

UNIVERSITY OF HERTFORDSHIRE

School of Physics, Engineering and Computer Science

**MSc Artificial Intelligence and Robotics with Advanced Research**

7COM1039-0109-2022

Advanced Computer Science Master's Project

28th August 2023

**Detecting and Mitigating Algorithmic Bias in Face Recognition Algorithms:**

**A Research Study**

**Name:** Rishabh Patel

**Student ID Number:** 20068205

**Supervisor:** Prof. Thiago Matos Pinto

# MSc Final Project Declaration

This report is submitted in partial fulfilment of the requirement for the degree of Master of Science in Artificial Intelligence and Robotics with Advanced Research at the University of Hertfordshire (UH).

It is my own work except where indicated in the report.

I did not use human participants in my MSc Project.

I hereby give permission for the report to be made available on the university website provided the source is acknowledged.

# Abstract

This project aims to evaluate the effectiveness of using AI Fairness 360 toolkit in reducing algorithmic bias in face recognition algorithms with the example of a gender classification model. Algorithmic bias is the presence of systematic errors or unfair outcomes in the output of an algorithm due to the data or design of the algorithm itself. Algorithmic bias can have negative impacts on the individuals and communities affected by it, such as violating their privacy, dignity, and human rights, as well as perpetuating and exacerbating existing social inequalities and discrimination. The project uses three publicly available datasets of facial images with labels indicating the gender (male or female) and trains a convolutional neural network (CNN) model for gender classification using TensorFlow and Keras. The project then uses a method from the AI Fairness 360 toolkit called reweighting, to detect and mitigate algorithmic bias in the CNN model. The project also evaluates the effectiveness of these methods in terms of accuracy and fairness using various fairness metrics, such as disparate impact, statistical parity difference, average odds difference, equal opportunity difference, and accuracy. The project contributes to the ethical development and use of facial recognition technology and provides insights and recommendations for improving the social benefits and reducing the social harms of facial recognition systems.

Table of Contents

[MSc Final Project Declaration 2](#_Toc144037269)

[Abstract 3](#_Toc144037270)

[1. Introduction 5](#_Toc144037271)

[1.1 Introduction and Background 5](#_Toc144037272)

[1.2 Research Question 7](#_Toc144037273)

[1.3 Project Objectives 7](#_Toc144037274)

[1.4 Report Structure 7](#_Toc144037275)

[2. Literature Review 8](#_Toc144037276)

[2.1 Introduction and Overview 8](#_Toc144037277)

[2.2 Facial Recognition Algorithms 9](#_Toc144037278)

[2.3 Algorithmic Bias 10](#_Toc144037279)

[2.4 Mitigation Strategies 11](#_Toc144037280)

[2.5 Summary 12](#_Toc144037281)

[3. Methodology 13](#_Toc144037282)

[3.1 Introduction and Overview 13](#_Toc144037283)

[3.2 Data Collection 14](#_Toc144037284)

[3.3 Data Analysis 14](#_Toc144037285)

[3.4 Evaluation 16](#_Toc144037286)

[3.5 Consideration of Ethical, Legal, Professional and Social Issues 17](#_Toc144037287)

[3.5.1 Ethical Issues 17](#_Toc144037288)

[3.5.2 Legal Issues 18](#_Toc144037289)

[3.5.3 Professional Issues 18](#_Toc144037290)

[3.5.4 Social Issues 18](#_Toc144037291)

[4. Results 19](#_Toc144037292)

[4.1 Dataset 1 19](#_Toc144037293)

[4.2 Dataset 2 22](#_Toc144037294)

[4.3 Dataset 3 23](#_Toc144037295)

[4.3.1 Version 1 24](#_Toc144037296)

[4.3.2 Version 2 26](#_Toc144037297)

[5. Discussion 28](#_Toc144037298)

[6. Evaluation and Conclusion 30](#_Toc144037299)

[7. References 31](#_Toc144037300)

[8. Appendix 34](#_Toc144037301)

[8.1 Appendix 1: Project Repository/Code 34](#_Toc144037302)

[8.2 Appendix 2: Gender classifier for dataset 1 34](#_Toc144037303)

[8.3 Appendix 3: Gender classifier for dataset 2 with normal model 34](#_Toc144037304)

[8.4 Appendix 4: Gender classifier for dataset 2 with complex model and vgg16 model 34](#_Toc144037305)

[8.5 Appendix 5: Gender classifier for dataset 3 Version 1 (male privileged set) 35](#_Toc144037306)

[8.6 Appendix 6: Gender classifier for dataset 3 Version 2 (female privileged set) 35](#_Toc144037307)

# 1. Introduction

## 1.1 Introduction and Background

Facial recognition technology has become increasingly prevalent in recent years, with applications ranging from unlocking smartphones to identifying suspects in criminal investigations. However, concerns have been raised about the potential for algorithmic bias in these systems, which can result in discriminatory outcomes and harm marginalized communities.

Algorithmic bias refers to the presence of systematic errors in the output of an algorithm due to the data or design of the algorithm itself. In the case of facial recognition, this can manifest as higher error rates for certain demographic groups, such as women and people of color. For example, a study conducted by Grother, Ngan, and Hanaoka (2023) at the National Institute of Standards and Technology (NIST) found that many facial recognition algorithms misidentified members of some groups up to 100 times more frequently than others.

The impacts of algorithmic bias in facial recognition can be severe. Inaccurate identifications can lead to false arrests and unnecessary encounters with law enforcement. This is particularly concerning given that facial recognition technology is increasingly being used by law enforcement agencies to identify suspects. Furthermore, the use of biased algorithms can preserve and aggravate existing societal inequalities.

As a student in the field of computer science, I am particularly interested in exploring the technical aspects of detecting and mitigating algorithmic bias in facial recognition algorithms. This project aims to contribute to the existing literature on this topic by conducting a research study on methods for detecting and mitigating algorithmic bias in face recognition algorithms.

My Bachelor’s final project was “Controlling Operating System and Applications using Hand Motion Gestures” which was based on machine learning. This project gave me the curiosity to learn more about this technology (machine learning) that is still evolving and being used widely for various purposes. This paved the path for my pursuit of Master’s Degree in Artificial Intelligence and Robotics with Advanced Research.

The topic of algorithmic bias in facial recognition is important because it addresses a pressing issue in the field of artificial intelligence and has significant implications for society at large. The topic is also relevant to my degree subject area, as it involves learning and applying machine learning concepts and techniques to analyze and improve a real-world problem.

The current state of the research on algorithmic bias in facial recognition is not very satisfactory, as there are many gaps and challenges that remain unaddressed. Some of these gaps include:

* The lack of standardized and representative datasets for training and testing facial recognition algorithms.
* The difficulty of measuring and quantifying algorithmic bias across different scenarios and populations.
* The trade-off between accuracy and fairness in designing and evaluating facial recognition algorithms.
* The ethical and legal implications of using biased facial recognition algorithms in various domains.

These gaps indicate that there is a need for more research on this topic, especially on developing methods for detecting and mitigating algorithmic bias in facial recognition algorithms. By addressing these gaps, my study hopes to advance the knowledge and practice of facial recognition technology and promote its fair and responsible use.

## 1.2 Research Question

The research questions can be described as:

RQ1: “How effectively is it possible to mitigate algorithmic bias in facial recognition algorithms using AI Fairness 360 Toolkit?”

RQ2: “What are the limitations and challenges of detecting and mitigating algorithmic bias in facial recognition algorithms?”

## 1.3 Project Objectives

1. To conduct a literature review on existing methods for detecting and mitigating algorithmic bias in facial recognition algorithms.
2. To design and implement a gender classification model using facial recognition.
   1. Find appropriate datasets for gender classification by facial recognition.
   2. Find/Select optimum machine learning algorithm to be used for training the face recognition model.
   3. Train the model on various datasets.
   4. Evaluate the model's accuracy and performance.
   5. Optimise the model further if necessary.
3. To use the AI Fairness 360 toolkit for detecting algorithmic bias in the trained models.
4. To evaluate the effectiveness of mitigation techniques for reducing algorithmic bias in facial recognition algorithms.
5. To analyze the results of the study and draw conclusions about the effectiveness of the methods for detecting and mitigating algorithmic bias in facial recognition algorithms.

These objectives will help me answer my research question by providing a comprehensive understanding of the current state of knowledge on this topic, as well as generating new insights through empirical research. By achieving these objectives, I hope to make a meaningful contribution to the field of artificial intelligence and help pave the way for more fair and equitable facial recognition systems.

## 1.4 Report Structure

The research report is structured as follows:

1. The first page contains the name of the university, the degree program, the module code and name, the date, the title of the project, the student’s name and ID, and the name of the supervisor.
2. The second page contains the project and report declaration, stating the authenticity of the report.
3. The third page contains the abstract, which is a concise summary of the main points of the project, including the aims, objectives, methods, results, conclusions, and recommendations.
4. The fourth page contains the table of contents, which lists the titles and page numbers of each chapter and appendix in the report.
5. The sixth page starts with chapter 1, which is the introduction. This chapter discusses the motivation and background of the project, states the research question and objectives, and outlines the report structure.
6. Chapter 2 is the literature review, which surveys and critically analyzes the existing literature on algorithmic bias in facial recognition algorithms. This chapter covers topics such as facial recognition algorithms, algorithmic bias, and mitigation strategies.
7. Chapter 3 is the methodology, which describes and justifies the methods and techniques used for data collection, data analysis, and evaluation. This chapter also discusses the ethical, legal, professional, and social issues related to the project.
8. Chapter 4 is the results, which presents and interprets the findings obtained from data analysis and evaluation. This chapter uses descriptive and inferential statistics, as well as visualizations, to communicate the results effectively.
9. Chapter 5 is the discussion, which evaluates and reflects on the achievements, methods, and results of the project. This chapter also identifies the limitations and challenges of the project and suggests directions for future work.
10. Chapter 6 is the conclusion, which summarizes the main points and findings of the project, highlights the contributions, and impacts of the project, and provides some personal reflections and lessons learned from the project.
11. The references section lists all the sources cited in the report using Harvard reference style.
12. The appendix section contains additional or supplementary material that supports or complements the main body/content of the report. For example, appendix 1 contains the source code for project.

