-39, and 60-130.

5.2.3.2 DeepFace

DeepFace, absent other test conditions, indicates potential bias in prediction of either gender. This potential remains, given race. Given age, the potential only remains in the age range of 10-69.

5.3 Identifying Specific Biases with F1 and Accuracy Scores

When examining p-values for potential areas of bias, our hypothesis testing results did not well-align with our F1 score calculations. E.g. a rejection of the null hypothesis did not directly translate to a low F1 score, with the inverse also being true. We proceeded to examine F1 scores, separate of p-value and power results from our hypothesis tests.

General trends for both models: many categories and sub-categories of protected classes fail to meet our selected definition of excellence (F1 score of 0.9 or more). FairFace had more results meeting our definition of excellence compared to DeepFace. Both models demonstrate preference in classification for specific age groups, races, and genders, and both seem to display biases against Indian and Other racial categories. Examining a particular class of subjects, given additional controlling variables, reveal nested biases for and against various classes.

5.3.1 Age

When examining the results of the F1 scores for age, no categories for DeepFace met our specification for excellence. This identifies potential points of improvement in age categorization on part of DeepFace. As DeepFace is unable to detect faces between the ages of 0-9 and 70-130, there is a bias against very young and very old faces. Additionally, The group with the highest F1 performance is 20-29, implying a favorable bias towards subjects in early adulthood.

FairFace's overall age calculations, absent other conditional variables, failed to produce any category that met our F1 threshold, implying lack of excellence in correct predictions for any one age group. However, the categories that did perform the best had a preferential bias towards the very young and very old faces, almost in opposition to DeepFace; FairFace displayed a preferential bias towards the ages of 0-9 and 70-130. When examining specific sub-categories, FairFace presented notable favorable bias to identify male faces between the ages of 0-2 in the White, Asian, and Other categories, as only those categories passed the F1 threshold.

These findings suggest that, were either model to be used in a decision-making process, it could result in age discrimination in accordance with United States Laws. Both FairFace and DeepFace have challenges with various age groups over 40 years old. In the case of employment, the Age Discrimination in Employment Act of 1967 (ADEA) protects certain applicants and employees 40 years of age and older from discrimination on the basis of age in hiring, promotion, discharge, compensation, or terms, conditions or privileges of employment. If used in law enforcement, DeepFace could misidentify an individual as being a suspect, if the suspects specifications fell in the age range ()

5.3.2 Race

Compared to age results, race performs significantly worse for both DeepFace and FairFace, due to the fact that no racial category on its own reaches our F1 threshold. Both models show preference for certain races. In order of preference, DeepFace shows a preferential bias for classifying White, Black, and Asian faces, and FairFace shows a similar bias for classifying Asian, Black and White faces. Indian and Other faces perform the worst overall for both models, with significantly lower F1 scores than the preferred categories, by at least 0.2 for FairFace and 0.3 for DeepFace. As such, these preferences are substantial.

In terms of race with additional control variables, DeepFace demonstrates exceedlingly poor performance. No category for race given age scores surpassed our F1 threshold. Overall, White faces score the highest, provided the identified faces are not 0-9. For FairFace, the only noted bias was a preference for Asian faces younger than 20, and White faces in the ranges of 0-9 and 60-130. For gender-specific biases, there are also no categories that meet or surpass our F1 threshold, but it should be emphasized that DeepFace identified female faces for all races better than male faces. FairFace had a similar performance, except for Indian faces, where male faces scored above female ones.

5.3.3 Gender

Gender shows a similar pattern as race for overall evaluation. DeepFace fails to have any category meet or exceed our F1 threshold, but male faces do show a slightly higher score than female ones. FairFace had both male and female faces score above 0.9, showing a notably positive performance, with little to no difference between males and females.

DeepFace did show preference for certain genders given age, with the range of 30-69 performing above 0.9 for male faces, but only females age 20-29 were significant. This implies a positive bias towards identifying older male faces, as well as bias towards younger adult women. FairFace was more balanced, with significant scores for most age groups except for females age 0-2, and males 0-9. This showcases a negative bias against very young people in general, and particularly male children. For Gender given race, DeepFace had no statistically significant f1 scores, but did show a positive bias towards White faces, and negative biases towards Asian faces of all genders, and

Black female faces. FairFace was far better in all categories, with f1 scores over 0.9 for all categories except Asian male faces. Therefore, it shows a significant negative bias against identifying Asian male faces.

5.4 So What?

Respective of our statistical hypothesis tests and bootstrap population proportion distributions, we can conclude that the models and the source data, in many cases and conditions, have differing source populations, as listed in the earlier portions of this section. Such differences in source population do not appear to be directly indiciative of bias against protected classes.

Some specific examples, but not all examples, include:

- FairFace
  - The proportion of 60-69 year olds, with a p-value of nearly 1, but with an F1 score of 0.354, and an accuracy score of 0.671. This shows that high p-values do not translate to high F1 scores for FairFace.
  - The proportion of Females, given they are "Other" for race, with a p-value of $2.58 \cdot 10^{-13}$, with an F1 score of 0.922, and accuracy of 0.914. This shows that low p-values do not translate to low F1 scores for FairFace.
- DeepFace
  - The proportion of 20-29 year olds, given they are Black, with a p-value of 0.10, with an F1 score of 0.588. This shows that a high p-value does not directly translate to a high F1 score for DeepFace.
  - The proportions of males age 40-49, 50-59, 60-69 with respective p-values of $2.01 \cdot 10^{-12}$, $9.43 \cdot 10^{-12}$ and $1.23 \cdot 10^{-5}$, but holding respective F1 scores of 0.946, 0.937, and 0.928. Showing that low p-values do not necessarily indicate a low F1 score for DeepFace.

