Bias in Facial Classification ML Models

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

Table of contents

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):

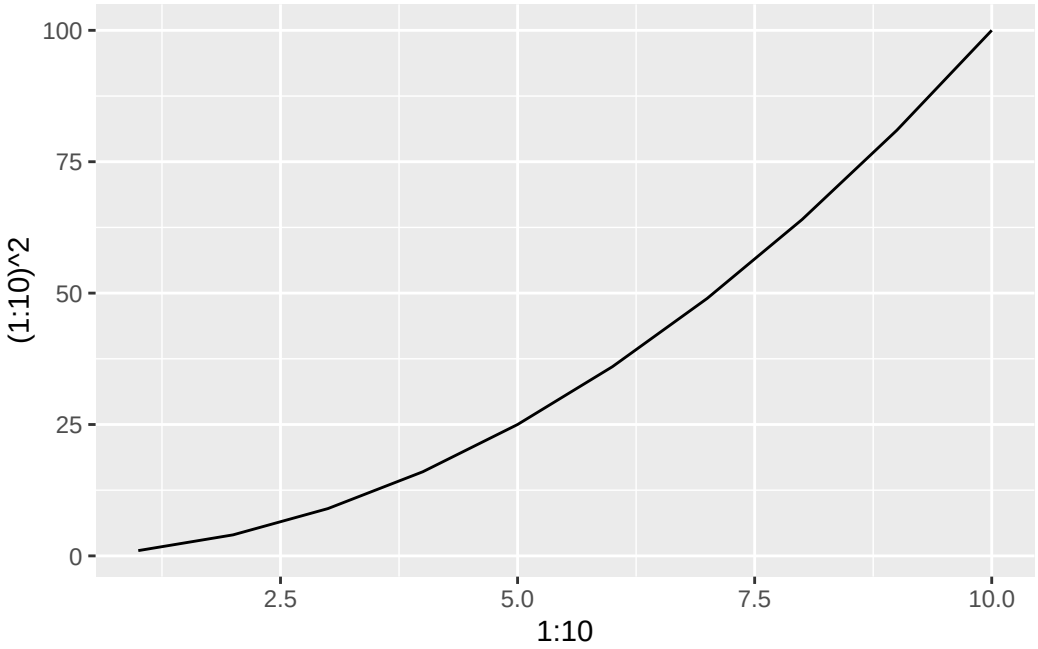


Figure 1: A caption for generated figure

- 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

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 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!!

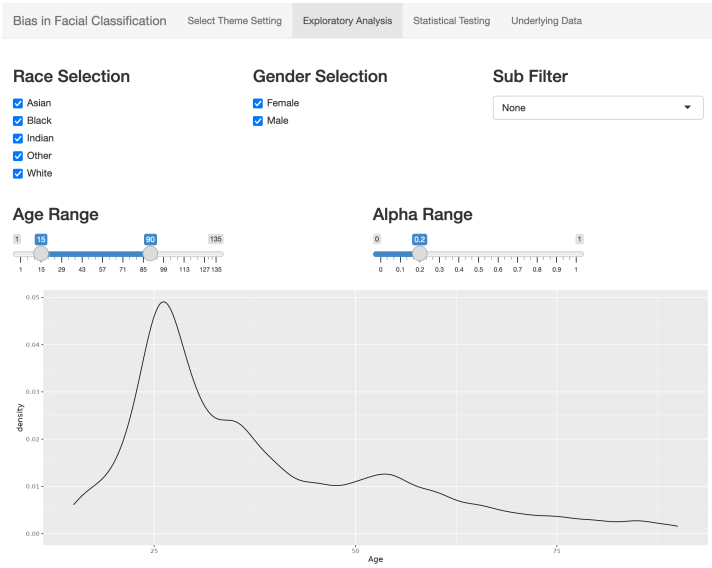


Figure 2.1: Image data EDA

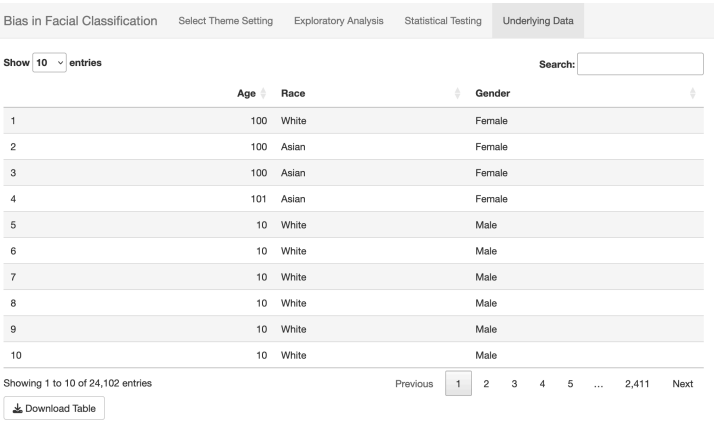


Figure 2.2: Image dataset visualization

Figure 2.3: 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

> **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

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.