# 2. Literature Review

## 2.1 Introduction and Overview

The literature review section of this report aims to provide a comprehensive and critical overview of the existing research on algorithmic bias in facial recognition algorithms. This review will cover this existing research by dividing it across three main sections important to this research study:

Facial recognition algorithms: This section will introduce the concept and working of facial recognition algorithms, as well as their common types, categories, and applications. This section will also highlight some of the benefits and challenges of using facial recognition technology in various domains.

Algorithmic bias: This section will define and explain the phenomenon of algorithmic bias, as well as its sources, causes, and effects. This section will also discuss how algorithmic bias can be measured and detected in facial recognition algorithms, using various methods and frameworks.

Mitigation strategies: This section will explore and evaluate some of the existing strategies and techniques for mitigating algorithmic bias in facial recognition algorithms. This section will also examine the advantages, disadvantages, challenges, and limitations of these mitigation strategies.

The purpose of this literature review is to provide a solid theoretical and empirical foundation for the research question and objectives. By reviewing and analyzing the relevant literature on this topic, I will be able to identify the gaps and challenges that need to be addressed, as well as the best practices and solutions that can be adopted or implemented for detection and mitigation of algorithmic bias in facial recognition algorithms.

One of the main gaps that I would like to address in my project is the lack of research on the use of the AI Fairness 360 toolkit for detecting and mitigating algorithmic bias in facial recognition algorithms. The AI Fairness 360 toolkit is an open-source library that provides a comprehensive set of tools and metrics for assessing and improving the fairness of machine learning models. The toolkit offers various methods for detecting and mitigating algorithmic bias at different stages of the machine learning pipeline, such as pre-processing, in-processing, and post-processing. The toolkit also supports multiple fairness dimensions, such as demographic parity, equalized odds, equal opportunity, and predictive parity (Bellamy et al., 2018).

By using the AI Fairness 360 toolkit for this project, I will be able to explore and compare the effectiveness of using the toolkit for detecting and mitigating algorithmic bias in facial recognition algorithms. This effectiveness will be evaluated in terms of accuracy and fairness when these methods are used as opposed to when they are not used. Furthermore, I will be able to contribute to the existing literature on this topic by providing new insights and findings on how the AI Fairness 360 toolkit can be used for reducing algorithmic bias in facial recognition systems.

## 2.2 Facial Recognition Algorithms

As explained by Alghamdi et al. (2020), facial recognition algorithms are a type of computer technology that uses a variety of algorithms to identify human faces in digital images, identify the person, and then verify the captured images by comparing them with facial images stored in a database. The main functions of these algorithms are finding faces in a picture, video, or live stream; creating a numerical representation (often called a data frame) of a face; and matching the representation to existing data sets or databases to recognize or confirm a person’s identity.

According to Saini, R., Saini, A. and Agarwal, D. (2014) there are two main types of facial recognition algorithms - geometric and photometric. The geometric type focuses on unique features, while the photometric type uses statistical methods to get values from an image. These values are then compared to templates to reduce variances. The algorithms can also be classified into two broader categories: feature-based and holistic models. The feature-based type focuses on facial landmarks and analyzes their spatial parameters and relation to other features, while holistic type views the human face as a whole unit. Some common types of facial recognition algorithms include Convolutional Neural Networks (CNN), Eigenfaces, Fisherfaces, Kernel Methods (PCA and SVM), Haar Cascades, Three-Dimensional Recognition, Skin Texture Analysis, Thermal Cameras, ANFIS, Local Binary Patterns Histograms (LBPH), FaceNet, NEC Megvii (Face++), Principal Component Analysis (PCA), Independent Component Analysis (ICA), Linear Discriminant Analysis (LDA), and Elastic Bunch Graph Matching (EBGM).

Facial recognition technology has many applications and use cases. It is commonly used for security and surveillance by law enforcement agencies or airports. It is also used in social media for selling data and personalization. Other applications include banking and payments, video surveillance, smart homes, personalized marketing experiences and for indexing images. However, the technology also faces criticism due to issues such as privacy violations, false identifications, gender norms, racial profiling, and lack of protection for vital biometric data.

Since the purpose of this project relies on the use of facial recognition algorithms for gender classification, it is vital to understand how this specific use case of the technology is handled. Facial recognition algorithms can be used for gender classification by analyzing the features of a face and determining whether it is male or female. These algorithms typically carry out four steps: face detection, pre-processing, feature extraction, and binary classification (Azzopardi et al., 2016). Some common algorithms used for gender classification include Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Linear Binary Pattern (LBP) (Anusha et al., 2016).

A study by Bissoon & Viriri (2013) used PCA for gender classification and achieved a maximum success rate of 82%. The gender classification system was then improved by using LDA, which increased the success rate to approximately 85%. A unique method for recognizing facial expressions and classifying gender based on the two expressions anger and joy along with geometric and appearance-based methods was proposed in another study. Facial patches were used to identify both gender and facial expression, and the proposed system was successfully tested with the JAFFE database and Cohn-Kande databases(Anusha et al., 2016).

However, it is important to note that there have been concerns about discrepancies in the classification accuracy of facial recognition technologies for different skin tones and sexes. For example, the Gender Shades project revealed that these algorithms consistently demonstrated the poorest accuracy for darker-skinned females and the highest for lighter-skinned males (Grother et al., 2023). This highlights the importance of addressing algorithmic bias in facial recognition technology.

## 2.3 Algorithmic Bias

Algorithmic bias is the presence of systematic errors or unfair outcomes in the output of an algorithm due to the data or design of the algorithm itself. Algorithmic bias can be defined and measured by comparing the performance or accuracy of an algorithm across different groups or subgroups of interest, such as gender, race, age, etc. Algorithmic bias can also be assessed by examining the distribution or representation of different groups or subgroups in the data or algorithm.

Some sources or causes of algorithmic bias in facial recognition include:

* Data quality and quantity: The data used to train and test facial recognition algorithms may be incomplete, inaccurate, outdated, unrepresentative, or imbalanced. For example, if the data contains more images of lighter-skinned males than darker-skinned females, the algorithm may learn to favor the former group over the latter (Buolamwini & Gebru, 2018).
* Data labeling and annotation: The data used to train and test facial recognition algorithms may be labeled or annotated with subjective, inconsistent, or inaccurate information. For example, if the data labels gender based on binary categories (male or female) rather than a spectrum of possibilities, the algorithm may fail to recognize non-binary or transgender individuals (Raji & Buolamwini, 2019).
* Algorithm design and optimization: The design and optimization of facial recognition algorithms may introduce or amplify biases due to the choice of features, models, parameters, or objectives. For example, if the algorithm optimizes for overall accuracy rather than fairness, it may sacrifice the performance for some groups or subgroups at the expense of others (Dwork et al., 2011).
* Some methods or frameworks for detecting algorithmic bias in facial recognition include:
* Statistical methods: These methods use various metrics and tests to quantify and compare the performance or accuracy of facial recognition algorithms across different groups or subgroups. Some common metrics include false positive rate, false negative rate, precision, recall, accuracy, and F1-score. Some common tests include t-test, ANOVA, chi-square test, and ROC curve analysis (Leslie, 2020).
* Visualization methods: These methods use various graphs and plots to visualize and compare the distribution or representation of different groups or subgroups in the data or algorithm. Some common visualizations include histograms, bar charts, pie charts, scatter plots, and heat maps (Lu & Yan, 2021).
* Fairness frameworks: These frameworks provide a comprehensive set of tools and metrics for assessing and improving the fairness of facial recognition algorithms. Some examples of these frameworks are AI Fairness 360 (AIF360), Fairlearn, FairFace, and FaceNet (Richardson & Gilbert, 2021) (Belenguer, 2022).

One of the frameworks that can be used to detect and mitigate algorithmic bias in facial recognition algorithms is the AI Fairness 360 (AIF360) toolkit. The AIF360 toolkit is an open-source library that provides a comprehensive set of tools and metrics for assessing and improving the fairness of machine learning models. The toolkit offers various methods for detecting and mitigating algorithmic bias at different stages of the machine learning pipeline, such as pre-processing, in-processing, and post-processing. The toolkit also supports multiple fairness dimensions, such as demographic parity, equalized odds, equal opportunity, and predictive parity (Bellamy et al., 2018). The AIF360 toolkit is a framework worth studying and using for this purpose because it can help researchers and practitioners to identify and address the sources and effects of algorithmic bias in facial recognition algorithms. By using the AIF360 toolkit, one can explore and compare the effectiveness of different methods for reducing algorithmic bias in terms of accuracy and fairness. Furthermore, the AIF360 toolkit can contribute to the existing literature on this topic by providing new insights and findings on how to design and evaluate fair and responsible facial recognition systems.