However, the cases in which we reject the null hypothesis could be indicative that the training data for the facial recognition models, and the data we provided to them from UTKFace, have little to no overlap between one another (in terms of features and qualities of the images). This could be a topic for further research - i.e. are the differences in source population a result of feature differences between a model's training data and the models classification predictions on novel images?

Absent further research, with Accuracy and F1 scores accepted as best practice - the results of our two-sample hypothesis tests can only truly tell us that there is a difference in the source populations for each of our samples.

On the examination of F1 scores, we find that

- FairFace displays preferential biases in classifying:
  - correct race, given the subject is male, in almost every racial group.
  - Asian race, given they are under 20.
- DeepFace displays preferential bias in classifying:
  - correct age, given the subject is White, for nearly every age group.
  - correct gender, given the subject is generally an adult (30-69)

These preferential biases can result in discrimination:

- FairFace combination of racial prediction given a male subject could result in a combination of racial discrimination against Indians, and gender discrimination against females, if used in decision-making processess.

- DeepFace's combination of correct age given White as race could result in a combination of racial and age discrimination if used in decision-making processes.

5.5 Areas for Further Research

- As mentioned previously - Are differences in source populations between an evaluated dataset and a facial recognition model a result of feature differences between a model's training data and the model's classification predictions on novel images?
-

5.6 More Conclusions, for the Mathematically Inclined

F1 and Accuracy scores are generally accepted as best practice in evaluating the efficacy of machine learning models. From our tests, we saw contradictions between two-sample proportion tests and F1/Accuracy scores with respect to each model. This is directly evident from Figure 4.4 and Figure 4.5 with a clear lack of correlation of any type between the variables, for all our 432 hypothesis tests.

We can examine this further. An Accuracy or F1-score of 0.9 is a reasonable threshold for an "excellent" peforming model. We could set this this threshold as analogous to the outcomes of a hypothesis test. If a model is performing well, we would expect there wouldn't be enough evidence to reject the null hypothesis (i.e. equal proportions between the source and model could not be statistically rejected). If a model is is not peforming well, we would expect there to be enough evidence suggesting we should reject the null hypothesis in favor of the alternative hypothesis (i.e. there was enough evidence the proportions between the source and model were not equal).

In that perspective, if we assume that the sample outputs' F1 scores should reject the null hypotheses when below a certain threshold, and fail to reject when above 0.9, we can build a confusion matrix of "prediction" to reject or fail to reject the null using two-sample proportion tests, in comparison to a "correct" result using sample out F1 scores. We should use this same threshold, as it's the same that we set for each model in evaluating protected classes.

Pursuing such an evaulation is an appropriate approach, because the methods we've leveraged for attempting to examine bias using proportionality testing is a model, just as classification of inputs and outputs using confusion matrices is a model. A standard method of evaluating model performance is via confusion matrices.

Such matrices produce the following results when evaluating our sample outputs:

model	accuracy	F1	threshold
DeepFace	0.4454369	0.5514563	0.9
DeepFace	0.6257857	0.4255319	0.9
DeepFace	0.3886014	0.0124224	0.9
FairFace	0.5198658	0.5207373	0.9
FairFace	0.5223073	0.5174825	0.9
FairFace	0.5000000	NA	0.9

Assuming that a correct decision to reject or fail to reject the null should be based upon an F1 and Accuracy scores at multiple thresholds (0.9, 0.8, or 0.7), we see substantially low accuracy and F1 scores for two-sample proportionality tests as a model for predicting machine learning model performance. This highlights the contradictions we witnessed in our results for two-sample proportion tests vs. leveraging accuracy and F1 scores. Given These results, we find that two-sample proportionality testing is likely not a strong indicator to identify issues and errors in machine learning models.

References

Buolamwini, Joy. 2023. "Gender Shades: Intersectional Accuracy Disparities in." *MIT Media Lab*. https://www.media.mit.edu/publications/gender-shades-intersectional-accuracy-disparities-in-commercial-gender-classification.

Georgetown Law. 2016. "The Perpetual Line-Up: Unregulated Police Face Recognition in America." *Center on Privacy & Technology*. https://www.perpetuallineup.org.

Huilgol, Purva. 2021. "Accuracy vs. F1-Score - Analytics Vidhya - Medium." *Medium*, December. https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2.

Karkkainen, Kimmo, and Jungseock Joo. 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*, 1548–58.

Lohr, Steve. 2018. "Facial Recognition Is Accurate, if You're a White Guy." *N.Y. Times*, February. https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html.

NIST. 2020. "NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software | NIST." *NIST*. https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software.

Serengil, Sefik Ilkin, and Alper Ozpinar. 2021. "HyperExtended LightFace: A Facial Attribute Analysis Framework." In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, 1–4. IEEE. https://doi.org/10.1109/ICEET53442.2021.9659697.

"UTKFace." 2021. *UTKFace*. https://susanqq.github.io/UTKFace.