Bias in Facial Classification ML Models

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

Table of contents

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

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.

Features of Quarto:

How we should write this report

- 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`.

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;

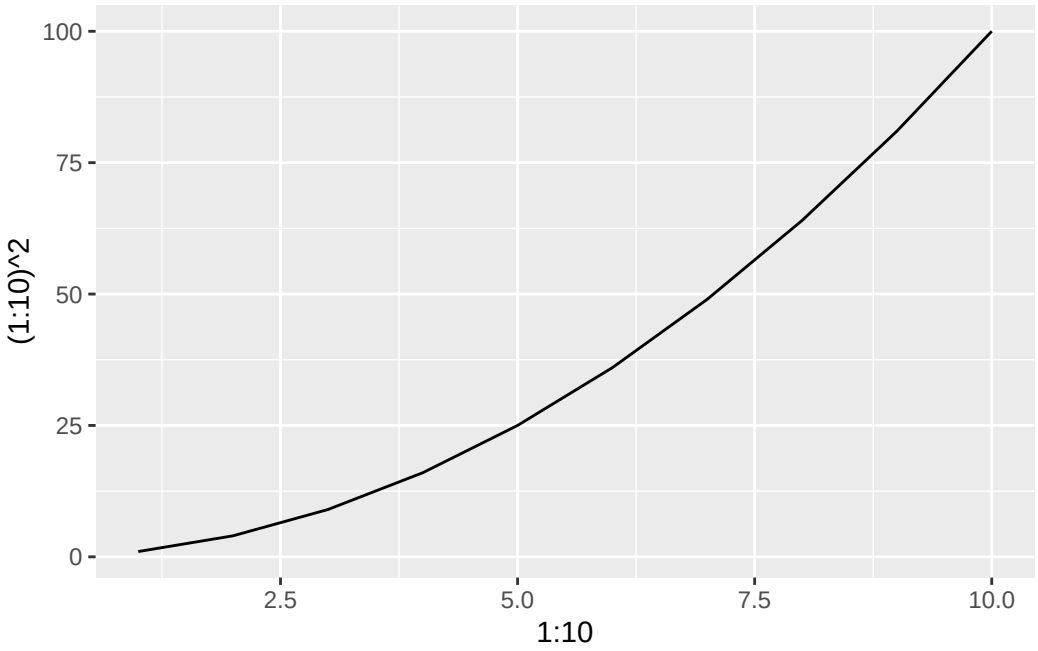


Figure 1: A caption for generated figure

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

From the report requirements

This section introduces your problem to a **non-expert** audience, describes the context and history of the problem.
For example, if your overall project topic is on Diabetes Prevention and Prediction, then you would use the Introduction to introduce what diabetes is, who it affects, why prevention is important, history on diabetes prevention, etc.
Some questions that you could answer in the introduction:

- What is the "research question"? why is it interesting or worth answering?
- What is the relevant background information for readers to understand your project? Assume that your audience
  is not an expert in the application field.
- Is there any prior research on your topic that might be helpful for the audience?

The goal of the introduction it to capture the audience's interest in your paper. An introduction that starts with "Diabetes kills over 87 thousand people each year and in many cases may be preventable" is more engaging than "This paper is about diabetes prevention".
The introduction should be 2-4 paragraphs long.

2 Data

We describe the data here. Note that the default global setting for Quarto is set to NOT output the code into the rendered document, aka only including the results of any R code.

We should include a print of the head of the dataframe of our data, along with some sample 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.

> **i** 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?
> - What are the sources and influences of bias in the data?
> - What are the important features (=columns) that you are using in your analysis? What do they mean?
>
> Feel free to add anything else that you think is necessary for understanding the paper and the context of the problem.