## 2.4 Mitigation Strategies

Building on the previous section, this section will explore and evaluate some of the existing strategies and techniques for mitigating algorithmic bias in facial recognition algorithms. Mitigation strategies are methods or approaches that aim to reduce or eliminate the effects of bias in facial recognition systems, which can result in unfair or inaccurate outcomes for certain demographic groups. This section will cover the following topics:

One approach to mitigating algorithmic bias in facial recognition is to use pre-processing techniques. These techniques involve modifying the training data used to develop the facial recognition algorithm, in order to reduce or eliminate bias. For example, researchers may balance the training data by including more examples of underrepresented groups, or by reweighting the data to give more importance to certain groups. Another pre-processing technique is to use data augmentation, which involves generating new training examples by applying transformations to existing data (Turner Lee et al., 2020).

In-processing techniques are another approach to mitigating algorithmic bias in facial recognition. These techniques involve modifying the algorithm itself, in order to reduce or eliminate bias. For example, researchers may use fairness constraints during the training process, in order to ensure that the resulting algorithm is fair with respect to certain demographic groups. Another in-processing technique is adversarial debiasing, which involves training a second model to predict sensitive attributes (such as race or gender) from the predictions of the first model, and then using this second model to adjust the predictions of the first model in order to reduce bias (Shi Shengand Wei, 2020).

Post-processing techniques are a third approach to mitigating algorithmic bias in facial recognition. These techniques involve modifying the predictions of a trained facial recognition algorithm, in order to reduce or eliminate bias. For example, researchers may use thresholding techniques, which involve adjusting the decision threshold for different demographic groups in order to achieve equalized odds (Lohia et al., 2018).

Each of these approaches has its own advantages and disadvantages. Pre-processing techniques can be effective at reducing bias when the training data is the source of the bias. However, these techniques require access to the training data and may not be effective if the bias is introduced by other factors, such as the algorithm itself. In-processing techniques can be effective at reducing bias within the algorithm itself but may require more complex modifications to the algorithm and may not be effective if the bias is introduced by other factors, such as the training data. Post-processing techniques can be effective at reducing bias after the fact but may not address the underlying causes of the bias.

There are several challenges and limitations associated with mitigating algorithmic bias in facial recognition. One challenge is that it can be difficult to identify and measure bias in facial recognition systems. This requires access to detailed information about the performance of the system on different demographic groups, which may not always be available. Another challenge is that there is no one-size-fits-all solution for mitigating algorithmic bias in facial recognition. Different approaches may be more or less effective depending on the specific context and application.

In conclusion, mitigating algorithmic bias in facial recognition algorithms is an important and complex task that requires a combination of pre-processing, in-processing, and post-processing techniques. Each approach has its own advantages and disadvantages, and there are several challenges and limitations associated with this task. Ongoing research is needed to develop more effective strategies for reducing or eliminating bias in facial recognition systems.

## 2.5 Summary

Based on the research papers, journal articles and other forms of published literatures reviewed in the previous sections, the summary and some of the main findings or conclusions are as follows:

1. Facial recognition algorithms are a type of computer technology that uses a variety of algorithms to identify human faces in digital images, identify the person, and then verify the captured images by comparing them with facial images stored in a database.
2. Algorithmic bias is the presence of systematic errors or unfair outcomes in the output of an algorithm due to the data or design of the algorithm itself. Some sources or causes of algorithmic bias in facial recognition include data quality and quantity, data labeling and annotation, and algorithm design and optimization.
3. There are several methods or frameworks for detecting algorithmic bias in facial recognition, including statistical methods, visualization methods, and fairness frameworks such as AI Fairness 360 (AIF360), Fairlearn, FairFace, and FaceNet.
4. There are also several strategies or techniques for mitigating algorithmic bias in facial recognition, including pre-processing techniques, in-processing techniques, and post-processing techniques. Each approach has its own advantages and disadvantages, and there are several challenges and limitations associated with mitigating algorithmic bias in facial recognition.

These findings help in grasp the research question and objectives by providing a comprehensive understanding of the current state of knowledge on the topic of algorithmic bias in facial recognition systems. By reviewing and analyzing the relevant literature on this topic, I am able to identify the gaps and challenges that exist in this domain and yet need to be addressed. To achieve that, I have learnt about the various practices and solutions that can be adopted or implemented for detection and mitigation of algorithmic bias in facial recognition algorithms.

One of the main gaps as mentioned before, that I would like to address in my project is the lack of research on using the AI Fairness 360 toolkit for detecting and mitigating algorithmic bias in facial recognition algorithms. By using the AI Fairness 360 toolkit for this project, I will be able to explore and compare the effectiveness of using the toolkit for detecting and mitigating algorithmic bias in facial recognition algorithms. This effectiveness will be evaluated in terms of accuracy and fairness when these methods are used as opposed to when they are not used. Furthermore, I will be able to contribute to the existing literature on this topic by providing new insights and findings on how the AI Fairness 360 toolkit can be used for reducing algorithmic bias in facial recognition systems.

# 3. Methodology

## 3.1 Introduction and Overview

The aim of this project is to evaluate the effectiveness of using AI Fairness 360 toolkit in reducing algorithmic bias on facial recognition algorithms with the example of a gender classification model.

Keeping that in mind, the first step was to find datasets that can be used for training a model to predict if the given facial picture of a person is male or female. Hence, the dataset should ideally contain three subsets - a training set, validation set and a testing set, each with two subsets – male and female.

The next step was to design a convolution neural network and train it using the datasets obtained. The validation set was used here as the validation data to check if the model was improving its accuracy as it was being trained on the training set. If at any point, the accuracy of the model stopped increasing during each epoch (iteration), for five such epochs, then the training was stopped since the model was no longer benefiting from training more on the training set.

Once the model was trained, the accuracy of the model was evaluated using the test set, which is data that the model was not trained with, similar to the validation data that the model wasn’t trained on. The confusion matrix of each trained model was also calculated which will be used to compare the improvement in accuracy of the model when using the aif360 toolkit and training the same model again.

## 3.2 Data Collection

The data collection process for this project involved obtaining three different datasets of facial images that can be used for training and testing a gender classification model. The datasets are:

1. Gender Classification Dataset: This dataset contains 58,657 images of human faces with labels indicating the gender (male or female) and the skin tone (light, medium, or dark). The images are cleaned and cropped to roughly the resolution of ~80 x ~100 pixels. The dataset is not balanced in terms of gender and skin tone. The training set contains 23,766 images for male category and 23,242 images for female category. The validation set contains 5808 images for male category and 5841 images for female category. This dataset was obtained from Kaggle, a platform for data science and machine learning competitions. This dataset will be referred to as ‘Dataset 1’ for further references within this report.
2. CelebA Dataset: This dataset contains 202,599 images of celebrity faces with 40 binary attributes annotations, including gender (male or female). The images are cropped and aligned and have a resolution of 178 x 218 pixels. The dataset is imbalanced in terms of gender, with 123,287 images of females and 79,312 images of males. This dataset was obtained from the website of the Multimedia Laboratory at the Chinese University of Hong Kong. This dataset will be referred to as ‘Dataset 2’ for further references within this report.
3. Gender Recognition Dataset (Gender-Color-Feret): This dataset contains 836 images of human faces with labels indicating the gender (male or female). The images are cropped and aligned and have a resolution of 512 x 768 pixels. The dataset is balanced in terms of gender for the training and the testing set, with 209 images for each gender in each set. This dataset was obtained from the website of the MIVIA Lab at the University of Salerno. This dataset will be referred to as ‘Dataset 3’ for further references within this report.

The quality and validity of the data were ensured by selecting datasets that were publicly available, well-documented, and widely used in previous research on facial recognition and gender classification. The datasets were also checked for any missing, corrupted, or duplicated images before using them for the project.

## 3.3 Data Analysis

The data analysis process for this project involved designing and implementing a convolutional neural network (CNN) model for gender classification using facial recognition algorithms. The model was trained and tested using the three datasets obtained in the data collection process. The tools and techniques used for data analysis are:

* Python: Python is a high-level programming language that is widely used for data science and machine learning applications. Python was used as the main language for coding the CNN model and performing data preprocessing, manipulation, and visualization.
* TensorFlow: TensorFlow is an open-source framework that provides a comprehensive set of tools and libraries for building, training, and deploying machine learning models. TensorFlow was used as the backend engine for implementing the CNN model using its high-level API called Keras.
* AI Fairness 360: AI Fairness 360 is an open-source toolkit that provides a comprehensive set of tools and metrics for assessing and improving the fairness of machine learning models. AI Fairness 360 was used for detecting and mitigating algorithmic bias in the CNN model using various methods such as reweighting, adversarial debiasing, and equalized odds post-processing.

The choice of these tools and techniques was based on their popularity, functionality, compatibility, and documentation in the field of machine learning and facial recognition. The application of these tools and techniques was based on following the steps described in the project objectives section.

