Bias in Facial Classification ML Models

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

2023-12-10

Table of contents

Abstract

In this study, our research team examines the performance of two facial recognition models, FairFace and DeepFace, for potential biases against the protected classes of age, gender, and race. Our objective is to determine whether the source data and model outputs originate from the same population and determine if any such difference is indicative of bias in one or both models. We employ two-sample proportionality tests to evaluate the parent populations for the input data and each model's output in terms of gender, race, and age, incorporating model F1 and Accuracy scores in tandem with our test results to evaluate presence of bias. Our goal in this effort is to contribute to ongoing research and discourse on ethics in computing and machine learning, and to share any insights or significant findings we encounter. Further details on our source data and models, methods, results, and implications are detailed in the subsequent sections of this paper.

Report PDF and Code Location

A link to download the PDF version of this report, and a link to the Github source code for this report, are both available as icons in the top right corner of this website.

How we should write this report TODO: REMOVE

- See Karkkainen and Joo (2021) , that is an example on how to cite a bibliography.
- Sections/title headings are automatically numbered.
- Any changes you make, make sure to make a comment of your initials at the top of your work (INCLUDING written text) like so:

```
<!-- BJ !-->
Blah blah etc ....

OR
#BJ
r_var <- ...
```

- Make sure to add a unique name to all code cells, and to also enable the following (the quarto way) (In order for a figure to be cross-referenceable, its label must start with the fig- prefix):

- You can then refer to figures like this `@fig-sec1-unique-name` Figure 1

- Format tables doing the following Link here

- Do all your r work initially in your own custom `.rmd` file in this directory, so that it can be copy-pasted over later into the appropriate section (written descriptions/words can go straight into the `.qmd` files though). For example, Bhav's work is in `5000-final/BJ_work.rmd`.

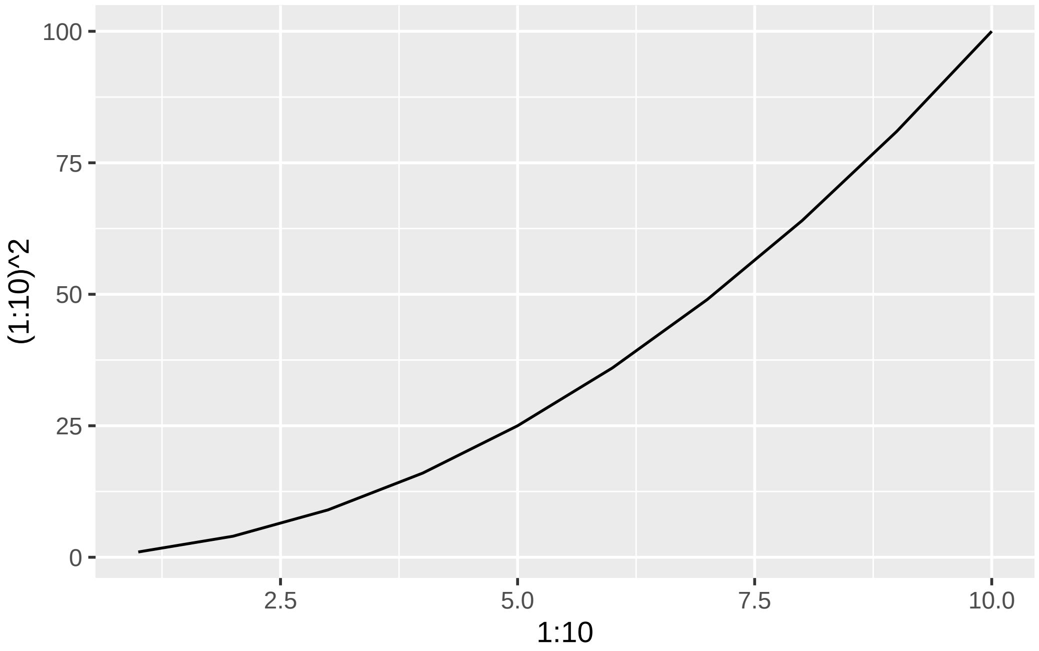


Figure 1: A caption for generated figure

ℹ From the report requirements

A 3-5 summary of the paper. It should address the research question, the methods, and the conclusions of your analysis.

“A good recipe for an abstract is: first sentence: specify the general area of the paper and encourage the reader; second sentence: specify the dataset and methods at a general level; third sentence: specify the headline result; and a fourth sentence about implications.”

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. It was trained on a balanced dataset that featured equal numbers of different races, including Middle Eastern faces. The creators point out in their work that the majority of other training data is overwhelmingly white and male, lending those algorithms a bias (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. The creators have shown an accuracy rate of up to 97% on at least one dataset, but replication is key when it comes to testing code (Serengil and Ozpinar 2021).