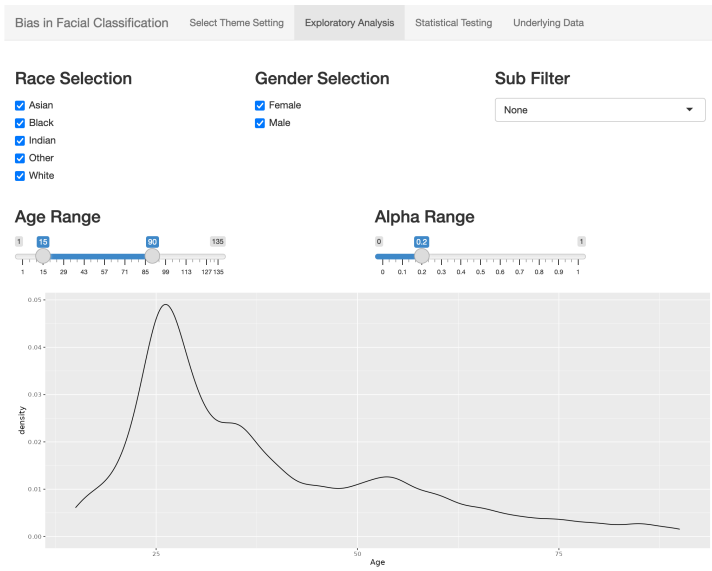


Figure 2.2: Image data EDA

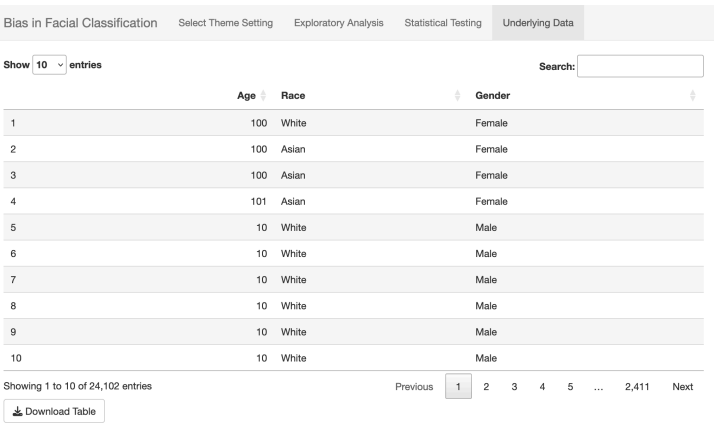


Figure 2.3: Image dataset visualization

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

2.2 More information on data

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.2     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.3     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

filenames	purpose	recommendations
Croped_ff_np.csv	Permutation evaluation (older version) for fairface, no preprocessing on cropped images. Updated this file to look at the same files as the uncropped dataset.	Remove from github data folder.
MasterDataFrame.csv	Final master data file containing all input and output files	Keep as-is with no changes
crop_df_np.csv	Permutation evaluation for DeepFace, cropped images, no pre-processing	Retain; rename to PERM_DF_c_np.csv
crop_df_p_mtcnn.csv	Permutation evaluation for DeepFace, cropped images, preprocessed with MTCNN backend.	Retain; rename to PERM_DF_c_p_mtcnn.cs
crop_df_p_opencv.csv	Permutation evaluation for DeepFace, cropped images, preprocessed with OpenCV backend.	Retain; rename to PERM_DF_c_p_opencv.c
cropped_UTK.csv	Permutation evaluation (older version), list of cropped files to perform evaluation.	Remove from github data folder
cropped_UTK_dataset.csv	Permutation evaluation (newest version), list of cropped files to perform evaluation.	Retain with no changes

filenames	purpose	recommendations
cropped_ff_p.csv	Permutation evaluation (older version), used older version of cropped images dataset.	Remove from github data folder.
joined_permutations.csv	Permutation evaluation (newest version), joined all permutation outputs from DeepFace and FairFace to a single file	Retain with no changes
new_ff_c_np.csv	Permutation evaluation (newest version), FairFaice outputs for cropped images with no preprocessing	Retain; rename to PERM_FF_c_np.csv
new_ff_c_p.csv	Permutation evaluation (newest version), FairFaice outputs for cropped images with dlib preprocessing	Retain; rename to PERM_FF_c_p.csv
new_ff_uc_np.csv	Permutation evaluation (newest version), FairFaice outputs for uncropped images with no preprocessing	Retain; rename to PERM_FF_uc_np.csv
new_ff_uc_p.csv	Permutation evaluation (newest version), FairFaice outputs for uncropped images with dlib preprocessing.	Retain; rename to PERM_FF_uc_p.csv
non_normalized_DeepFace_dataset.csv	Final dataset of DeepFace outputs (non-normalized)	Retain; rename to Master_DF_non_normalized.cs
non_normalized_FairFace_dataset.csv	Final dataset of FairFace outputs (non-normalized)	Retain; rename to Master_FF_non_normalized.cs
uncropped_DF_all.csv	Final normalized output for DeepFace - used to build MasterDataFrame.csv	Retain with no changes
uncropped_FF_all.csv	Final normalized output for FairFace - used to build MasterDataFrame.csv	Retain with no changes
uncropped_UTK.csv	Permutation evaluation (older version) - source data file for iteration script	Remove from github data folder.
uncropped_UTK_dataset.csv	Permutation evaluation (newest version) - source data file for uncropped images in iteration script	Retain with no changes
uncropped_df_np.csv	Permutation evaluation (newest version) - DeepFace uncropped images with no preprocessing	Retain; rename to PERM_DF_uc_np.csv
uncropped_df_p_mtcnn.csv	Permutation Evaluation (newest version) - DeepFace uncropped images with mtcnn preprocessing	Retain; rename to PERM_DF_uc_p_mtcnn.c
uncropped_df_p_opencv.csv	Permutation Evaluation (newest version) - DeepFace uncropped images with opencv preprocessing	Retain; rename to PERM_DF_uc_p_opencv
uncropped_ff_np.csv	Permutation Evaluation (older version) - FairFace uncropped images with no preprocessing	Remove from github data folder.
uncropped_ff_p.csv	Permutation Evaluation (older version) - FairFace uncropped images with dlib preprocessing.	Remove from github data folder.

2.3 Data Selection (UTK Dataset) - LN DV

Motivation - has there been progress against bias?