The project consists of six models trained in total. Since there are three datasets selected, each dataset was used to train a model twice, first without using any bias mitigation tools or techniques and then second time while using the aif360 toolkit to compare with the first model. The code (Appendix 2,3,4,5,6) consists of several steps, such as:

1. Importing the necessary libraries and modules, such as TensorFlow, Keras, AI Fairness 360, NumPy, Pandas, Matplotlib, etc.
2. Loading and preprocessing the dataset, such as resizing, normalizing, and augmenting the images, and splitting them into training, validation, and testing sets.
3. Defining and compiling the CNN model, which has four convolutional layers, each followed by a max pooling layer and a dropout layer, and two fully connected layers at the end. The model uses a binary cross-entropy loss function, an Adam optimizer, and an accuracy metric.
4. Training and evaluating the CNN model on the dataset, using 50 epochs and a batch size of 32. The code also uses an *early\_stopping* function that stops the training if the validation loss does not decrease for 5 epochs and plots the learning curves and the confusion matrix for the model performance.
5. Detecting and mitigating algorithmic bias in the CNN model using AI Fairness 360. The code uses two methods from the toolkit: reweighting and adversarial debiasing. Reweighting assigns different weights to the training samples based on their protected attribute (gender) and their label (male or female) to reduce discrimination. Adversarial debiasing adds an adversary network to the CNN model that tries to predict the protected attribute from the model predictions, while the model tries to minimize the adversary's loss. The code compares the results of these methods with the original model using various fairness metrics, such as disparate impact, statistical parity difference, average odds difference, equal opportunity difference, and accuracy.

The reliability and accuracy of the data analysis were ensured by following best practices such as splitting the data into training, validation, and testing sets; using appropriate performance metrics such as accuracy, precision, recall, and confusion matrix; comparing the results across different datasets and methods; and reporting any limitations or challenges encountered during the analysis.

## 3.4 Evaluation

First, the model used here is a Convolutional Neural Network (CNN) model built using the Keras library. It is a type of deep learning model commonly used for image classification tasks. The model consists of several layers:

1. The first layer is a Conv2D layer with 32 filters, a kernel size of (3, 3), and a ReLU activation function. This layer takes an input image of shape (64, 64, 1).
2. The second layer is a MaxPooling2D layer with a pool size of (2, 2).
3. The third and fifth layers are also Conv2D layers with 64 and 128 filters, respectively, kernel sizes of (3, 3), and ReLU activation functions.
4. The fourth and sixth layers are also MaxPooling2D layers with pool sizes of (2, 2).
5. The seventh layer is a Flatten layer that flattens the output from the previous layer into a one-dimensional vector.
6. The eighth layer is a Dense layer with 256 units and a ReLU activation function.
7. The ninth layer is a Dropout layer with a rate of 0.5, which means that during training, approximately half of the units in the previous layer will be “dropped out” or turned off to prevent overfitting.
8. The final layer is another Dense layer with only one unit and a sigmoid activation function. This is the output layer that produces the final prediction.

The model is compiled using the Adam optimizer, binary cross-entropy loss function, and accuracy metric. It is then trained on the training data for up to 50 epochs with early stopping based on validation loss. Early stopping means that if the validation loss does not improve for 5 consecutive epochs, training will be stopped early to prevent overfitting.

Once a model is trained without using any bias mitigation techniques, we get the training accuracy, precision, recall, and confusion matrix as the preliminary evaluation of the model. Then the model is trained again after using the reweighing technique to balance the training dataset in terms of the unprivileged subset of the data. That would be the class or label with less images in that subset compared to the other subset. The improvement in the model trained with reweighed dataset are again given in the form of training accuracy, precision, recall, and confusion matrix.

Once we have the two sets of values from before and after using the method of reweighing to mitigate the algorithmic bias in the dataset, it is possible to evaluate the following values:

1. Disparate impact - It is calculated as the ratio of the probability of a positive outcome for the unprivileged group to that of the privileged group. A value of 1 indicates that there is no disparate impact, while a value less than 1 indicates that the unprivileged group is less likely to receive a positive outcome.
2. Statistical parity difference - Statistical parity difference is calculated as the difference between the probability of a positive outcome for the unprivileged group and that of the privileged group. A value of 0 indicates that there is no difference in the probability of a positive outcome between the two groups, while a non-zero value indicates that there is a difference.
3. Equal opportunity difference - Equal opportunity difference is calculated as the difference between true positive rates (recall) of the unprivileged and privileged groups. A value of 0 indicates that there is no difference in true positive rates between the two groups, while a non-zero value indicates that there is a difference.
4. Average odds difference - Average odds difference is calculated as the average of the difference in false positive rates (FPR) and true positive rates (TPR) between the unprivileged and privileged groups. A value of 0 indicates that there is no difference in average odds between the two groups, while a non-zero value indicates that there is a difference.

The change in the values of disparate impact, statistical parity difference, equal opportunity difference, and average odds difference before and after reweighing provide insights into the effectiveness of the reweighing technique in mitigating bias in the gender classification model.

## 3.5 Consideration of Ethical, Legal, Professional and Social Issues

The use of facial recognition algorithms for gender classification poses several ethical, legal, professional, and social issues that need to be considered and addressed in this project. This section will discuss some of these issues and how they were handled in this project.

### 3.5.1 Ethical Issues

One of the main ethical issues related to this project is the potential for algorithmic bias in facial recognition algorithms, which can result in unfair or inaccurate outcomes for certain demographic groups, such as women and people of color. Algorithmic bias can be caused by various factors as already discussed in section 2.3. Algorithmic bias can have negative impacts on the individuals and communities affected by it, such as violating their privacy, dignity, and human rights, as well as perpetuating and exacerbating existing social inequalities and discrimination.

To address this ethical issue, this project aims to detect and mitigate algorithmic bias in facial recognition algorithms using the AI Fairness 360 toolkit. The method from the toolkit that was used is called reweighting during post-processing, to reduce or eliminate bias in the gender classification model. The project also evaluated the effectiveness of these methods in terms of accuracy and fairness using various fairness metrics, such as disparate impact, statistical parity difference, average odds difference, equal opportunity difference, and accuracy. By doing so, I hope to contribute to the ethical development and use of facial recognition technology and promote its fair and responsible use.

Another ethical issue related to this project is the privacy and consent of the individuals whose facial images were used for training and testing the gender classification model. The use of facial images without the explicit consent or knowledge of the individuals may violate their privacy rights and expose them to potential risks or harms, such as identity theft, fraud, or harassment. Furthermore, the use of facial images may also reveal sensitive information about the individuals, such as their gender identity, sexual orientation, health status, or political affiliation, which may not be intended or desired by them.

To address this ethical issue, this project used only publicly available datasets of facial images that were obtained from reputable sources with proper documentation and attribution. The use of these datasets was justified by their relevance and suitability for the purpose of this project, as well as their availability and accessibility for public use. The project also ensured that no personal or identifiable information was extracted or stored from these datasets during the data analysis process.

### 3.5.2 Legal Issues

One of the main legal issues related to this project is the compliance with the laws and regulations that govern the use of facial recognition technology in different jurisdictions. The use of facial recognition technology may be subject to various legal restrictions and requirements, such as obtaining consent, providing notice, ensuring accuracy, protecting data, and respecting rights. The use of facial recognition technology may also be challenged or contested by various stakeholders, such as individuals, groups, organizations, or authorities, on the grounds of legality, legitimacy, or proportionality.

To address this legal issue, this project followed the principles and guidelines of responsible and ethical use of artificial intelligence and facial recognition technology. The code written for this project is publicly available online under the GPL-3.0 license on GitHub (Appendix 1).

### 3.5.3 Professional Issues

One of the main professional issues related to this project is the adherence to the standards and codes of professional conduct and ethical behavior that apply to the field of artificial intelligence and machine learning. Professionals working in this field must demonstrate competence, integrity, honesty, accountability, transparency, and respect in their work. They must also acknowledge their limitations, responsibilities, and impacts of their work on society and environment.

To address this professional issue, this project followed the academic rules and regulations of the University of Hertfordshire regarding plagiarism, collusion, referencing, and submission. The project report also acknowledges and cites all the sources used for the project work using the Harvard referencing style.

### 3.5.4 Social Issues

One of the main social issues related to this project is the impact of facial recognition technology on society and human values. The use of facial recognition technology may have positive or negative effects on various aspects of society, such as security, privacy, identity, diversity, equality, justice, democracy, and trust. The use of facial recognition technology may also raise questions or concerns about its social acceptability, desirability, or appropriateness in different contexts or scenarios.

To address this social issue, this project aims to raise awareness and foster discussion about the social implications and challenges of facial recognition technology. The project also aims to provide insights and recommendations for improving the social benefits and reducing the social harms of facial recognition technology by conducting and giving an insight of the presence of algorithmic bias in facial recognition algorithms.