It is this paper's goal to test the accuracy of those claims and compare these algorithm's ability to guess the race, gender, and age of a given dataset, both with and without controlling for the race of the given faces. Both will be tested against the UTKFace dataset, which consists of over 20,000 labeled faces that can be used for purposes such as this ("UTKFace" 2021). We will examine the bias present in either tested model using hypothesis testing and by inspecting performance metrics such as F1 score and accuracy, and compare this bias against the baseline (aka, the bias inherently present in the UTK face dataset).

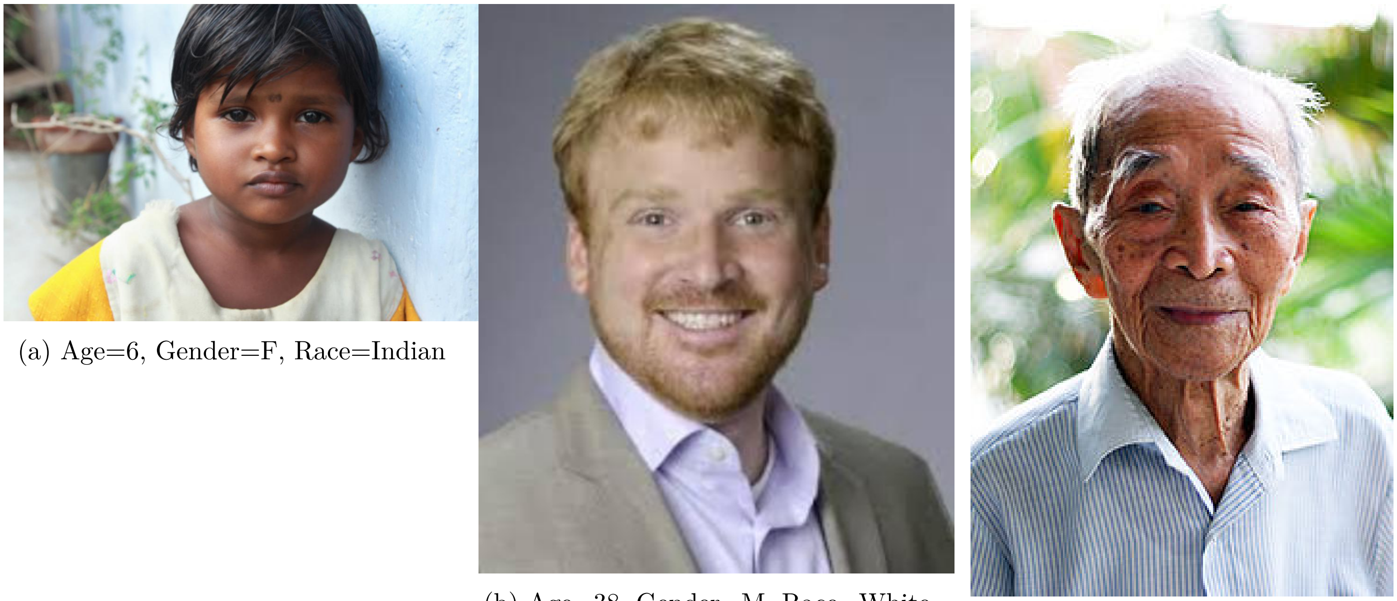
2 Data

Pursuant to the study, the team sought out multiple datasets on which we could evaluate the performance of two selected recognition models (Facebook's DeepFace and K. Karkkainen & J. Joo's Fair Face models) 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: https://susanqq.github.io/UTKFace/

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 Exploration of Source Data



(a) Age=6, Gender=F, Race=Indian

(b) Age=38, Gender=M, Race=White

(c) Age=80, Gender=M, Race=Asian

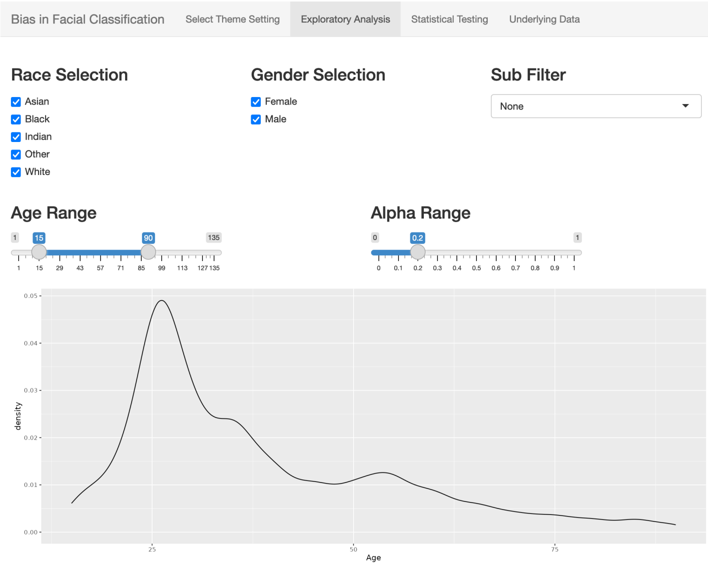
Figure 2.1: Example face images from the UTK dataset ("UTKFace" 2021) with their associated given labels.