Reference Articles from Ethics Class - Joy B's work, etc. Lots of disparity back in 2018, how much of that still exists with free models?

2.4 Selected Models (LN DV)

Lorem ipsum

2.4.1 FairFace

Motivation as to how / why we came across this

2.4.2 DeepFace

Motivation as to how / why we came across this

2.4.3 Permutation Analysis Information

3 Methods

Karkkainen and Joo (2021)

3.1 The Big Picture

- 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?
  - This is where we can dive into "conventional" bias

> **Thoughts on Bias**
>
> We need to be careful how we define and use bias. Statistical bias is essentially error, and we could be crossing our definitions between statistical bias and conventional bias.

3.2 Measuring Performance

> **Note**
>
> This performance section is important in choosing the correct models to ensure data integrity, however for the actual statistical tests, we'll focused on more common statistics like mean and proportion.

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, however, to begin to explain the differences between these models we should 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 on a confusion matrix. In the image below, green are correct predictions while red are incorrect predictions.

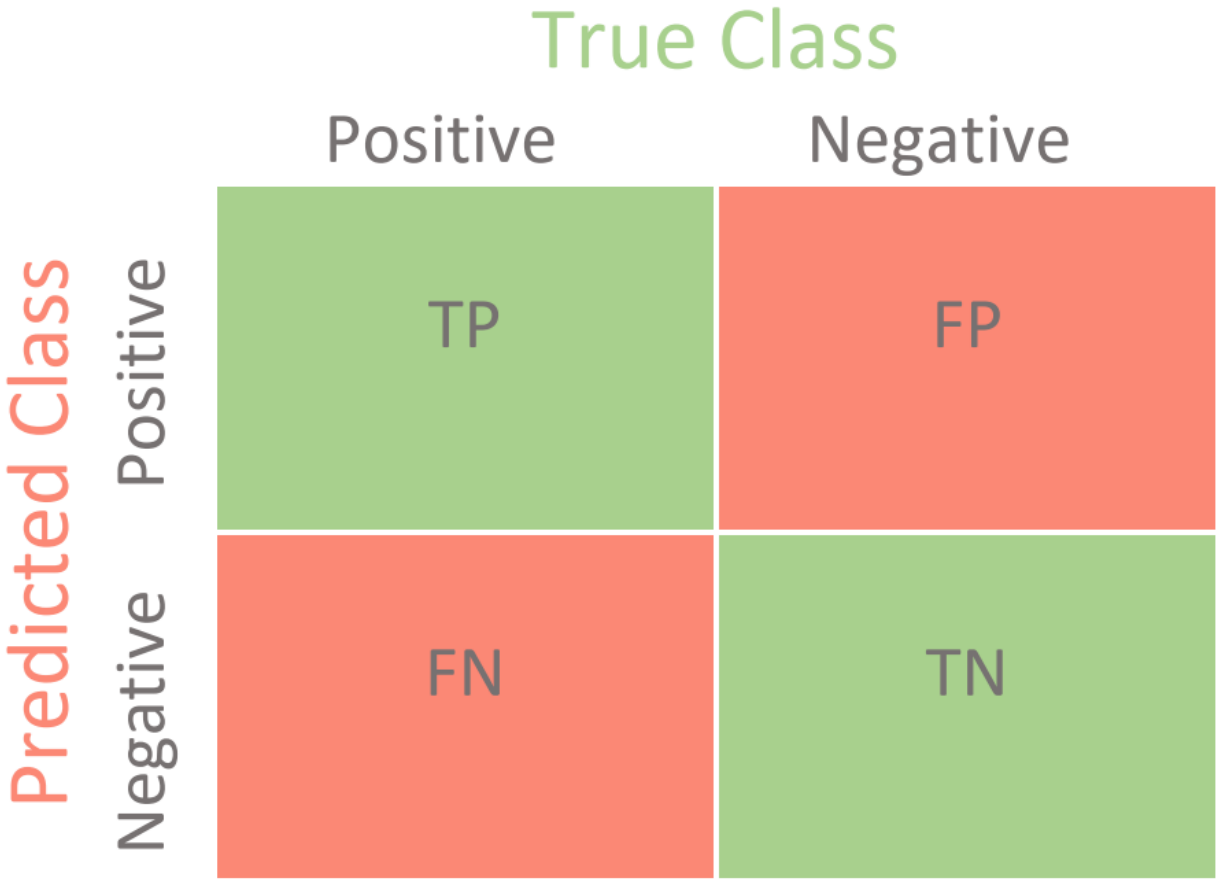


Figure 3.1: confusion_matrix

3.2.1 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.

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

3.2.2 Precision

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

$Precision = \frac{TP}{TP+FP}$

3.2.3 Recall

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

$Recall = \frac{TP}{TP+FN}$

3.2.4 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 $= \frac{2*Precision*Recall}{Precision+Recall}$

3.3 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 treating the source data as the basis for the original assumptions (our *null hypotheses*), and then using the output from Fairface and Deepface to test for statistically significant differences. 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 model.

