Detecting and Mitigating Bias in Machine Learning Image Data through Semantic Description of the Attention Mechanism

The use-case Gender Bias in Profession Prediction from Images

Georgios Dimitropoulos

TU Delft

Detecting and Mitigating Bias in Machine Learning Image Data

through Semantic Description of the Attention Mechanism
The use-case Gender Bias in Profession Prediction from Images

by

Georgios Dimitropoulos

to obtain the degree of Master of Science in Computer Science with Specialization in Data Science at the Delft University of Technology

to be defended publicly on Monday September 16, 2019 at 15:00 PM.

Student number:	4727657
Project duration:	November 1, 2018-September 16, 2019
Thesis committee:	
Chair:	Prof. Dr. G.-J. Houben, Faculty EEMCS, TU Delft
University Supervisor:	Prof. Dr. A. Bozzon, Faculty IDE, TU Delft
Company Supervisor	Dr. P. Pawełczak, Faculty EEMCS, TU Delft
Committee Member:	Dr. Z. Szlávik, Center for Advanced Studies IBM Benelux

An electronic version of this thesis is available at http://repository.tudelft.nl/.

PREFACE

Artificial Intelligence (AI) and one of its popular application, Machine learning are becoming an increasingly ubiquitous part of our daily lives. Machine Learning is a tool which is utilized in a plethora of learning methods in a variety of domains. For instance, it is used in health care (prediction of cancer, drug discovery), financial services (financial trading, prediction of stocks), retail (product recommendations, customer service) and also in a vast amount of other sectors and applications. Furthermore, a lot of Machine Learning models are increasingly used to assist or replace humans in a variety of decision-making domains like financial services, health care and criminal justice. For example, decisions about whether or not an unemployed person should receive some social welfare benefits or decision about determining whether a prior defendant who has been set up free from the jail is going to commit a crime again.

Therefore, algorithmic bias is gained a significant popularity over the last years as these decisions in the aforementioned life-affecting scenarios may have important impacts on the lives of people who are involved. A lot of concerns have been raised about the fairness of these decisions with respect to different groups of people and about possible harmful discrimination that may be apparent. Discrimination is more objectionable in cases that certain privileged groups have systematic advantages and in comparison some unprivileged groups have systematic disadvantages. For example, this kind of unfairness is more evident in cases like loan eligibility or risk of recidivism, where there exist some sensitive or protected features such as race or gender that lead to predictions which are biased towards groups that are represented in the training set in a disproportionate way.

In this thesis project, we are going to focus on coming up with a methodology to identify and mitigate gender bias in Machine Learning data through semantically describe the reason that a particular prediction of Machine Learning model is made. We focus on this specific angle of the problem, because we strongly believe that it is extremely important to tackle these aspects of the problem properly, as in a opposite case there is a high risk of harmful consequences and negative impacts towards the lives of the groups of people who receive biased, unfair and undesirable decisions. Therefore we seek an answer to the following research question: How to provide training data in Machine Learning algorithms that are balanced in terms of gender bias with respect to the content of the images to the output of these models?

We start by investigating which the current methods and their limitations related to bias in Machine Learning data with respect to protected attributes of people are. After that, we focus on ways of how to compensate for gender bias that is related with the content of the image data in Machine Learning systems. In order to do so, we focus on two angles. Firstly, to find to what extent the content of the image is correlated to the prediction errors with respect to the gender in a Machine Learning system. Secondly, to discover a way in using crowdsourcing to help uncover potential unknown elements of gender bias that may reside in Machine Learning data. Finally, we want to semantically describe at scale the reason that a particular prediction of a Machine Learning system is made in a human interpretable way. Although our work can be applied to a variety of visual tasks in which gender bias may be appear, we focus on the specific use case of the profession prediction from images.

Our results show that the metrics that we adopt enable to observe whether there is discrimination in the predictions of our classification model with respect to the gender. Also, through the experiments that we make, we are able to semantically describe at scale the reason that a particular prediction of a Machine Learning classification model is made in a human interpretable way. Finally, we also highlight, that the methodology that we propose, helps us towards compensating for gender bias that is related with the content of the image data.

Georgios Dimitropoulos
Delft, September 2019

ACKNOWLEDGEMENT

I would like to acknowledge all people and institutions that we worked together during this year. I strongly believe, that in case that I did not have such a support of many people, and without the Delft University of Technology and the Center for Advanced Studied of the IBM Benelux, the completion of this thesis would only be an imaginary situation.
Particularly, I would like to thank my thesis supervisors from the university Alessandro Bozzon and from the company Zoltán Szlávik for their precious and frequent feedback. They were always here for me to answer all the questions that I had about my research. More specifically, they helped me a lot in giving structure to my thoughts, in suggesting new ideas to me and in presenting in a better way my work.
I would also like to thank some other people who helped me towards completing this thesis project. Particularly, Benjamin Timmermans who provided some precious feedback of how to design successfully a crowdsourcing task and Agathe Balayn who was always here for me in order to listen and give feedback to my presentations. I also wish to acknowledge the other members of the committee, Geert-Jan Houben and Przemysław Pawełczak, without who I would not be able to defend my thesis.
Furthermore, working at the Center of Advanced Studies (CAS) of the IBM Benelux was undoubtedly a great opportunity to interact and discuss about our works with more people during the large interval of the thesis project. Therefore, I wish to acknowledge all the other interns from CAS and all the students from university who made the period during the thesis very funny and more motivating.
Finally, I should express my deep gratitude to my family for giving eternal courage, motivation, power and support to me throughout this year. This accomplishment would not have been real without them.