From the report requirements

This section should describe the data you'll be using. Answer **at least all** of the following questions:

- How was the data collected?

The dataset used in this research is a publicly non-commercial available dataset on Github called "UTKFace". The data was collected by the The University of Tennessee, Knoxville. It is specified on its Github page that the images were gathered from the internet. They are likely to be obtained through technique such as web scrapping. The dataset contains more than 20,000 images, representing a highly diversed demographic. However, face images are vary in pose, facial expression, lighting, and resolution.



Race Selection

☑ Asian
☑ Black
☑ Indian
☑ Other
☑ White

Gender Selection

☑ Female
☑ Male

Sub Filter

None

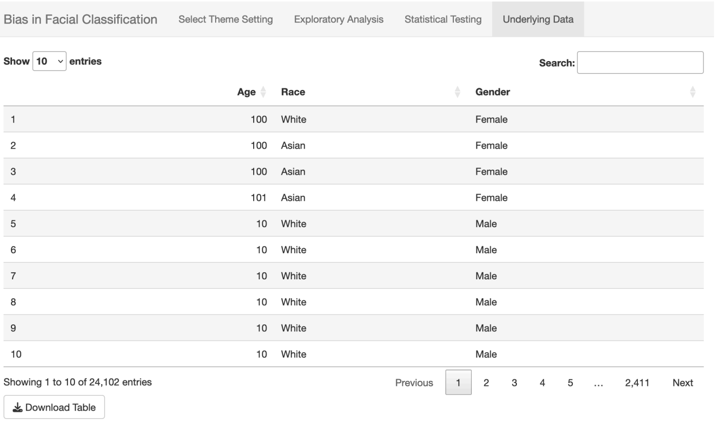
Age Range

15 – 90

Alpha Range

0.2

Figure 2.2: Image data EDA



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

Figure 2.3: Image dataset visualization

Figure 2.4: 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

- What are the sources and influences of bias in the data?

The distribution of each demographic groups are not normally distributed. By plotting a distribution of each demographic group, it is evident that the dataset contains an uneven high volume of older White men. While a smaller porportion of female among ...(input race)... is present.

- What are the important features (=columns) that you are using in your analysis? What do they mean?

There are three features in the dataset which are essential to our analysis. They are Race, Gender, and Age. Race is categorized into five groups; Asian, Black, Indian, White, and Other. It should be noted by Asian group in this dataset mostly refers to people from East and Southeast Asia. Whereas, Other includes ethnicities such as Hispanic, Latino, and Middle Eastern.
Gender is divided into two groups, either male or female.
Lastly, Age is represented with an integer. This dataset contains people of all ages ranging from 0 to 116.
Feel free to add anything else that you think is necessary for understanding the paper and the context of the problem.

TODO: Not sure where this goes

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.

2.2 Data Selection (UTK Dataset) - LN DV

2.2.1 Motivation for the Selection of UTKFace Dataset

In 2018, Joy Buolamwini paper link, a PhD candidate at MIT Media Lab, published a thesis on gender and racial biases in facial recognition in algorithms. In her paper, she tested facial recongition softwares from multiple large technology companies such as Microsoft, IBM, and Amazom 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 ethinicty is required. Upon searching, UTKFace came out as the largest dataset which fit the preferred qualifications.

2.2.2 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 scrapping. The dataset contains more than 20,000 face images, representing a highly diversed demographic. However, face images are vary in pose, facial expression, lighting, and resolution.

2.2.3 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.

Overall, all of 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.

2.3 Selected Models (LN DV)

2.3.1 FairFace

Developed by researchers at University of California, Los Angeles, FairFace was specifically designed to mitigate gender and racial biases. The model 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 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.

2.3.2 DeepFace

DeepFace is a lightweight open-source model developed and used by Meta (Facebook). Since the model is used by one of the largest social media company, it is widely known among developers. Therefore, its popularity prompts us to evaluate its performance. It should be noted that this model of DeepFace is a free open source version. It is highly likely that this version is less advanced than what Meta is actually utilizing. Thus, we should not view the result of this model as a representative of Meta's algorithm.

2.4 Dataset Features

Understanding and selecting the appropriate features for our data is key to success in analysis. Furthermore, understanding feature differences and planning standardization across datasets is key in making sound comparisons and analyses upon the dataset.

2.4.1 Input data set

The input dataset, being UTKFace, 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 in the UTKFace Readme:

- "[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"

We processed each image to extract these features from each image and create a table of source information (link to source data file here).

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.

2.4.2 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 import 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 key 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 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_fair	String	The predicted race of the image subject	[White|Black|Latino_Hispanic|East Asian|Southeast Asian|Middle Eastern|Indian]
gender_preds_fair7	String	The predicted gender of the image subject	[Male|Female]
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+’]

2.4.3 DeepFace Outputs

Default outputs have 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 the purpose of our planned analyses, the score features are of less concern to us.