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

3.3.1 Demographics

- Age Group
- Gender
- Race

3.3.2 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

3.3.3 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:

$H_0 : p = p_{\text{Source Data Subset}}$

$H_A : p \neq p_{\text{Source Data Subset}}$

With the following test statistic:

$\frac{\sqrt{n}(\hat{p}-p)}{\sqrt{p(1-p)}}$

With the p-value being calculated by:

$P(|Z| > \hat{p}|H_0)$

$= P(|Z| > \frac{\sqrt{n}(\hat{p}-p)}{\sqrt{p(1-p)}})$,

where

- n: output data subset size
- $\hat{p}$: output data subset proportion
- p: source data subset proportion

3.3.4 Notation

Before we list the specific tests, we should introduce some notation.

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 $D \in \{Source, Fairface, Deepface\} = \{D_0, D_f, D_d\}$

3.3.5 More Specific Proportion Tests

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:

$H_0 : p_F = p_{F|D_0}$

$H_A : p_F \neq p_{F|D_0}$

P-value Calculation:

$P(|Z| > \frac{\sqrt{n}(\hat{p}-p)}{\sqrt{p(1-p)}})$,

where

- $p = p_{F|D_0}$: proportion of females from the source data
- $\hat{p} = p_{F|D_f}$: proportion of females from the fairface output
- $n = n_{F \cup M|D_f}$: number of data points in the gender subset form 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*, then we could write a hypothesis test like:

$H_0 : p_{F|B} = p_{F|D_0 \cap B}$

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

P-value Calculation:

$P(|Z| > \frac{\sqrt{n}(\hat{p}-p)}{\sqrt{p(1-p)}})$,

where

- $p = p_{F|D_0 \cap B}$: proportion of females from the source data, given they were black
- $\hat{p} = p_{F|D_f \cap B}$: proportion of females from the fairface output, given they were black
- $n = n_{F \cup M|D_f \cap B}$: number of data points in the gender subset form the fairface output, given they were black.

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

> **i** From the report requirements
>
> Also can be called "Analyses"
> This section might contain several subsections as needed.

- At least one subsection should describe the exploratory data analysis you did.
- What modifications were necessary to make the dataset ready for analysis? (e.g. dealing with missing values, removing certain rows, replacing/cleaning text values, binning, etc)
- Describe the analyses you did to answer the question of interest. **Explain why you believe these methods are appropriate.**

- At least one subsection should describe the exploratory data analysis you did.
- What modifications were necessary to make the dataset ready for analysis? (e.g. dealing with missing values, removing certain rows, replacing/cleaning text values, binning, etc)
- Describe the analyses you did to answer the question of interest. **Explain why you believe these methods are appropriate.**
- At least one subsection should describe the exploratory data analysis you did.
- What modifications were necessary to make the dataset ready for analysis? (e.g. dealing with missing values, removing certain rows, replacing/cleaning text values, binning, etc)
- Describe the analyses you did to answer the question of interest. **Explain why you believe these methods are appropriate.**

Some methods we learn in this class include distribution comparison, correlation analysis, and hypothesis testing. You are required to include hypothesis tests into the project, but feel free to use additional methods to tell a good story about the data.

3.4 Standardizing output

The model outputs for both FairFace and DeepFace do not conform to the categories provided within the University of Tennessee - Knoxville (UTK) dataset. We elected to take the outputs from each model and modify them based upon the categories specified in the UTK dataset, namely:

- "[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"

3.4.1 From FairFace

- **Race**: The FairFace classification model had two options - one for "fair7" and one for "fair4." The latter provided predictions of race in the following categories: [White, Black, Asian, Indian]. Of key note, the 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"
- **Age**: 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. This split will allow

us to determine whether or not the prediction correctly binned the age (i.e. $lowerBound \leq actualAge \leq upperBound$), and if not - how far outside of those bounds the actual age lay.

- **Gender**: no change to outputs of "Male" and "Female."

3.4.2 From DeepFace

- **Race**: Racial categorical output from DeepFace includes the following categories ["middle eastern", "asian", "white", "latino hispanic", "black", "indian"]
- **Age**: DeepFace provides a prediction of a single, specific, predicted age. We elected to match the predicted age to be the same range as would be predicted by Fair Face. For example, if DeepFace predicts an age like "19," we assign it the same matching category as it would have in FairFace - "10-19." From there, we also split this category into an upper and lower bound. In spite of the fact that DeepFace does not provide any bounds or ranges on its age prediction outputs, to have a similar and fair comparison of both models, we give it those same upper and lower bounds for equitable comparison.
- **Gender**: DeepFace outputs are "Man" and "Woman", and we refactor those values to "Male" and "Female" respectively.

3.5 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 an image in each of these categories.

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, and using OpenCV as the default detection backend). Running these initial scripting tests, we encountered a failure rate in DeepFace of approximately 70% in identifying and categorizing an image of a face.

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.

3.5.1 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, 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 permutation evaluation.

3.5.2 FairFace Analysis Options

The default script from FairFace provided no options via its command line script to change 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 gives 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.