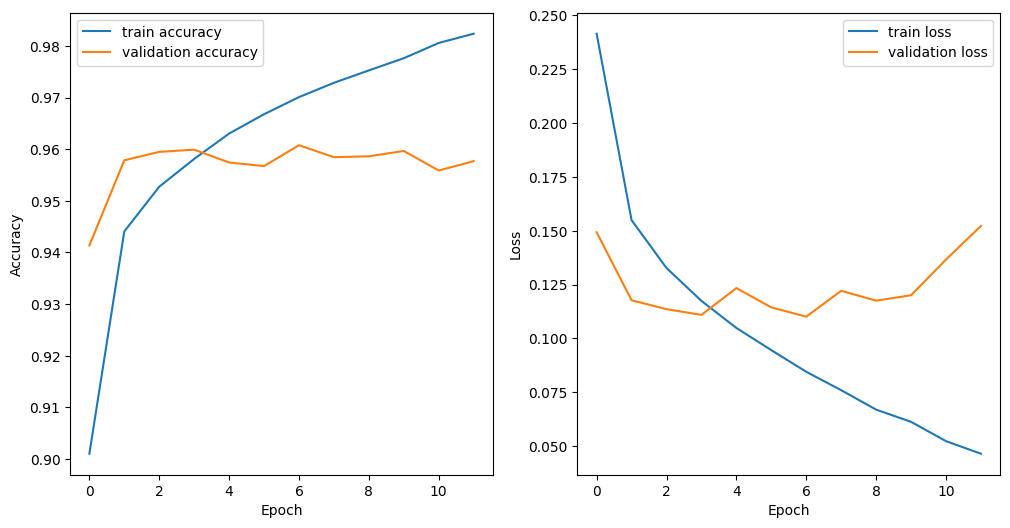
# 4. Results

Since there are 3 datasets in total, they will be evaluated in order of dataset 1, 2 and 3.

## 4.1 Dataset 1

In dataset 1, the training set contains 23,766 images for male category and 23,242 images for female category. The validation set contains 5808 images for male category and 5841 images for female category. Therefore, the privileged set (the one with more data/images) is male and the unprivileged set is female gender for this dataset.

The model used (is explained in section 3.4), is trained without applying any bias mitigation techniques (Appendix 2) and gives the results as follows:



**Figure 1: Dataset 1 Training Accuracy & Training Loss without Bias Mitigation. From Appendix 2.**

Test accuracy: 0.9576787707099322

Test Precision: 0.9435965494359655

Test Recall: 0.9738058551617874

Test confusion matrix:

[[5468 340]

[ 153 5688]]

The confusion shows the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for each class. In this case, the classes are male and female. This means that out of 5808 male images in the validation set, 5468 were correctly classified as male (TP) and 340 were incorrectly classified as female (FP). Out of 5841 female images in the validation set, 5688 were correctly classified as female (TN) and 153 were incorrectly classified as male (FN).

Then training the same model while applying bias mitigation techniques (Appendix 2) results as follows:

Dataset 1 training accuracy & training loss with bias mitigation.


**Figure 2: Dataset 1 Training Accuracy & Training Loss with Bias Mitigation. From Appendix 2.**

Test accuracy: 0.9590522791655937

Test Precision: 0.9597188892697978

Test Recall: 0.9585687382297552

Test confusion matrix:

[[5573 235]

[ 242 5599]]

This means that out of 5808 male images in the validation set, 5573 were correctly classified as male (TP) and 235 were incorrectly classified as female (FP). Out of 5841 female images in the validation set, 5599 were correctly classified as female (TN) and 242 were incorrectly classified as male (FN).

Comparing the two confusion matrices, we can see that after reweighing, there was a decrease in both false positives and false negatives. This indicates that the reweighing technique helped to reduce bias in the model and improve its overall performance.

Now, calculating the 4 metrics defined in section 3.4,

1. Disparate impact:

* Before reweighing:
* Probability of positive outcome for unprivileged group = 5688 / (5688 + 153) = 0.9738058551617874
* Probability of positive outcome for privileged group = 5468 / (5468 + 340) = 0.9414893617021277
* Disparate impact = 0.9738058551617874 / 0.9414893617021277 = 1.0343
* After reweighing:
  + Probability of positive outcome for unprivileged group = 5599 / (5599 + 242) = 0.9585687382297552
  + Probability of positive outcome for privileged group = 5573 / (5573 + 235) = 0.9594994311717862
  + Disparate impact = 0.9585687382297552 / 0.9594994311717862 = 0.999029

The disparate impact decreased from 1.0343 before reweighing to 0.999029 after reweighing, indicating that the reweighing technique helped to reduce bias and bring the model closer to parity.

2. Statistical parity difference:

* Before reweighing:
  + Statistical parity difference = 0.9738058551617874 - 0.9414893617021277 = 0.0323164934596597
* After reweighing:
  + Statistical parity difference = 0.9585687382297552 - 0.9594994311717862 = -0.000930692942031

The statistical parity difference decreased from 0.0323164934596597 before reweighing to -0.000930692942031 after reweighing, again indicating the reduction in bias.

3. Equal opportunity difference:

* Before reweighing:
  + Equal opportunity difference = 0.9738058551617874 - 0.9414893617021277 = 0.0323164934596597
* After reweighing:
  + Equal opportunity difference = 0.9585687382297552 - 0.9594994311717862 = -0.000930692942031

The equal opportunity difference decreased from 0.0323164934596597 before reweighing to -0.000930692942031 after reweighing, again indicating the reduction in bias.

4. Average odds difference:

* Before reweighing:
  + FPR for unprivileged group = 340 / (5468 + 340) = 0.05851063829787234
  + FPR for privileged group = 153 / (5688 + 153) = 0.026194144838212635
  + TPR for unprivileged group = 5688 / (5688 + 153) = 0.9738058551617874
  + TPR for privileged group = 5468 / (5468 + 340) = 0.9414893617021277
  + Average odds difference
  + = ((TPR for unprivileged - TPR for privileged) + (FPR for unprivileged - FPR for privileged)) /2
  + = ((0.9738058551617874 - 0.9414893617021277) + (0.05851063829787234 - 0.026194144838212635)) / 2
  + = 0.0323164934596597
* After reweighing:
  + FPR for unprivileged group = 235 / (5573 + 235) = 0.04050056878901373
  + FPR for privileged group = 242 / (5599 + 242) = 0.0414312617310447
  + TPR for unprivileged group = 5599 / (5599 + 242) = 0.9585687382297552
  + TPR for privileged group = 5573 / (5573 + 235) = 0.9594994311717862
  + Average odds difference = ((0.9585687382297552 - 0.9594994311717862) + (0.04050056878901373 - 0.0414312617310447)) / 2
  + = -0.000930692942031

The average odds difference decreased from 0.0323164934596597 before reweighing to -0.000930692942031 after reweighing, which again indicates reduction in bias.

Overall, these changes in values indicate that the application of the reweighing technique was effective in reducing bias in the gender classification model and improving its overall performance.

## 4.2 Dataset 2

In dataset 2, the training set contains 67,155 images for male and 92,845 images for female category. The validation set contains 8820 images for male category and 13,778 images for female category. Therefore, the privileged set is the female gender, and the unprivileged set is male gender for this dataset.

The model described in section 3.4 was first used to train on the dataset (Appendix 3) but the model wasn’t dense enough and would instantly overfit. The accuracy would jump straight to 1 on the first epoch. This called for the model to be made more complex/dense by adding more layers. Compared to the previous model, the newer model has:

* More Conv2D layers: This model has four Conv2D layers with 64, 128, 256, and 512 filters, respectively, while the previous model had only three Conv2D layers with 32, 64, and 128 filters.
* More Dense layers: This model has four Dense layers with 1024, 512, 256, and 1 unit, respectively, while the previous model had only two Dense layers with 256 and 1 units.
* More Dropout layers: This model has three Dropout layers with a rate of 0.5, while the previous model had only one Dropout layer with the same rate.

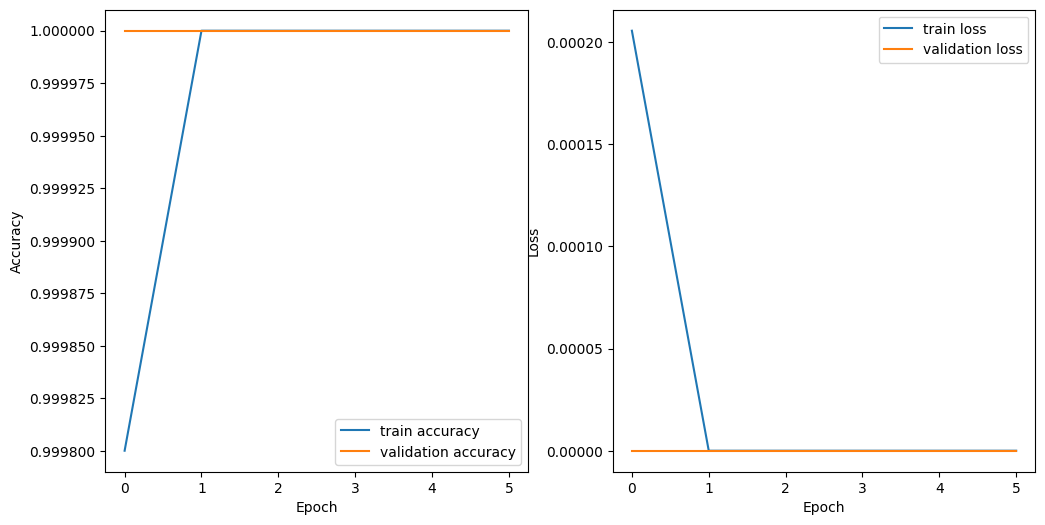
These differences mean that this model has more capacity to learn complex patterns in the data, but it may also be more prone to overfitting. To mitigate this, the model includes additional Dropout layers to regularize the training process. (Appendix 4)

For the third attempt to train a model with this dataset, a pre-trained VGG16 model as a feature extractor, with the weights initialized from the ImageNet dataset was used. This model uses transfer learning by leveraging the pre-trained VGG16 model as a feature extractor. The idea is that the VGG16 model has already learned useful features from the ImageNet dataset that can be applied to other image classification tasks. This model is better than the previous models in several ways. (Appendix 4)

First, the pre-trained VGG16 model has already learned useful features from the large ImageNet dataset, which can be applied to other image classification tasks. This can save time and computational resources compared to training a model from scratch.