We hone in 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]

Column Name	Data Type	Significance	Valid Values
Dominant Race	String	The predicted race of the image subject	[middle eastern|asian|white|latino hispanic|black|indian]

2.5 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 - the Asian Face Age Dataset (**need citation here**). 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 what settings may provide the most fair and equitable comparison of the models prior to proceeding to further analysis.

The ultimate 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. In lamens terms, 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.

2.5.1 DeepFace Analysis Options

DeepFace has a robust degree of avaialble 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 exeuction. 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.

2.5.2 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 pre-processing, 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, similar to 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.

2.5.3 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 datset. 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 conlcusion 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 came to the conclusion 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 excluded 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 MasterScript.py that allowed us to apply multiprocessing across the list of images and evaluate all items in a reasonable amount of time. (cite FairFace here, may also need to reference that we rebuilt their script into a class format).

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 all of 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.

2.6 Model Evaluation Data

The final listing of all inputs and outputs from each model, with standardization methods discussed in this section applied, are stored here: MasterDataFrame.csv

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 section 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

Our EDA performed on the source UTK dataset can be seen in the previous section in **?@fig-data-eda**. 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 bias prevalent in facial recognition machine learning models?
- Can one model be shown to have statistically significant less bias than the other?
- Does one model outperform the other in a statistically significant manner, in all aspects?
- Does one model outperform the other in a statistically significant manner, in certain aspects?
- Are there disparate outcomes (i.e. lower chances of correct classification) 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 hpyothesis will be no difference in means. 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 a bias in a model. We use a significance level of 99.7% to accurately judge the strength of the test statistic.

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) in cases of inconclusive results from p-value and power should assist us in identifying potential error.

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

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
- $\hat{p_2} = p_{F,D_f}$: 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 providing more, or fewer, correct 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 favor of correctly predicting race, given the subject is young." - which also suggests a bias against correctly predicting the race of older subjects in a model. Such a bias, if used in a decision making process, could result in age discrimination.

"Model Y demonstrates bias against predicting correct gender, given the subject image is Black, Asian, or Other." - which also suggests a bias for correctly predicting gender, given the subject is White or Indian. Such a bias, if used in a decision making process, could result in 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.

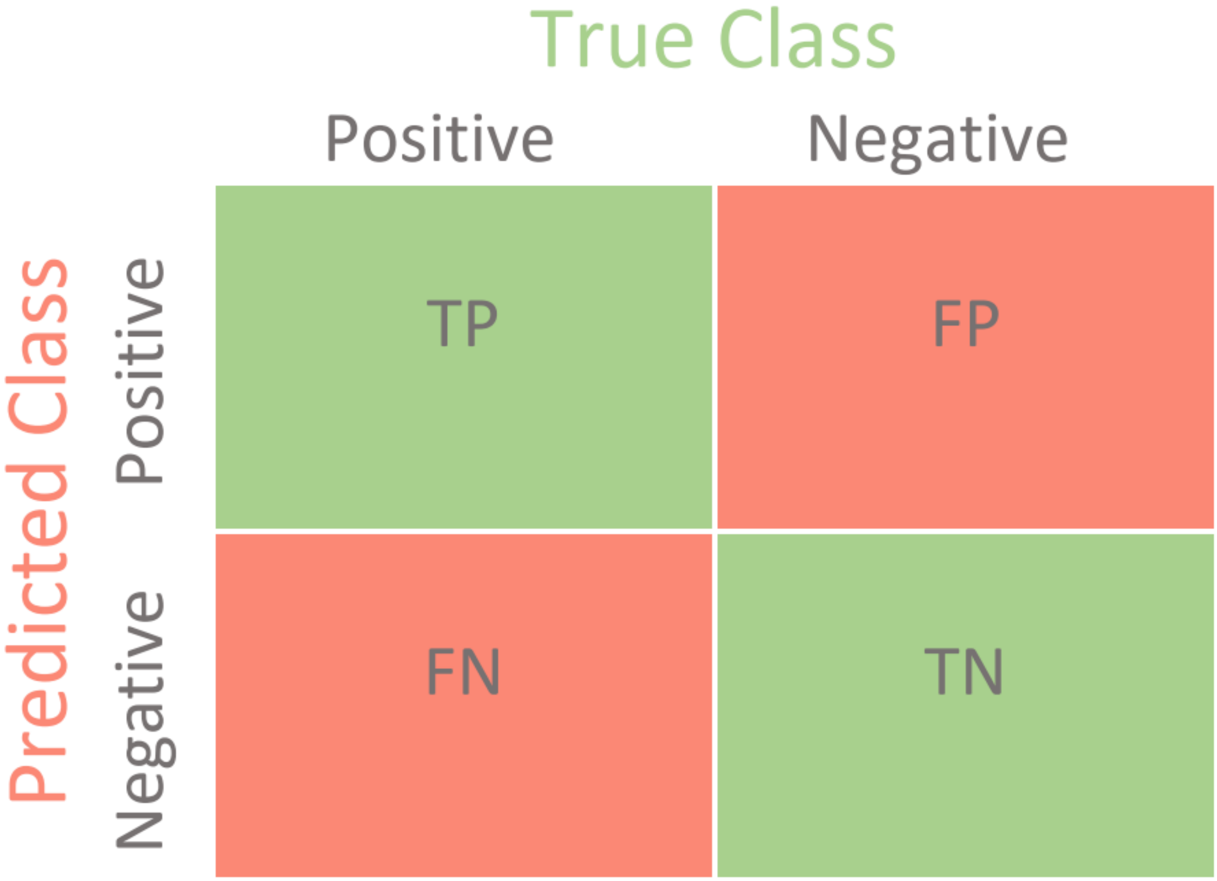
There are four main measures of performance when evaluating a model:

- **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.



		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

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.

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

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.

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

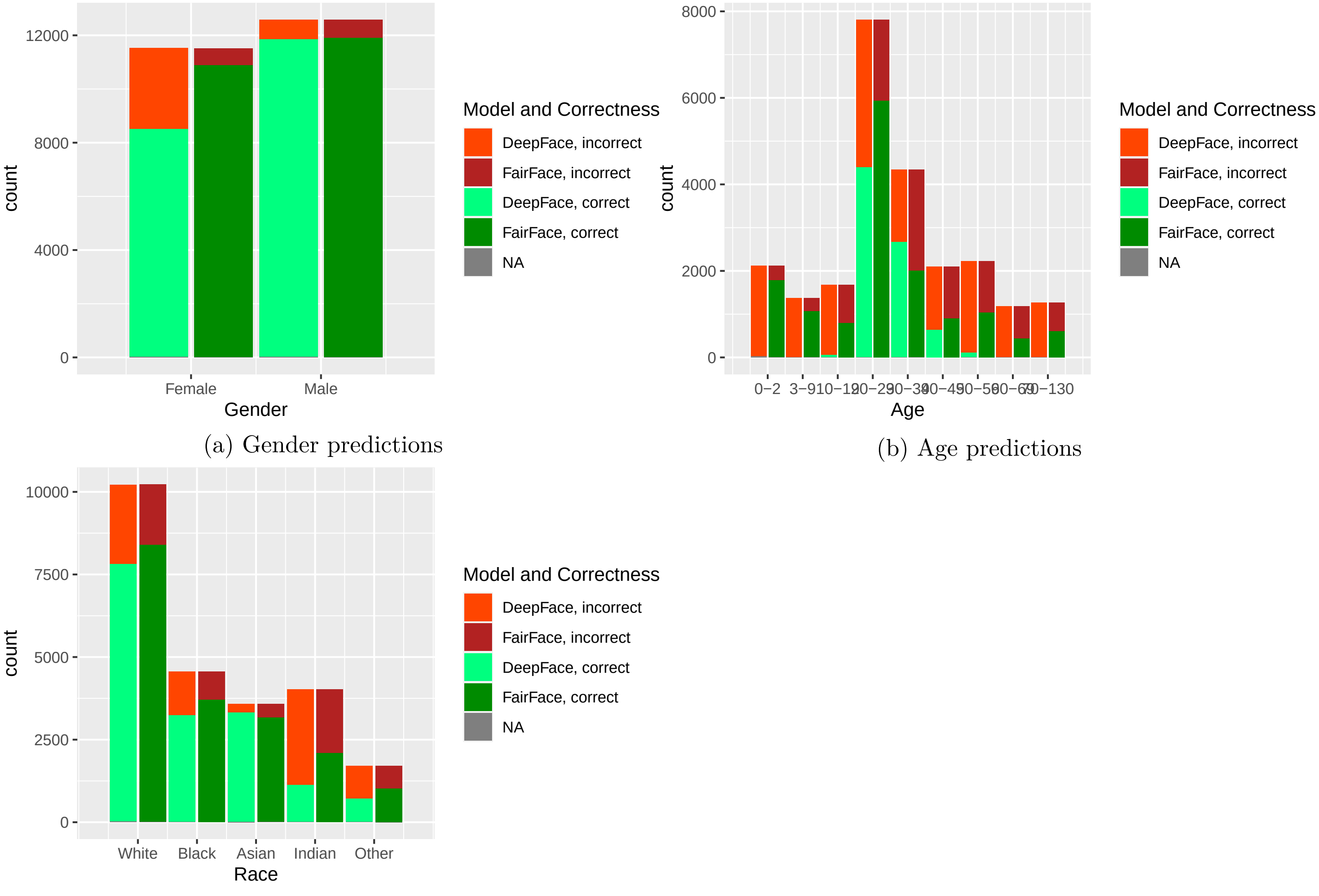
When considering the classification of a subject by protected classes of age, gender, and race, we believe that stronger penalties should be assigned in making an improper classification decision. Due to F1 being the harmonic mean of precision and recall, incorrect classification will more directly impact 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 scores to assess p-values in cases of low statistical power should assist us in identifying potential Type-II Errors. We set a 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.