3.5.3 Specific Permutations

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

Detection	Detection Model	Image Source	Results Output
Enabled	FairFace=Dlib; DeepFace=OpenCV	Pre-cropped	new_ff_c_p.csv, crop_df_p_opencv.csv
Enabled	FairFace=Dlib; DeepFace=OpenCV	In-The-Wild	new_ff_uc_p.csv, uncropped_df_p_opencv.csv
Enabled	FairFace=Dlib; DeepFace=mtcnn	Pre-cropped	new_ff_c_p.csv, crop_df_p_mtcnn.csv
Enabled	FairFace=Dlib; DeepFace=mtcnn	In-The-Wild	new_ff_uc_p.csv, uncropped_df_p_mtcnn.csv
Disabled	FairFace,DeepFace=None	Pre-cropped	new_ff_c_np.csv, cropped_df_np.csv
Disabled	FairFace,DeepFace=None	In-The-Wild	new_ff_uc_np.csv, uncropped_df_np.csv

We processed each of the above setting permutations againnst 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.

3.5.4 Permutation Sample Results (LN & DV)

(enforcement of facial detection, detection backend model, and cropped images vs. faces in-the-wild)

3.5.5 Setting Selection

Upon completion of our evaluation, we determined the settings that gave both models the best chance of success included enabling facial detection with mtcnn for DeepFace and Dlib for FairFace on uncropped images.

From there, we proceeded to process the entirety of the UTK dataset using these settings. The only exception are 4 images that did not conform to UTK's naming convention to identify age, gender, and race of the subject in the image.

We wrote a script, MasterScript.py, to enable us to perform batch iteration of images and generate output files. When processing, we generated both the non-normalized output content and normalized output content.

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 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.

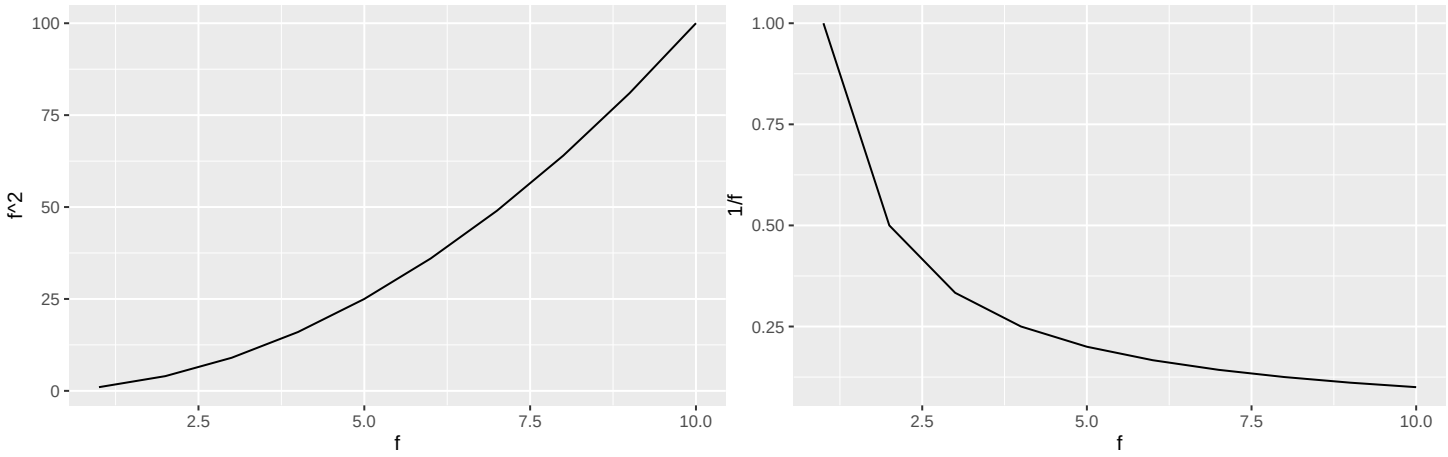
The following table outlines the output files.

The last file, MasterDataFrame.csv, is the final output of our evaluation. This file is in the following format, with the following column definitions:

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 age ranges from the FairFace outputs) for each image in the UTK datas
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
pred_gender	The gender of the subject in the image predicted by the given prediction model
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
pred_age_lower	The integer lower bound of the predicted age group
pred_age_upper	The integer upper bound of the predicted age group