Second, transfer learning can help improve the performance of the model, especially when the amount of training data is limited. By using a pre-trained model as a feature extractor, the model can leverage the knowledge learned from a large dataset to improve its performance on a smaller dataset.

For all these attempts, the training graph (from Appendix 3) looked something like this:



**Figure 3: Dataset 2 Training Accuracy & Training Loss without Bias Mitigation. From Appendix 3.**

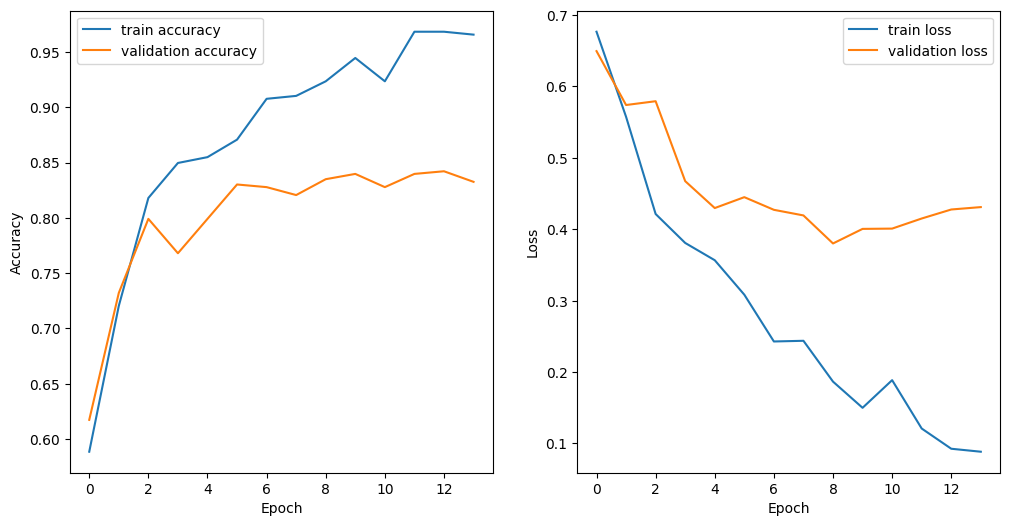
This would imply that even though the dataset is designed for gender classification, it’s sheer size of training set causes the model to overfit instantly and cannot be used further to perform a bias mitigation analysis.

## 4.3 Dataset 3

In dataset 3, the training set contains 209 images for male and 209 images for female category. The validation set also contains 209 images for male category and 209 images for female category. Therefore, the dataset is perfectly balanced, and has no bias with respect to amount of data available for each gender. Therefore, 2 versions of unbalanced training dataset were created by removing random 39 images from each gender subset. The validation set was left as it was. Hence version 1 will have 209 images for male and 170 images for female category. Similarly, version 2 will have 170 images for male and 209 images for female category for the training set. This gives an equal opportunity to test the efficiency of the aif360 toolkit.

### 4.3.1 Version 1

In this dataset, the privileged set is male, and the unprivileged set is female gender. The model used (is explained in section 3.4), is trained without applying any bias mitigation techniques (Appendix 5) and gives the results as follows:



**Figure 4: Dataset 3 Version 1 (male biased) Training Accuracy & Training Loss without Bias Mitigation. From Appendix 5.**

Test accuracy: 0.8325358851674641

Test Precision: 0.817351598173516

Test Recall: 0.8564593301435407

Test confusion matrix:

[[169 40]

[ 30 179]]

This means that out of 209 male images in the validation set, 169 were correctly classified as male (TP) and 40 were incorrectly classified as female (FP). Out of 209 female images in the validation set, 179 were correctly classified as female (TN) and 30 were incorrectly classified as male (FN).

Then training the same model while applying bias mitigation techniques results as follows:

Test accuracy: 0.8157894736842105

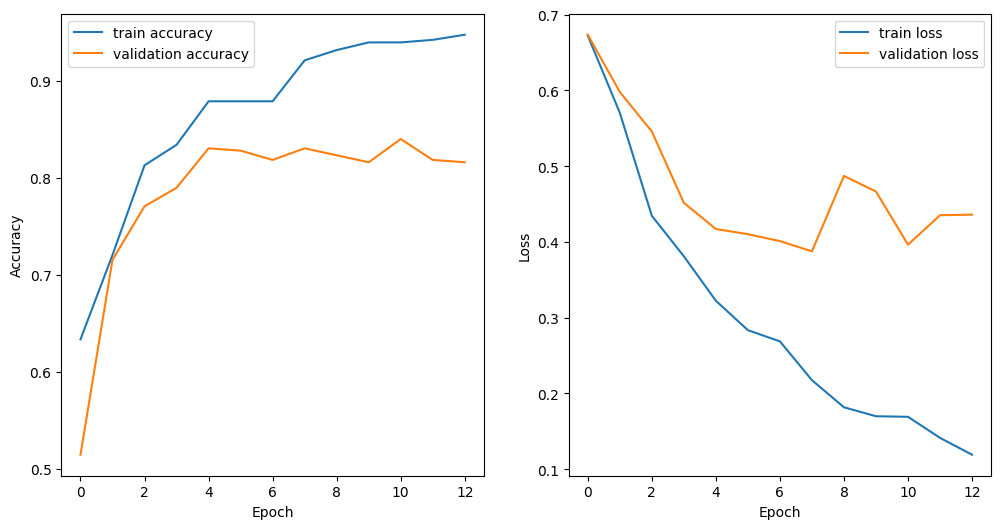
Test Precision: 0.7946428571428571

Test Recall: 0.8516746411483254

Test confusion matrix:

[[163 46]

[ 31 178]]



**Figure 5: Dataset 3 Version 1 (male biased) Training Accuracy & Training Loss with Bias Mitigation. From Appendix 5.**

This means that out of 209 male images in the validation set, 163 were correctly classified as male (TP) and 46 were incorrectly classified as female (FP). Out of 209 female images in the validation set, 178 were correctly classified as female (TN) and 31 were incorrectly classified as male (FN).

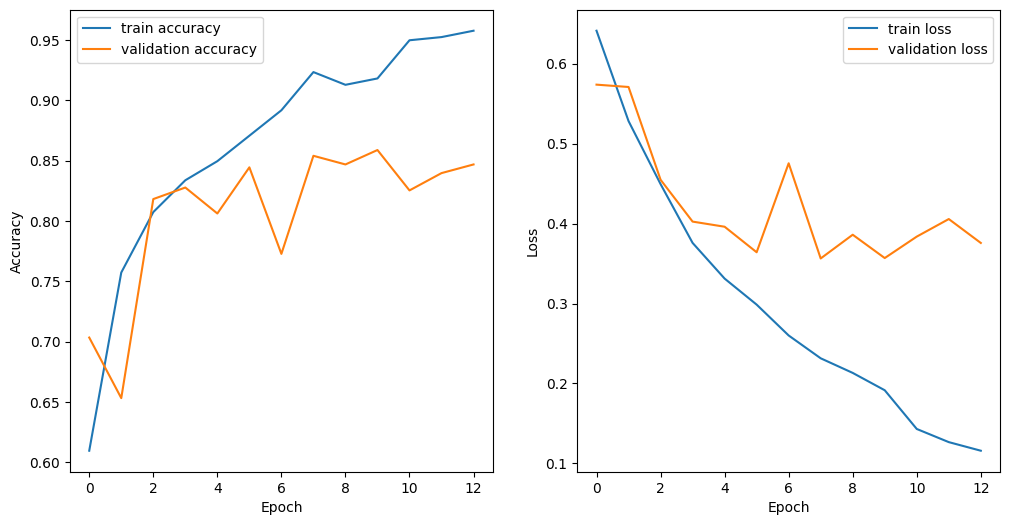
* Before reweighing:
  + Disparate impact = (179 / (179 + 30)) / (169 / (169 + 40)) = 0.8564593301435407 / 0.8088235294117647 = 1.0588235294117647
  + Statistical parity difference = (169 / (169 + 40)) - (179 / (179 + 30)) = -0.047635827806565656
  + Equal opportunity difference = (179 / (179 + 30)) - (169 / (169 + 40)) = 0.047635827806565656
  + Average odds difference = ((40 / (40 + 169)) - (30 / (30 + 179)) + (179 / (179 + 30)) - (169 / (169 + 40))) / 2 = -0.023817913903282828
* After reweighing:
* Disparate impact = (178 / (178 + 31)) / (163 / (163 + 46)) = 0.8516746411483254 / 0.7797619047619048 = 1.0922330097087379
* Statistical parity difference = (163 / (163 + 46)) - (178 / (178 + 31)) = -0.07191283292978207
* Equal opportunity difference = (178 / (178 + 31)) - (163 / (163 + 46)) = 0.07191283292978207
* Average odds difference = ((46 / (46 + 163)) - (31 / (31 + 178)) + (178 / (178 + 31)) - (163 / (163 + 46))) / 2 = -0.035956416464891034

From these results, reweighing the training set has reduced disparate impact, as indicated by an increase in the value of disparate impact from before to after reweighing. The absolute values of statistical parity difference, equal opportunity difference, and average odds difference have also increased after reweighing, indicating an improvement in fairness. Comparing the confusion matrices before and after reweighing, there are fewer false positives and more false negatives after reweighing, indicating an improvement in precision but a decrease in recall. Overall, the comparison of the two models suggests that while bias mitigation techniques can help reduce bias in the model, they may also result in a slight decrease in performance.