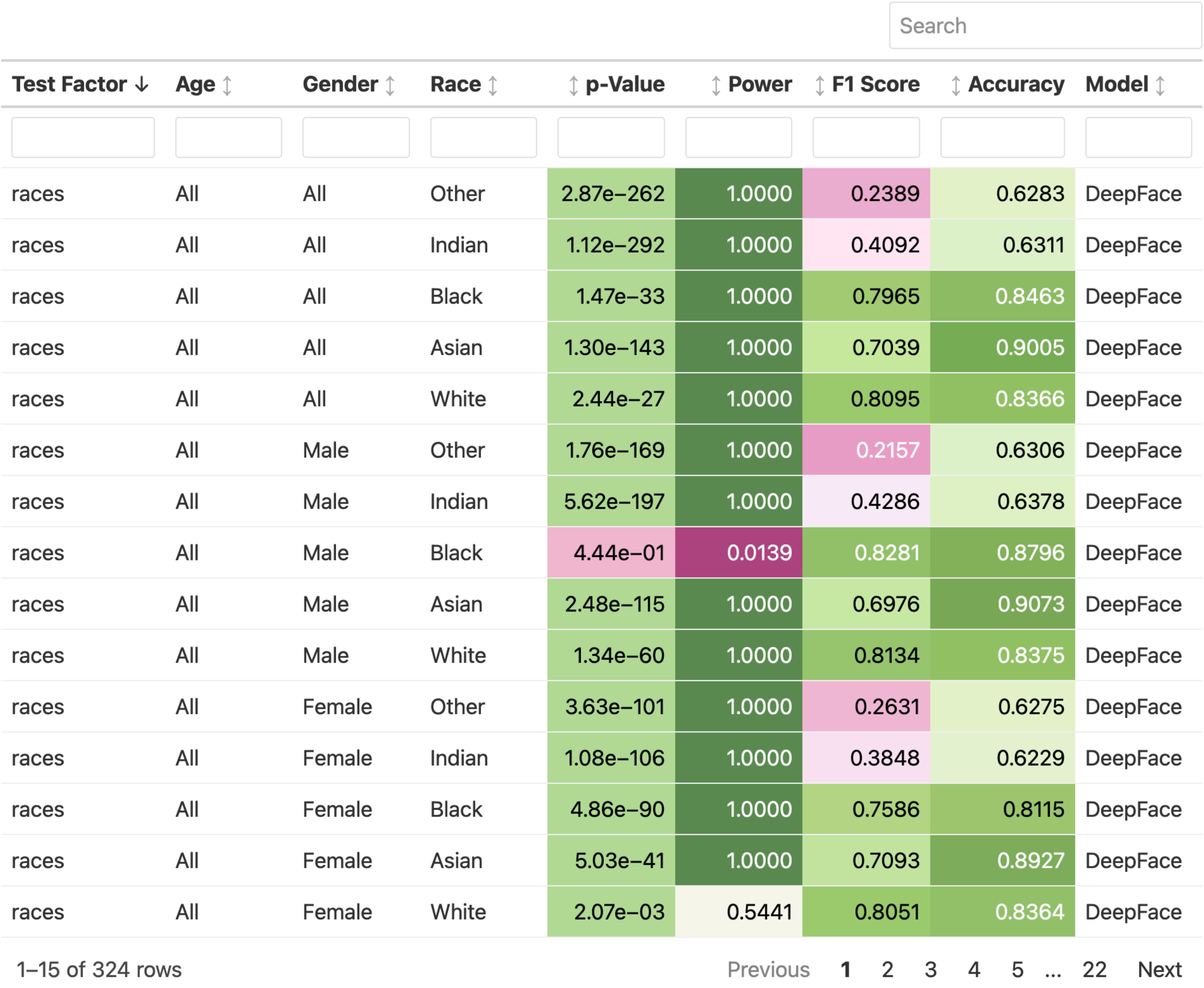
4.2 Model Performance, Hypothesis Testing

For each category and model, we calculate the F1 score, accuracy, and p-value, as described in section 3. Cell values are colored according to the strength of the metric. We also specifically looked at the performance metrics of the models, when controlled for specific source categories; TODO

4.2.1 TODO

- Add color key
- Better description of signifiance of numerical values of Accuracy, F1 score, power

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

Figure 4.2: Screenshot of the interactive table showing TODO. To see and interact with this table, go to the website link