4 Results

4.1 Tabbed example output



(a) Subcap

(b) Subcap 2

Figure 4.1: ANother example caption

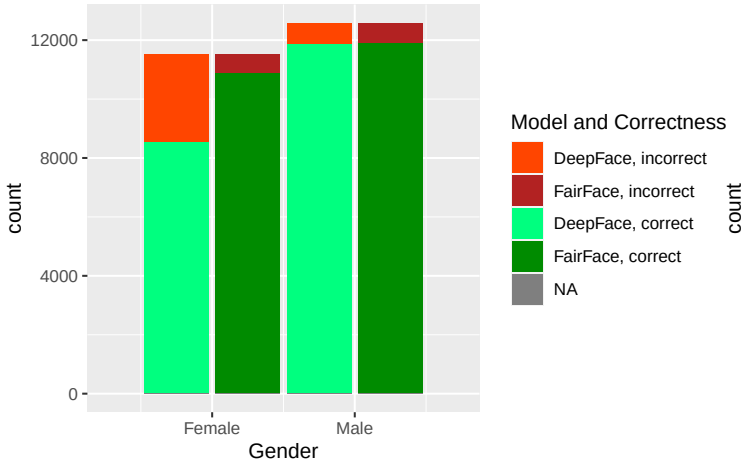
4.2 Example outout

```
[1]  1  2  3  4  5  6  7  8  9 10
```

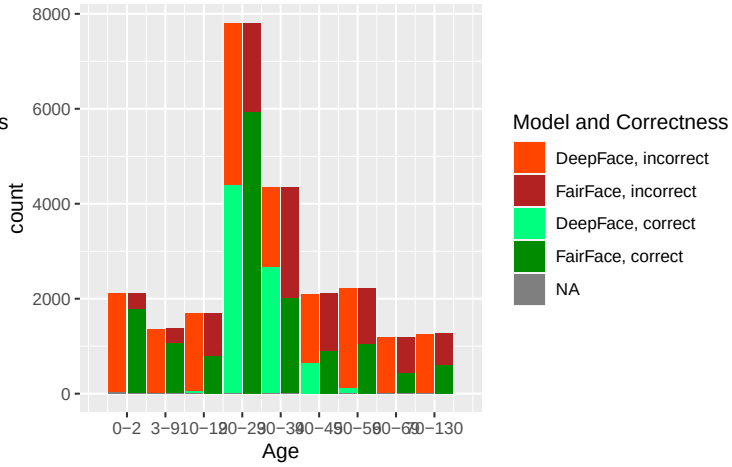
> **From the report requirements**
>
> Describe the results of your analysis using visualizations, descriptive statistics, tables and similar.
> Don't focus too much on the implications in this section – that's what the next section is for. Just present the numbers/graphs.

4.3 Model Output

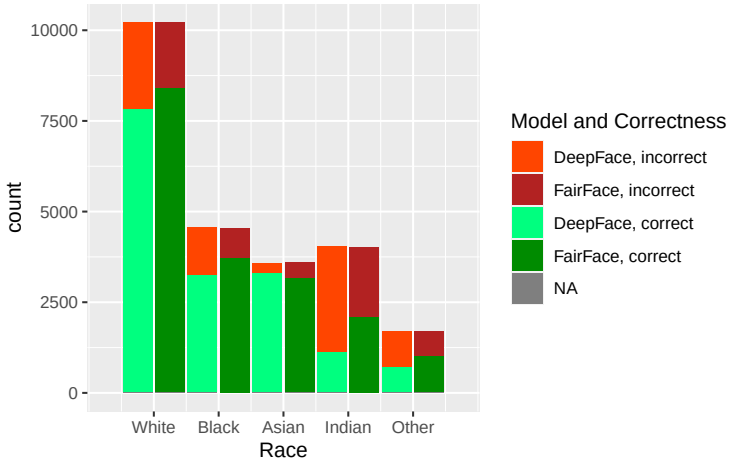
The two models, DeepFace and FairFace, were run on the dataset described previously. In Figure 4.2, one can see the results of the predictions done by each model, by each factor that was considered: age, gender, and race. Note that the histogram distributions match the correct (source dataset) distributions, 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 4.2: Histograms of the output from DeepFace and FairFace, with correct vs incorrect values colored. Note that the distributions match the correct (source dataset) distributions.

4.4 Model Performance

4.4.1 TODO: Remove

4.5 Hypothesis Testing

4.5.1 TODO: Remove

Test cateogry	Null Proportion	FairFace Proportion	FairFace P-Value	DeepFace Proportion	DeepFace P-Value
0-2	0.0880010	0.0762941	0.0000000	0.0000000	0
3-9	0.0568832	0.0664564	0.0000000	0.0000000	0
10-19	0.0697867	0.0606451	0.0000000	0.0206211	0
20-29	0.3238735	0.3480138	0.0000000	0.3987029	0
30-39	0.1802755	0.1762899	0.1075659	0.4047312	0
40-49	0.0872542	0.1023619	0.0000000	0.1467177	0
50-59	0.0923160	0.0930638	0.6884418	0.0268158	0
60-69	0.0490831	0.0490640	0.9890528	0.0024113	0
70-130	0.0525268	0.0278112	0.0000000	0.0000000	0

Test cateogry	Null Proportion	FairFace Proportion	FairFace P-Value	DeepFace Proportion	DeepFace P-Value
White	0.4240727	0.3854136	0.0000000	0.3757120	0
Black	0.1891129	0.1652484	0.0000000	0.1479233	0
Asian	0.1487843	0.1448259	0.0842652	0.2408847	0
Indian	0.1670816	0.1030675	0.0000000	0.0611150	0
Other	0.0709485	0.2014445	0.0000000	0.1743649	0