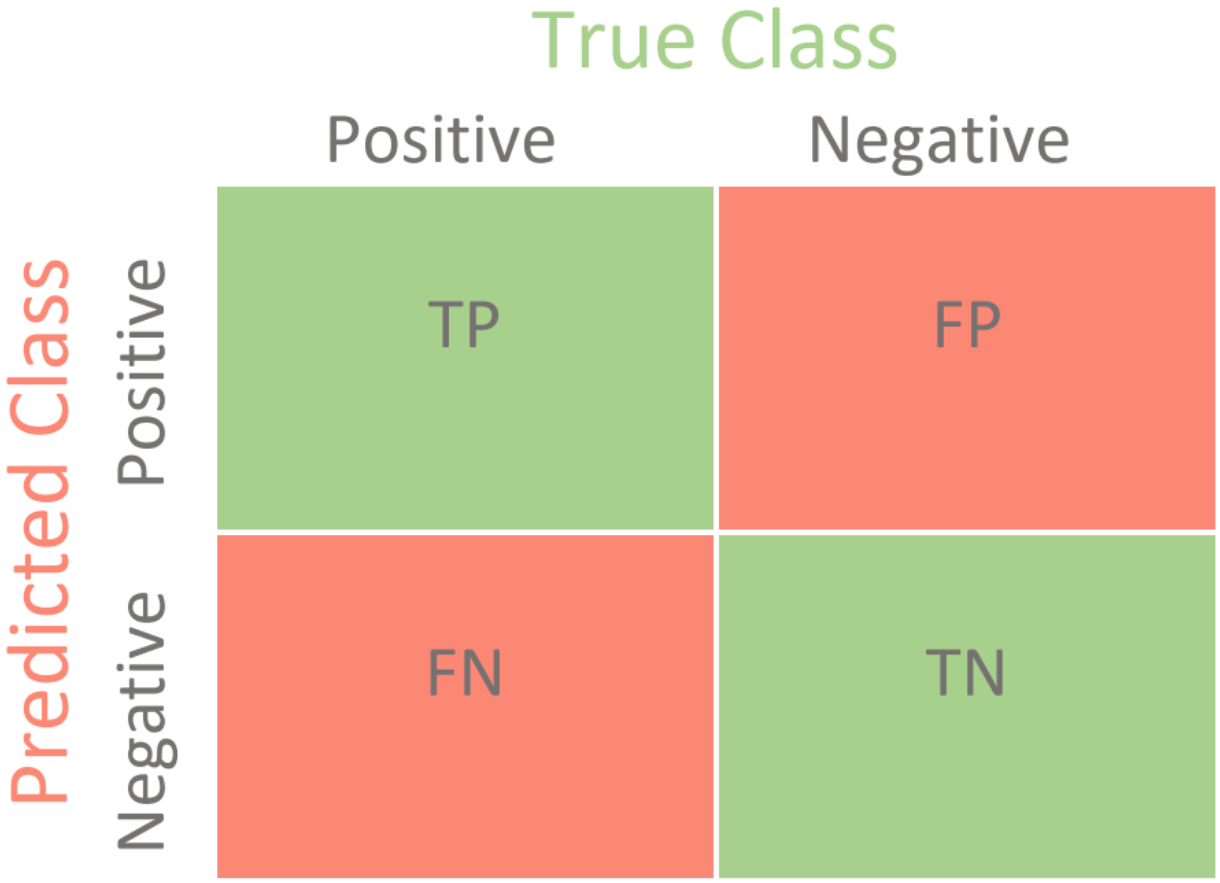


Figure 3.1: confusion_matrix

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 Statistical Testing

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 statistics in a cascading fashion, starting with the overall and then zeroing in on subgroups, we can get an idea if and where bias is worth investigating.

3.3.1 Overall

- Compare the mean age of the input to the mean age of the Deepface output
- Bin the input ages, compare bins between the input and the output of both models
- Bin the input ages, compare bins between the output of both models

Note: We can think of the binned data as categorical variables or as a small subset of discrete numerical data by ordering the bins.

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 Resultseepface 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 Difference in Means Test

The **Difference in Means Test** will be useful when measuring differences between the mean of ages. Whether that be the aggregate data means or means of demographic subsets.

Depending on the relationship between the data, there are different variants of this test.

3.4.1.1 Unpaired / Independent Samples: Unpooled Variances

This test is best for independent and separate groups.

The *hypothesis test* is:

- $H_0 : \mu_2 - \mu_1 = 0$
- $H_A : \mu_2 - \mu_1 \neq 0$

A look at the test statistic:

$T = \frac{\bar{x_2} - \bar{x_1} - \mu_o}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

and has degrees of freedom, v, where

$v = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{(\frac{s_1^2}{n_1})^2/(n_1 - 1) + (\frac{s_2^2}{n_2})^2/(n_2 - 1)}$

3.4.1.2 Unpaired / Independent Samples: Pooled Variances

We can improve the precision of a test if we can assume the equivalence of variances.

A look at the different test statistic:

- $T = \frac{\bar{x_2} - \bar{x_1} - \mu_0}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}}$,

with degrees of freedom $v = n_1 + n_2 - 2$,

and where $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ is known as the *pooled estimate of variance*.