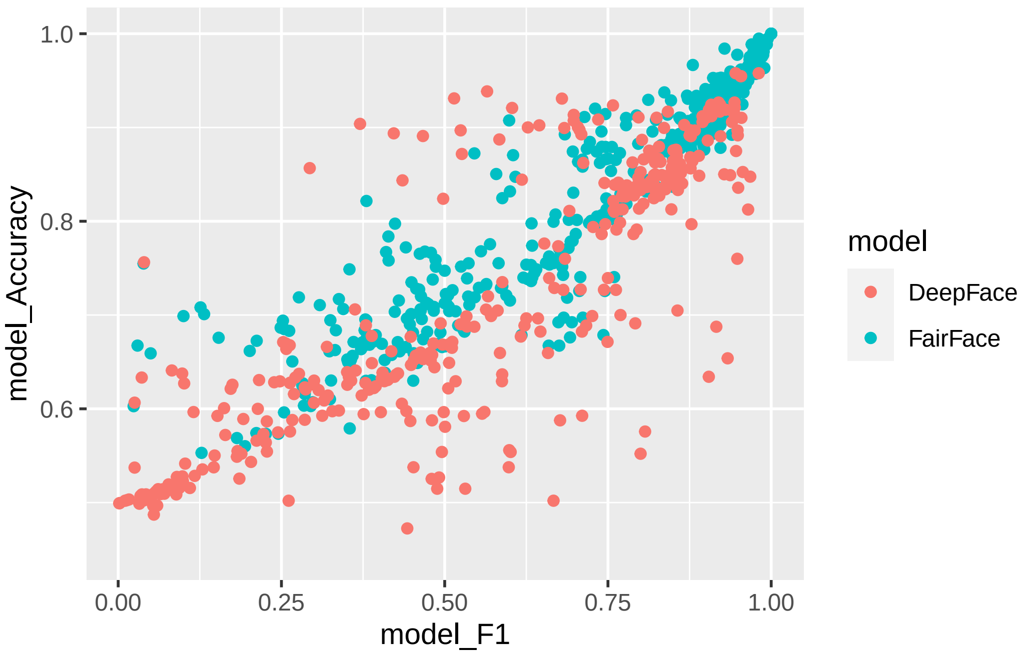
4.2.2 p-value Critical Values

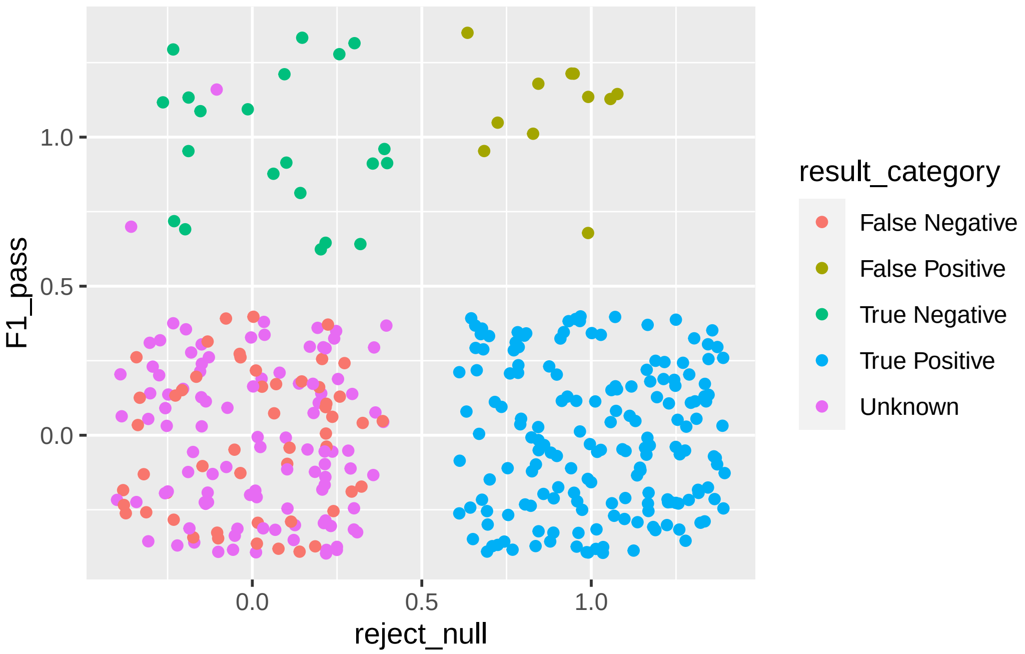
From the previous table, we extract and highlight key values; namely, the p-values where we reject the null hypothesis at a significance level of 99.7%, and where there is no conditional filtering based on another category. The values are displayed in Table 4.1.