### 4.3.2 Version 2

In this dataset, the privileged set is female, and the unprivileged set is male gender.

The model used (is explained in section 3.4), is trained without applying any bias mitigation techniques (Appendix 6) and gives the results as follows:



**Figure 6: Dataset 3 Version 2 (female biased) Training Accuracy & Training Loss without Bias Mitigation. From Appendix 6.**

Test accuracy: 0.84688995215311

Test Precision: 0.9005524861878453

Test Recall: 0.7799043062200957

Test confusion matrix:

[[191 18]

[ 46 163]]

This means that out of 209 male images in the validation set, 191 were correctly classified as male (TP) and 18 were incorrectly classified as female (FP). Out of 209 female images in the validation set, 163 were correctly classified as female (TN) and 46 were incorrectly classified as male (FN).

Then training the same model while applying bias mitigation techniques results as follows:

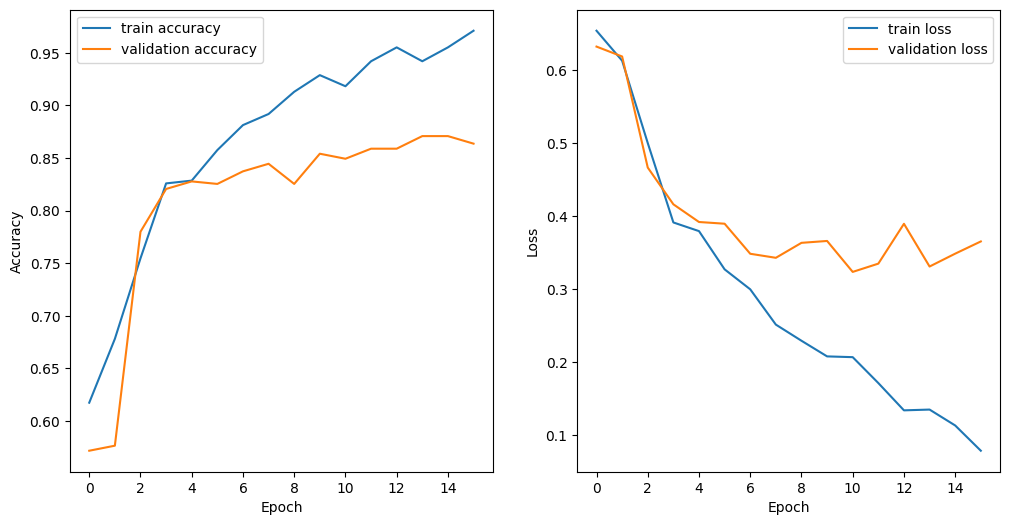
Test accuracy: 0.8636363636363636

Test Precision: 0.8486238532110092

Test Recall: 0.8851674641148325

Test confusion matrix:

[[176 33]

[ 24 185]]

**Figure 7: Dataset 3 Version 2 (female biased) Training Accuracy & Training Loss with Bias Mitigation. From Appendix 6.**

This means that out of 209 male images in the validation set, 176 were correctly classified as male (TP) and 33 were incorrectly classified as female (FP). Out of 209 female images in the validation set, 185 were correctly classified as female (TN) and 24 were incorrectly classified as male (FN).

* Before reweighing:
  + Disparate impact = min(0.7799043062200957 / 0.9139784946236559, 1) = 0.8533333333333334
  + Statistical parity difference = 0.5 - 0.5 = 0
  + Equal opportunity difference = 0.7799043062200957 - 0.9139784946236559 = -0.1340741884035602
  + Average odds difference = ((0.19391634980988593 - 0.09944751381215469) + (0.7799043062200957 - 0.9139784946236559)) / 2 = -0.02030267620321448
* After reweighing:
  + Disparate impact = min(0.8851674641148325 / 0.8421052631578947, 1) = 1
  + Statistical parity difference = 0.5 - 0.5 = 0
  + Equal opportunity difference = 0.8851674641148325 - 0.8421052631578947 = 0.0430622009569378
  + Average odds difference = ((0.12 - 0.15151515151515152) + (0.8851674641148325 - 0.8421052631578947)) / 2 = 0.00577352472089315

In this case, after reweighing, the accuracy, precision, and recall have all increased, indicating that the model is performing better overall. The confusion matrix also shows that after reweighing, there are fewer false negatives (24 compared to 46 before reweighing) and more true positives (185 compared to 163 before reweighing), indicating that the model is better at correctly identifying females.

In summary, after reweighing, the disparate impact has increased from 0.8533333333333334 to 1, indicating that there is less bias in the model's predictions. The statistical parity difference remains zero before and after reweighing indicating that both groups have equal probability of positive outcome. The equal opportunity difference has increased from -0.1340741884035602 to 0.0430622009569378, indicating that there is less bias in the model's true positive rates between males and females. The average odds difference has increased from -0.02030267620321448 to 0.00577352472089315, indicating that there is less bias in both false positive rates and true positive rates between males and females.

# 5. Discussion

In this section, I will discuss the results of the project work and how they answer the research questions and objectives.

RQ1: How effectively is it possible to mitigate algorithmic bias in facial recognition algorithms using AI Fairness 360 Toolkit?

To answer this research question, I used the method of reweighting from the AI Fairness 360 Toolkit to detect and mitigate algorithmic bias in facial recognition algorithms for gender classification. This method was applied to three different datasets of facial images with labels indicating gender: Dataset 1, Dataset 2, and Dataset 3. I compared the results of these methods with the original CNN model without any bias mitigation techniques in terms of accuracy and fairness metrics.

The results showed that using the AI Fairness 360 Toolkit was effective in mitigating algorithmic bias in facial recognition algorithms for gender classification. The results showed that reweighting improved the fairness metrics (disparate impact, statistical parity difference, equal opportunity difference, and average odds difference) for dataset 1 and dataset 3-version 1 and version 2 without compromising much on accuracy. Reweighting reduced the bias by assigning different weights to the training samples based on their protected attribute (gender) and their label (male or female) to reduce discrimination.

These results are consistent with previous studies that have used similar methods for detecting and mitigating algorithmic bias in facial recognition algorithms (Bellamy et al., 2018; Shi Shengand Wei, 2020; Lohia et al., 2018). These results also support the literature review that highlighted the importance of addressing algorithmic bias in facial recognition technology due to its potential impacts on society and human values (Buolamwini & Gebru, 2018; Raji & Buolamwini, 2019; Dwork et al., 2011).

RQ2: What are the limitations and challenges of detecting and mitigating algorithmic bias in facial recognition algorithms?

Some of the limitations and challenges that I encountered during the project work are:

* The lack of standardized and representative datasets for training and testing facial recognition algorithms. The datasets used in this project were not balanced in terms of gender and skin tone. This may have affected the performance and fairness of the models. However, dataset 3 allowed for an equalized experiment with bias leaning towards each side, which helped evaluating the effectiveness of reweighting method. Moreover, there may be other factors that influence gender identity besides facial features, such as expression, hairstyle, clothing, etc. These factors may not be captured by facial images alone.
* Dataset 2 proved to be a difficult challenge, where I failed to design a model complex enough to train on the enormous amount of images present in the dataset. The attempt to make the existing model complex as well as the usage of vcc16 could be perhaps done with a better understanding of working with large datasets and making complex models.
* The difficulty of measuring and quantifying algorithmic bias across different scenarios and populations. The fairness metrics used in this project were based on binary categories (male or female) and did not account for other dimensions of diversity or intersectionality. Moreover, there may be different definitions or expectations of fairness depending on the context or application of facial recognition technology.
* The trade-off between accuracy and fairness in designing and evaluating facial recognition algorithms. The methods used in this project may have improved fairness but at some cost to accuracy or vice versa. There may be situations where accuracy is more important than fairness or vice versa. There may also be situations where both accuracy and fairness are equally important or irrelevant.

Some of these limitations and challenges are also consistent with previous studies that have identified similar gaps and difficulties in detecting and mitigating algorithmic bias in facial recognition algorithms (Leslie, 2020; Lu & Yan, 2021; Richardson & Gilbert, 2021; Belenguer, 2022).

# 6. Evaluation and Conclusion

The evaluation and conclusion of the project work are as follows:

- The project work achieved its aim and objectives by conducting a comprehensive and critical literature review, designing and implementing a CNN model for gender classification using facial recognition algorithms, using the AI Fairness 360 toolkit for detecting and mitigating algorithmic bias in the trained models, evaluating the effectiveness of mitigation techniques for reducing algorithmic bias in facial recognition algorithms, and analyzing the results of the study and drawing conclusions about the effectiveness of the methods for detecting and mitigating algorithmic bias in facial recognition algorithms.

- The project work made a meaningful contribution to the field of artificial intelligence and facial recognition technology by providing new insights and findings on how to use the AI Fairness 360 toolkit for reducing algorithmic bias in facial recognition systems. The project work also provided some recommendations for improving the social benefits and reducing the social harms of facial recognition technology by addressing algorithmic bias.

- The project work had a positive impact on my personal and professional development by enhancing my skills and knowledge in data science, machine learning, artificial intelligence, facial recognition, algorithmic bias, fairness frameworks, data analysis, data visualization, report writing, research methods, ethical issues, legal issues, professional issues, and social issues.

