

# **Bias in Facial Classification ML Models**

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

# Table of contents

<b>Abstract</b>	<b>3</b>
How we should write this report . . . . .	3
<b>1 Introduction</b>	<b>5</b>
<b>2 Data Exploration</b>	<b>6</b>
<b>3 Methods</b>	<b>7</b>
3.1 The Big Picture . . . . .	7
3.2 Measuring Performance . . . . .	7
3.2.1 Accuracy . . . . .	7
3.2.2 Precision . . . . .	9
3.2.3 Recall . . . . .	9
3.2.4 F1-Score . . . . .	9
3.3 Conditional Peformance . . . . .	9
3.3.1 Performance - Overall . . . . .	9
3.3.2 Performance - Main Demographics . . . . .	9
3.3.3 Performance - Demographics' Subgroups . . . . .	10
3.4 Hypothesis Testing . . . . .	10
3.4.1 Proportion Test . . . . .	10
3.4.2 Chi-Squared Test . . . . .	11
3.4.3 Non Parametric Cohort . . . . .	11
<b>4 Describe the analyses you did to answer the question of interest. Explain why you believe these methods are appropriate.</b>	<b>13</b>
4.1 Standardizing output . . . . .	13
4.1.1 From FairFace . . . . .	13
4.1.2 From DeepFace . . . . .	13
4.2 Evaluating Permutations of Inputs and Models for Equitable Evaluation . . . . .	14
4.2.1 DeepFace Analysis Options . . . . .	14
4.2.2 FairFace Analysis Options . . . . .	14
4.2.3 Specific Permutations . . . . .	15
<b>5 Results</b>	<b>16</b>
5.1 Tabbed example output . . . . .	16
5.2 Example outout . . . . .	16
<b>6 Conclusions</b>	<b>17</b>
<b>References</b>	<b>18</b>

# Abstract

Here is where we will put the abstract.

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):

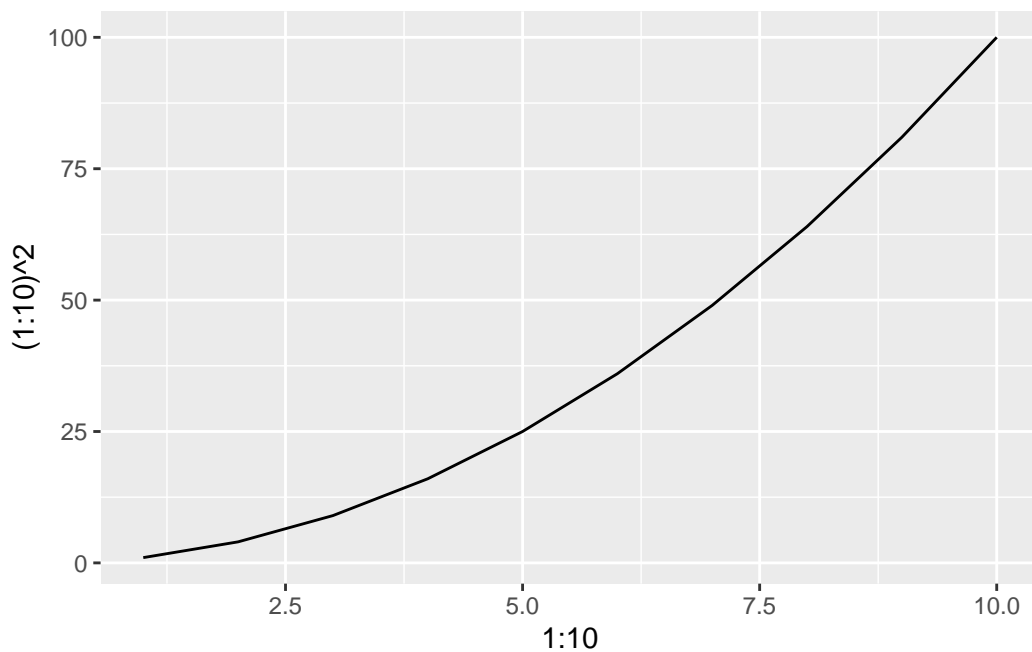


Figure 1: A caption for generated figure

- 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`.

**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

**i** 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 is 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 Exploration

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!!**

**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.

## 3 Methods

### 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

#### **i** 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

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.

#### 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.

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

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

Figure 3.1: confusion\_matrix



### 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 Conditional Performance

Just as with the exploratory analysis on the data set prior to running either model, it's helpful to garner an understanding of what the results were using basic statistics and metrics. By calculating metrics in a cascading fashion, starting with the overall performance and then zeroing in on subgroups, we can get an idea if and where bias is worth investigating.

### 3.3.1 Performance - Overall

- Confusion Matrix / Performance Measures for Fairface (Fairface Results)
- Confusion Matrix / Performance Measures for Deepface (Deepface Results)

### 3.3.2 Performance - Main Demographics

- Age Group
  - Fairface Results
  - Deepface Results
- Gender
  - Fairface Results
  - Deepface Results
- Race
  - Fairface Results
  - Deepface Results

### 3.3.3 Performance - Demographics' Subgroups

- Age Group (9 groups)
  - Fairface Results
  - Deepface Results
- Gender (2 groups)
  - Fairface Results
  - Deepface Results
- Race (5 groups)
  - Fairface Results
  - Deepface Results

## 3.4 Hypothesis Testing

Now that we have an idea how each of these models performed in general and across localized subsets, let's see if our models truly have statistically significant results and if bias is present.

Our tests will focus on proportions, and will discover if there is statistically significant differences between and within the models.