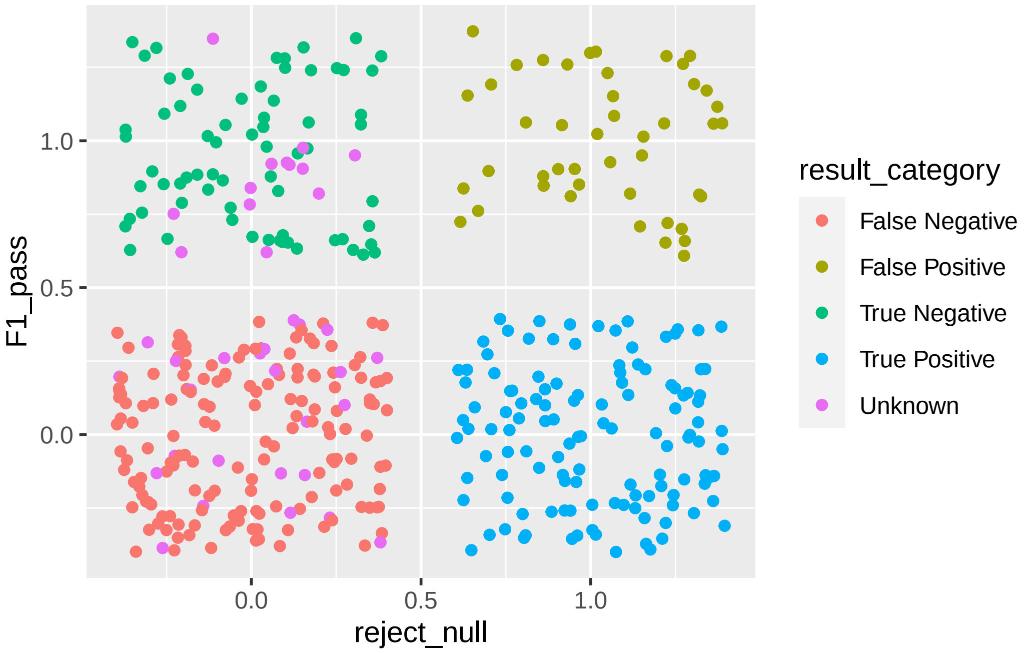
Table 4.1: TODO

test_prop	test_subcat	model_p_value	model_power	model_F1	model_Accuracy	model
ages	70-130	2.825510e-43	1.0000000	0.627066116	0.7383514	FairFace
ages	3-9	1.365839e-05	0.9197828	0.717603501	0.8772778	FairFace
ages	10-19	5.220244e-05	0.8640459	0.505249761	0.7211461	FairFace
ages	0-2	3.114548e-06	0.9568115	0.895975702	0.9172888	FairFace
ages	20-29	2.138176e-08	0.9958932	0.733292156	0.8050592	FairFace
ages	40-49	1.650236e-08	0.9965034	0.394397023	0.6786302	FairFace
races	White	5.830310e-18	1.0000000	0.861039893	0.8788455	FairFace
races	Black	7.462843e-12	0.9999512	0.868485771	0.8997692	FairFace
races	Indian	8.841041e-94	1.0000000	0.640245776	0.7488102	FairFace
races	Other	0.000000e+00	1.0000000	0.308747333	0.7105889	FairFace
ages	70-130	1.081059e-283	1.0000000	NA	0.5000000	DeepFace
ages	3-9	9.200211e-293	1.0000000	NA	0.5000000	DeepFace
ages	10-19	2.520269e-148	1.0000000	0.047860101	0.5055825	DeepFace
ages	0-2	0.000000e+00	1.0000000	NA	0.5000000	DeepFace
ages	20-29	2.003003e-65	1.0000000	0.505432596	0.6217793	DeepFace
ages	30-39	0.000000e+00	1.0000000	0.378631811	0.6275447	DeepFace
ages	40-49	1.647085e-91	1.0000000	0.227627841	0.5866155	DeepFace
ages	50-59	3.663906e-202	1.0000000	0.080167306	0.5137145	DeepFace
ages	60-69	9.807576e-229	1.0000000	0.001612903	0.4991769	DeepFace
genders	Female	1.176698e-97	1.0000000	0.819770248	0.8402892	DeepFace
genders	Male	1.176698e-97	1.0000000	0.863662918	0.8402892	DeepFace
races	White	2.699204e-27	1.0000000	0.809546093	0.8365916	DeepFace
races	Asian	1.748655e-143	1.0000000	0.703897491	0.9005150	DeepFace
races	Black	1.705494e-33	1.0000000	0.796499445	0.8462797	DeepFace
races	Indian	1.897319e-292	1.0000000	0.409248134	0.6310597	DeepFace
races	Other	4.642205e-262	1.0000000	0.238902067	0.6283106	DeepFace

4.2.3 Comparison Plots






F1_pass
1.0
0.5
0.0
0.0
0.5
1.0
reject_null
result_category
False Negative
False Positive
True Negative
True Positive
Unknown

5 Conclusions

From the report requirements

- Summarize what the paper has done, and discuss the implications of your Results.
- Explicitly connect the results to the research question.
- Discuss how you would you extend this research

Like the introduction, this section should be written with a **non-expert** in mind. A person should be able to read Introduction+Conclusion and get a rough idea of the meaning and significance of your paper

References

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.