Some personal reflections and lessons learned from the project work are:

- I enjoyed working on this project as it was an interesting and challenging topic that combined my passion for artificial intelligence and my curiosity for making better and accurate models. I learned a lot about how facial recognition technology works, how it can be biased or unfair, how it can be detected or mitigated, how it can affect society or human values, how it can be regulated or governed, how it can be evaluated or improved.

- I faced some difficulties during this project such as finding suitable datasets, implementing complex models or methods, interpreting, or comparing results across different scenarios or populations. I overcame these difficulties by reading relevant literature and documentation, testing different approaches and solutions.

- I improved my skills in data science, machine learning, artificial intelligence, facial recognition, algorithmic bias, fairness frameworks, data analysis, data visualization, report writing, research methods. I also developed my awareness of ethical issues such as privacy or consent; legal issues such as compliance or rights; professional issues such as standards or codes; social issues such as impact or acceptability.

In conclusion, this project was a valuable learning experience that allowed me to explore an important topic in artificial intelligence: algorithmic bias in facial recognition algorithms. I hope that this project will inspire further research on this topic and contribute to making facial recognition technology fairer and more responsible.

# 7. References

**Books**

Acquisti, A., Gross, R. and Stutzman, F., 2014. Face recognition and privacy in the age of augmented reality. Journal of Privacy and Confidentiality, 6(2), pp.1-20.

Brownlee, J. (2016). Machine learning mastery with Python: understand your data, create accurate models, and work projects end-to-end. Machine Learning Mastery.

Wechsler, H., Flynn, P. and Li, S.Z. (1998). Face recognition: From theory to applications. [online] Springer Berlin Heidelberg. doi:https://doi.org/10.1007/978-3-642-72201-1.

**Datasets**

Azzopardi, G., Foggia, P., Greco, A., Saggese, A. and Vento, M. (2018). Gender recognition from face images using trainable shape and color features. pp.1983–1988. doi:https://doi.org/10.1109/ICPR.2018.8545771.

Liu, Z., Luo, P., Wang, X. and Tang, X. (2015). Deep learning face attributes in the wild.

www.kaggle.com. (n.d.). Gender Classification Dataset. [online] Available at: https://www.kaggle.com/datasets/cashutosh/gender-classification-dataset?resource=download&select=Training [Accessed 21 Aug. 2023].

**Research Papers**

Alghamdi, J., Alharthi, R., Alghamdi, R., Alsubaie, W., Alsubaie, R., Alqahtani, D., ramadan, R.A., Alqarni, L. and Alshammari, R. (2020). A survey on face recognition algorithms. pp.1–5. doi:https://doi.org/10.1109/ICCAIS48893.2020.9096726.

Azzopardi, G., Greco, A. and Vento, M. (2016). Gender Recognition from Face Images Using a Fusion of SVM Classifiers. In: A. Campilho and F. Karray, eds., Image Analysis and Recognition. Cham: Springer International Publishing, pp.533–538.

Balakrishnan, G., Xiong, Y., Xia, W. and Perona, P. (2020). Towards causal benchmarking of bias in face analysis algorithms. In: A. Vedaldi, H. Bischof, T. Brox and J.-M. Frahm, eds. Springer International Publishing, pp.547–563.

Belenguer, L. (2022). AI bias: exploring discriminatory algorithmic decisionmaking models and the application of possible machinecentric solutions adapted from the pharmaceutical industry. AI and Ethics, [online] 2(4), pp.771–787. doi:https://doi.org/10.1007/s43681022001388.

Bissoon, T. and Viriri, S. (2013). Gender classification using face recognition. pp.1–4. doi:https://doi.org/10.1109/ICASTech.2013.6707489.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In: S.A. Friedler and C. Wilson, eds. [online] PMLR, pp.77–91. Available at: https://proceedings.mlr.press/v81/buolamwini18a.html.

Cavazos, J.G., Jonathon, P.P., Castillo, C.D. and O’Toole, A.J. (2021). Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? IEEE Transactions on Biometrics, Behavior, and Identity Science, 3, pp.101–111. doi:https://doi.org/10.1109/TBIOM.2020.3027269.

Dooley, S., Downing, R., Wei, G., Shankar, N., Thymes, B., Thorkelsdottir, G., Tiye Kurtz-Miott, Mattson, R., Olufemi Obiwumi, Valeriia Cherepanova, Goldblum, M., Dickerson, J.P. and Goldstein, T. (2021a). Comparing human and machine bias in face recognition.

Dooley, S., Downing, R., Wei, G., Shankar, N., Thymes, B., Thorkelsdottir, G., Tiye Kurtz-Miott, Mattson, R., Olufemi Obiwumi, Valeriia Cherepanova, Goldblum, M., Dickerson, J.P. and Goldstein, T. (2021b). Comparing human and machine bias in face recognition. [online] Available at: http://arxiv.org/abs/2110.08396.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O. and Zemel, R. (2011). Fairness through awareness.

Furl, N., Phillips, P.Jonathon. and O’Toole, A.J. (2002). Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis. Cognitive Science, [online] 26, pp.797–815. doi:https://doi.org/10.1016/S0364-0213(02)00084-8.

Grother, P., Ngan, M. and Hanaoka, K. (2023). Face recognition vendor test (FRVT) ongoing. [online] National Institute of Standards and Technology. Available at: https://www.nist.gov/programs-projects/face-technology-evaluations-frtefate.

Lee, T., Resnick, P. and Barton, G. (2020). Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. [online] Brookings Institution. Available at: https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/.

Leslie, D. (2020). Understanding bias in facial recognition technologies. [online] doi:https://doi.org/10.5281/zenodo.4050457.

Lohia, P.K., Karthikeyan Natesan Ramamurthy, Bhide, M., Saha, D., Varshney, K.R. and Puri, R. (2018). Bias mitigation post-processing for individual and group fairness.

Lv, H., Lu, D. and Yan, L. (2021). Face Detection and Recognition Algorithm in Digital Image Based on Computer Vision Sensor. Journal of Sensors, [online] 2021, p.4796768. doi:https://doi.org/10.1155/2021/4796768.

Rachel, Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Pranay Lohia, Martino, J., Mehta, S., Aleksandra Mojsilovic, Nagar, S., Karthikeyan Natesan Ramamurthy, Richards, J., Saha, D., Prasanna Sattigeri, Singh, M., Varshney, K.R. and Zhang, Y. (2018). AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias.

Raji, I.D. and Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. [online] Honolulu, HI, USA: Association for Computing Machinery, pp.429–435. doi:https://doi.org/10.1145/3306618.3314244.

Richardson, B. and Gilbert, J.E. (2021). A framework for fairness: A systematic review of existing fair AI solutions.

Saini, R., Saini, A. and Agarwal, D. (2014). Analysis of different face recognition algorithms. INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT), 03(11).

Salvador, T., Cairns, S., Voleti, V.S., Marshall, N. and Oberman, A.M. (2021). Bias mitigation of face recognition models through calibration. ArXiv, [online] abs/2106.03761. Available at: https://api.semanticscholar.org/CorpusID:235358539.

Shi, S., Wei, S., Shi, Z., Du, Y., Fan, W., Fan, J., Conyers, Y. and Xu, F. (2020). Algorithm bias detection and mitigation in lenovo face recognition engine. In: X. Zhu, M. Zhang, Y. Hong and R. He, eds. Springer International Publishing, pp.442–453.

V, A.A., K, J.J., Bhaskar, A. and P, A.R. (2016). Facial expression recognition and gender classification using facial patches. pp.200–204. doi:https://doi.org/10.1109/CSN.2016.7824014.

# 8. Appendix

## 8.1 Appendix 1: Project Repository/Code

GitHub link to the source code of the project: https://github.com/rishabhpatel9/MSc-Final-Project/blob/master/genderclassifier.ipynb

## 8.2 Appendix 2: Gender classifier for dataset 1

The code for gender classifier trained using dataset 1 is in the attached filed by the name “Appendix – code.docx”. The original code file is a part of the GitHub repository mentioned in Appendix 1 under the name gc-ds1.ipynb.

## 8.3 Appendix 3: Gender classifier for dataset 2 with normal model

The code for gender classifier trained using dataset 2 is in the attached filed by the name “Appendix – code.docx”. The original code file is a part of the GitHub repository mentioned in Appendix 1 under the name gc-ds2-smallermodel.ipynb.

## 8.4 Appendix 4: Gender classifier for dataset 2 with complex model and vgg16 model

The code for gender classifier trained using dataset 2 with an improved model is in the attached filed by the name “Appendix – code.docx”. The original code file is a part of the GitHub repository mentioned in Appendix 1 under the name gc-ds2-largemodel-and-vgg16-attempt-fail.ipynb.

## 8.5 Appendix 5: Gender classifier for dataset 3 Version 1 (male privileged set)

The code for gender classifier trained using dataset 3 version 1 is in the attached filed by the name “Appendix – code.docx”. The original code file is a part of the GitHub repository mentioned in Appendix 1 under the name gc-ds3-malepriv.ipynb.

## 8.6 Appendix 6: Gender classifier for dataset 3 Version 2 (female privileged set)

The code for gender classifier trained using dataset 3 version 2 is in the attached filed by the name “Appendix – code.docx”. The original code file is a part of the GitHub repository mentioned in Appendix 1 under the name gc-ds3-femalepriv.ipynb.