Test cateogry	Null Proportion	FairFace Proportion	FairFace P-Value	DeepFace Proportion	DeepFace P-Value
Female	0.4780101	0.4797642	0.5857219	0.3833202	0
Male	0.5219899	0.5202358	0.5857219	0.6166798	0

Test Category	Test Condition	Null Proportion	FairFace Proportion	FairFace P-Value	DeepFace Proportion	DeepFace P-Value
Female	White	0.4572938	0.4888530	0.0000000	0.4355428	3.32e-05
Female	Asian	0.5415505	0.5431356	0.8509504	0.3879876	0.00e+00
Female	Black	0.4872751	0.4649586	0.0048467	0.2405846	0.00e+00
Female	Indian	0.4325801	0.4430125	0.2940541	0.3496599	0.00e+00
Female	Other	0.5508772	0.4477643	0.0000000	0.3972341	0.00e+00
Male	White	0.5427062	0.5111470	0.0000000	0.5644572	3.32e-05
Male	Asian	0.4584495	0.4568644	0.8509504	0.6120124	0.00e+00
Male	Black	0.5127249	0.5350414	0.0048467	0.7594154	0.00e+00
Male	Indian	0.5674199	0.5569875	0.2940541	0.6503401	0.00e+00
Male	Other	0.4491228	0.5522357	0.0000000	0.6027659	0.00e+00

4.5.2 Updated Table Version with Data from Carl, Bhav

4.5.2.1 TODO: Remove

Race	Category	F1_d	F1_f	Accuracy_d	Accuracy_f	prop_d	prop_f	p_value_d	p_valu
All	0-2	NA	0.8959757	0.5000000	0.9172888	0.0000000	0.0762941	0.0000000	0.0000
All	3-9	NA	0.7176035	0.5000000	0.8772778	0.0000000	0.0664564	0.0000000	0.0000
All	10-19	0.0478601	0.5052498	0.5055825	0.7211461	0.0206211	0.0606451	0.0000000	0.0000
All	20-29	0.5054326	0.7332922	0.6217793	0.8050592	0.3987029	0.3480138	0.0000000	0.0000
All	30-39	0.3786318	0.4670003	0.6275447	0.6741504	0.4047312	0.1762899	0.0000000	0.1075
All	40-49	0.2276278	0.3943970	0.5866155	0.6786302	0.1467177	0.1023619	0.0000000	0.0000
All	50-59	0.0801673	0.4633983	0.5137145	0.7049843	0.0268158	0.0930638	0.0000000	0.6884
All	60-69	0.0016129	0.3739425	0.4991769	0.6708204	0.0024113	0.0490640	0.0000000	0.9890
All	70-130	NA	0.6270661	0.5000000	0.7383514	0.0000000	0.0278112	0.0000000	0.0000
All	White	0.8095461	0.8610399	0.8365916	0.8788455	0.3757120	0.3854136	0.0000000	0.0000
All	Black	0.7964994	0.8684858	0.8462797	0.8997692	0.1479233	0.1652484	0.0000000	0.0000
All	Asian	0.7038975	0.8948932	0.9005150	0.9338128	0.2408847	0.1448259	0.0000000	0.0842
All	Indian	0.4092481	0.6402458	0.6310597	0.7488102	0.0611150	0.1030675	0.0000000	0.0000
All	Other	0.2389021	0.3087473	0.6283106	0.7105889	0.1743649	0.2014445	0.0000000	0.0000
All	Female	0.8197702	0.9429153	0.8402892	0.9453080	0.3833202	0.4797642	0.0000000	0.5857
All	Male	NA	NA	NA	NA	0.6166798	0.5202358	0.0000000	0.5857
White	0-2	NA	0.9039010	0.5000000	0.9334307	0.0000000	0.0737749	0.0000000	0.0000
White	3-9	NA	0.7503392	0.5000000	0.8668432	0.0000000	0.0770059	0.0000000	0.0526
White	10-19	0.0634648	0.5638298	0.5102546	0.7330315	0.0163771	0.0693592	0.0000000	0.0000
White	20-29	0.4256326	0.6697460	0.6363584	0.8072440	0.3311940	0.2455574	0.0000000	0.0000
White	30-39	0.3884765	0.4731553	0.6486930	0.6821608	0.4022353	0.1628433	0.0000000	0.0410
White	40-49	0.2224248	0.3847156	0.5730236	0.6683039	0.2100255	0.1186861	0.0000000	0.0000
White	50-59	0.0890599	0.4832502	0.5086819	0.7046059	0.0376231	0.1419494	0.0000000	0.0252
White	60-69	NaN	0.3545817	0.4978778	0.6482054	0.0025451	0.0680668	0.0000000	0.0064
White	70-130	NA	0.6342183	0.5000000	0.7403000	0.0000000	0.0427571	0.0000000	0.0000
White	Male	0.8892356	0.9595281	0.8697687	0.9566327	0.5644572	0.5111470	0.0000332	0.0000
White	Female	0.8556585	0.9526238	0.8697687	0.9566327	0.4355428	0.4888530	0.0000332	0.0000
Asian	0-2	NA	0.9164589	0.5000000	0.9301909	0.0000000	0.2029235	0.0000000	0.0000
Asian	3-9	NA	0.7140255	0.5000000	0.9111790	0.0000000	0.0928633	0.0000000	0.0000
Asian	10-19	0.0395257	0.3798450	0.5023622	0.6944844	0.0457370	0.0366867	0.0000000	0.2105
Asian	20-29	0.5572885	0.8557951	0.5947638	0.8792496	0.5044874	0.4379478	0.0000000	0.0001
Asian	30-39	0.2992611	0.5069357	0.6258649	0.7042053	0.3206766	0.0988822	0.0000000	0.0000
Asian	40-49	0.1520190	0.3320463	0.5924407	0.6626378	0.0995858	0.0369733	0.0000000	0.3360
Asian	50-59	0.0898876	0.4608696	0.5273304	0.7272357	0.0250259	0.0315277	0.0066069	0.9201
Asian	60-69	NaN	0.4141414	0.4988555	0.7581786	0.0044874	0.0315277	0.0000000	0.0000
Asian	70-130	NA	0.7441860	0.5000000	0.8051278	0.0000000	0.0306678	0.0000000	0.0000
Asian	Male	0.7940330	0.8914286	0.7911354	0.8993780	0.6120124	0.4568644	0.0000000	0.8509
Asian	Female	0.7630232	0.9058670	0.7911354	0.8993780	0.3879876	0.5431356	0.0000000	0.8509
Black	0-2	NA	0.8854962	0.5000000	0.9026663	0.0000000	0.0180859	0.0000000	0.2466
Black	3-9	NA	0.7400881	0.5000000	0.8957332	0.0000000	0.0298920	0.0000000	0.0039
Black	10-19	0.0164609	0.3784787	0.5032694	0.6953003	0.0000000	0.0635519	0.0000000	0.0003
Black	20-29	0.5880567	0.6875152	0.6367203	0.7184594	0.4094997	0.4687265	0.0272015	0.0000
Black	30-39	0.4413203	0.4518681	0.5975793	0.6299581	0.4724564	0.2398895	0.0000000	0.0001
Black	40-49	0.1827542	0.3260274	0.5549318	0.6301481	0.1017426	0.0864104	0.0000300	0.3688
Black	50-59	0.0549451	0.3565062	0.5097384	0.6523810	0.0160202	0.0557649	0.0000000	0.0421
Black	60-69	0.0104712	0.3508772	0.5019317	0.6526201	0.0002811	0.0301432	0.0000000	0.0012
Black	70-130	NA	0.4086022	0.5000000	0.6383490	0.0000000	0.0075358	0.0000000	0.0000
Black	Male	0.8471279	0.9637681	0.8126869	0.9625782	0.7594154	0.5350414	0.0000000	0.0048
Black	Female	0.7724665	0.9615732	0.8126869	0.9625782	0.2405846	0.4649586	0.0000000	0.0048
Indian	0-2	NA	0.8132911	0.5000000	0.8443303	0.0000000	0.0318164	0.0000000	0.0000