### 3.4.1 Proportion Test

The **Proportion Test** will be useful when measuring difference in overall performance and single categories between models.

The *hypothesis test* we'll be using is:

- $H_0 : \pi_2 - \pi_1 = 0$
- $H_A : \pi_2 - \pi_1 \neq 0$

A look at the test statistic:

$$Z = \frac{p_2 - p_1 - \pi_0}{\sqrt{p^*(1-p^*)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \text{ where}$$

the *pooled proportion*,  $p^*$  is calculated as

$$p^* = \frac{x_1 + x_2}{n_1 + n_2}$$

A look at the code:

- x: vector of successes and failures
- n: vector of counts of trials (can be ignored if x is a matrix or a table)
- alternative: we can change if we have a specific equality to test
- conf.level: 0.95 is the default (not pertinent to specify)

We can also build **confidence intervals**. We have 95% confidence that the true difference between the props lies within the following:

### 3.4.2 Chi-Squared Test

The **Chi-squared Test** will be useful when testing across k-levels of a categorical variable. For instance, this could be useful when testing proportion within a demographic for a single model (i.e. accuracy across race in Fairface).

The *hypothesis test* we'll be using is:

*Equivalent Proportions (test for uniformity):*

- $H_0 : \pi_1 = \pi_2 = \dots = \pi_n = \pi$
- $H_A : H_0$  is incorrect (at least one of the proportions is not equivalent)

A look at the test statistic:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}, \text{ where}$$

$O_i$  is the observed count, and

$E_i$  is the expected count.

A look at the code:

Imagine using this test across the categories of race. A test which favored  $H_A$  could give credence to bias.

The standard is using the **Chi-squared Test** in testing for uniformity, however we can test specific proportions for each category. Here's a preview in R:

A possible use case of the non-uniform **Chi-Squared Test** could be to test if the proportion of inaccuracies follows the proportion of each demographic subcategory. For instance, does inaccuracy across race follow the proportions of race in the sample?

### 3.4.3 Non Parametric Cohort

**i** From the report requirements

Also can be called "Analyses"

This section might contain several subsections as needed.

«««< HEAD - 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.** ||||||| feac3d2 - 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)
-



## 4 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)

### 4.1 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”

#### 4.1.1 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 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.”

#### 4.1.2 From DeepFace

**Race:** Racial categorical output from DeepFace includes the following categories []

**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.

## 4.2 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). 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.

### 4.2.1 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`, `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 permutation evaluation.

### 4.2.2 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.

### 4.2.3 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	Enabled	Fairface=Dlib;	Deep-Face=[opencv,mtcnn,fastmtcnn,ssd]
Pre-cropped, in-the-wild	Disabled	None	Pre-cropped, in-the-wild				

For 8 total comparisons across a subset of X sampled images from UTK. ### Permutation sample results (LN & DV)

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

- Describe the analyses you did to answer the question of interest. **Explain why you believe these methods are appropriate.** »»»> d1ea1b4bc252c28fc569327e08036166d004f57a

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.

## 5 Results

This is where all the plots will go!!!! Here are some examples of plot layout:

### 5.1 Tabbed example output

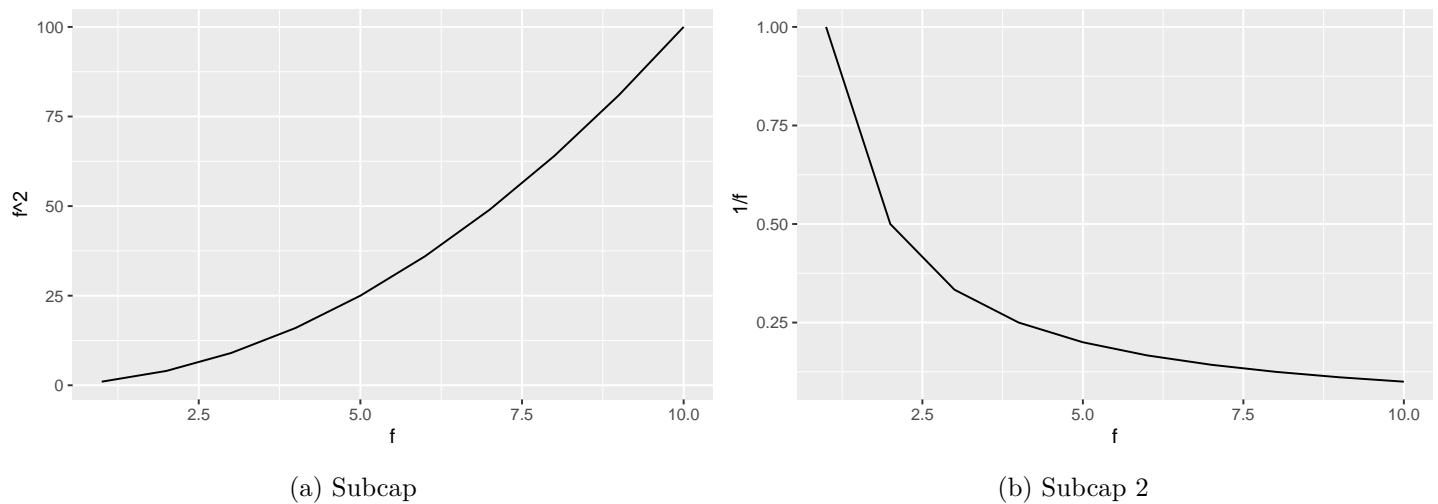


Figure 5.1: ANother example caption

### 5.2 Example output

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

**i** 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.



## 6 Conclusions

**i** 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 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.