3.4.1.3 Paired/Dependent Samples

In the previous tests, we were considering independent samples. In this case, we'll be considering the samples have a dependency between them. This is a common test for "before" and "after" type testing.

We can simplify the *hypothesis test* for this kind of sampling relationship to:

- $H_0 : \mu_2 - \mu_1 = \mu_d = 0$
- $H_A : \mu_d \neq 0$

With the test statistic being:

$T = \frac{\bar{d} - \mu_0}{\frac{s_d}{\sqrt{n}}}$,

where $\bar{d}$ is the mean of the pairwise differences and s_d is the sample standard deviation of the pairwise differences.

The degrees of freedom for this test are $df = n - 1$.

3.4.2 Proportion Test

The **Proportion Test** will be useful when measuring differences of 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:

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

3.4.3 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 singe model (i.e. accuracy across race in Fairface).

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

Equivalent Proportions (test for uniformity):

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

A look at the test statistic:

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

3.4.4 Non Parametric Cohort

> **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.

```
<!--PC!-->
## Standardizing output

The model outputs for both FairFace and DeepFace do not conform to the categories provided within the

"[race] is an integer from 0 to 4, denoting White, Black, Asian, Indian, and Others (like Hispanic, L

"[gender] is either 0 (male) or 1 (female)"

"[age] is an integer from 0 to 116, indicating the age"

### From FairFace

**Race**: The FairFace classification model had two options - one for "fair7" and one for "fair4."  T

**Age**: FairFace only provides a predicted age range as opposed to a specific, single, predicted age

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

### 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 th

Gender: DeepFace outputs are "Man" and "Woman", and we refactor those values to "Male" and "Femal

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

The need for this permutation evaluation rose from some initial scripting and testing of these models

We performed further exploratory analysis on both models in light of these facts, and sought some spe

DeepFace Analysis Options

DeepFace has a robust degree of avaialble settings when performing facial categorization and recognit

In a Python 3.8 environment, attempting to run detections using dlib, retinaface, mediapipe, yolov8,

FairFace Analysis Options

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

We converted the simple script to a class in Python without addressing any feature bugs or errors in

Specific Permutations

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

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
Enabled	FairFace=Dlib; DeepFace=fastmtcnn	Pre-cropped
Enabled	FairFace=Dlib; DeepFace=fastmtcnn	In-The-Wild
Enabled	FairFace=Dlib;DeepFace=ssd	Pre-cropped
Enabled	FairFace=Dlib;DeepFace=ssd	In-The-Wild
Disabled	FairFace,DeepFace=None	Pre-cropped
Disabled	FairFace,DeepFace=None	In-The-Wild

Detection	**Detection Model**	**Image source**
Enabled	Fairface=Dlib; DeepFace=[opencv,mtcnn,fastmtcnn,ssd]	Pre-cropped, in-the-wild
Disabled	None	Pre-cropped, in-the-wild

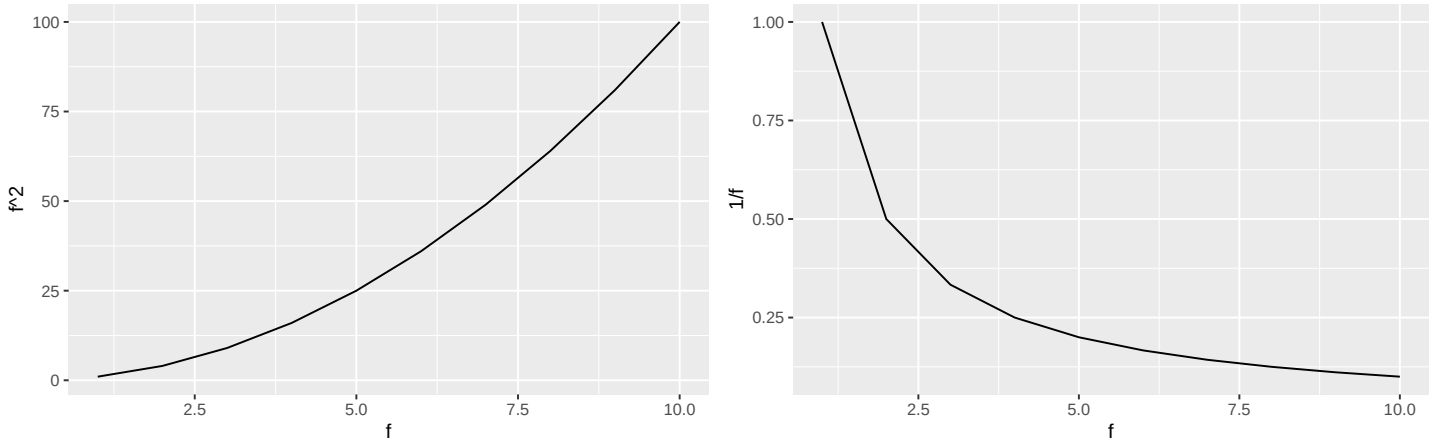
For 8 total comparisons across a subset of X sampled images from UTK.
<!--End PC!-->
Permutation sample results (LN & DV)

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

4 Results

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

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.

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.