4.6 Model Performance, Hypothesis Testing

For each category and model, we calculate the F1 score, accuracy, and p-value, as described in section 3. The results are summarized in **?@tbl-perf-pvalue**. 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 race groups;

4.6.1 TODO

- Add color key
- Color p-values based on sig level = 99.7%
- Better description of signifiance of numerical values of Accuracy, F1 score
- Add line plot of F1, Accuracy, p-value OR correlation matrix
- Make table caption work?

Static table:

- Values where we FAIL to reject null hypothesis

4.6.2 Statistical Power

$$\beta = P\left(\left|\frac{\sqrt{n}\cdot\hat{p}-p_a}{\sqrt{p_a(1-p_a)}}\right| \geq \frac{\sqrt{n}\cdot p_0 - p_a}{\sqrt{p_0(1-p_0)}}\right)$$

Our selected level of significance is 99.7% (3-sigma). Type-II error is denoted by β above, and Power will be $1-\beta$

With p_0 being our *assumed* population proportion (from the source dataset and what we used in our tests), p_a being the *actual* population proportion (from one or more of the below methods), n being the number of predicted members of a racial group (i.e. "Indian"),

- For Gender - assume that sex at birth is a bernoulli trial, over time, the proportion for both genders should be 0.5
- For age groups - assume that age has a true normal distribution. Each race may have different means and standard deviations for their distribution of age, but still adhere to a normal distribution. The "population" proportions may be a bit more challenging to calculate, but under this framework, we may be able to get there.
  - May be able to get via bootstrapping the source dataset, average age by race - I think that's what we did in our last project?
  - Could look at external data? May not have time to look through everything.

$$\frac{\sqrt{n_M}\cdot(\bar{p}_M - p_S)}{\sqrt{p_S\cdot(1-P_S)}}$$

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

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.

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