Thank you.

Georgios Dimitropoulos
Delft, Netherlands, September 2019

"Never say never, because limits, like fears, are often just an illusion."
— Michael Jordan

Contents

LIST OF FIGURES

List of Tables

1

INTRODUCTION

Artificial Intelligence (AI) and one of its popular application, Machine Learning are becoming an increasingly ubiquitous part of our daily lives. Machine Learning is a tool which is utilized in a plethora of learning methods. It is used in a variety of domains, like health care (prediction of cancer, drug discovery), financial services (financial trading, prediction of stocks), retail (product recommendations, customer service) and also in a vast amount of other sectors and applications. Furthermore, a lot of Machine Learning models are increasingly used to assist or replace humans in a variety of decision-making domains like financial services, health care and criminal justice. For instance, decisions about whether or not an unemployed person should receive some social welfare benefits or decision about determining whether a prior defendant who has been set up free from the jail is going to commit a crime again.

A lot of concerns have been raised about the bias of the decisions in life-affecting scenarios with respect to different groups of people and about possible harmful discrimination that may be apparent. Discrimination is more objectionable in cases that certain privileged groups have systematic advantages and in comparison some unprivileged groups have systematic disadvantages. For example, this kind of unfairness is more evident in cases like loan eligibility or risk of recidivism, where there exist some sensitive or protected features such as race or gender that lead to predictions which are biased towards groups that are represented in the training set in a disproportionate way.

As a motivating example, a survey was conducted investigating the risk of recidivism of a number of prior offenders based on their historical data since criminal history is a good predictor for future recidivism. It was found that black defendants were predicted to be at a higher risk of recidivism than they actually were compared to white defendants. More specifically, black defendants experienced two times higher false positive rate.

Another motivating example is the classification of some black people as gorillas based on image recognition algorithms. Also, in another case it has been found that the embedding model "word2vec" contains some gender bias. More specifically, it was found that some occupations related to male people were captain, doctor and boss and the ones that related to female occupations were homemaker, nurse and housekeeper.

Hence, in these cases evaluation metrics like accuracy have only minor meaning. It is more important to know whether our model has random or systematic errors. For these reasons, our main motivation, is to come up with ways to identify, reason upon and mitigate bias in Machine Learning data. Being able to so, constitutes a significant contribution towards the fairness and interpretation of these models in order to be used successfully and without discrimination in decision-making domains like financial services, health care and criminal justice.

Most of these Machine Learning models are built from human-generated data and in that way human biases result in a skewed and sometimes unbalanced distribution in the training data. More specifically, in case that this data concern groups of people, there might be a high probability that a specific subset of these people to be represented in a disproportionate way. This situation has as a direct consequence that algorithms trained with this kind of data encoding human bias, to reproduce and not eliminate this bias. Thus, in such a case, this unintended bias is propagated from the training data to the resulting models and finally leading to unfair applications towards certain groups. Also, it might be the case (as we will show later on), that imposing an equal distribution of the data with respect to the protected attribute of people, does not always solve the problem. Hence, it is crucial to be able to investigate the data, the algorithm and the output of Machine

Learning models which are made for predicting decisions that are highly related to different groups of people in order to identify potential issues that are related to bias towards them. A general overview of the whole **bias-mitigation classifier-build pipeline** and the **focus of our work** can be depicted in Figure 1.1. As it can be depicted in this Figure, we focus on the data aspect of this pipeline.

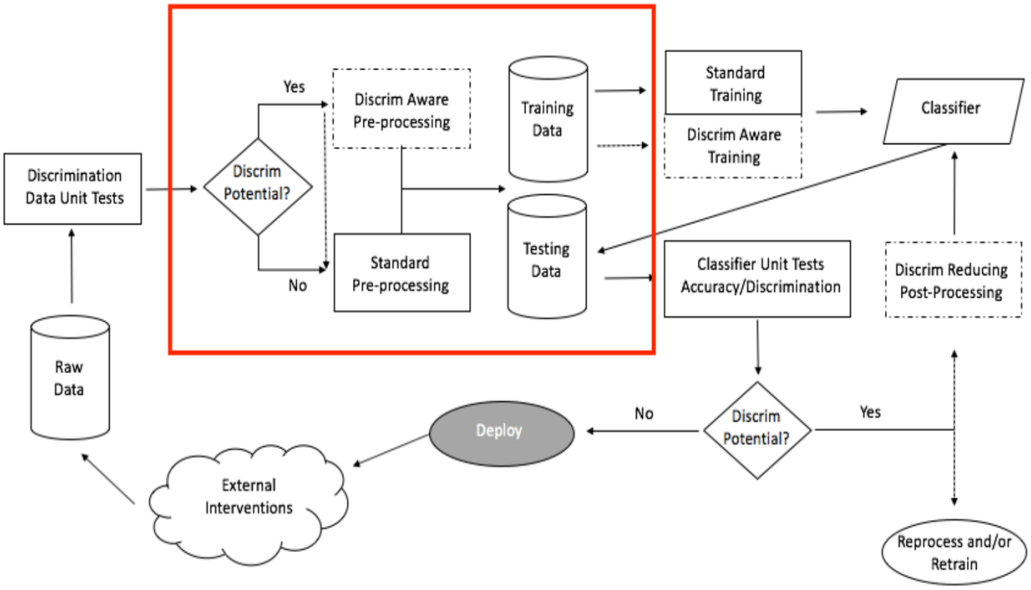


Figure 1.1: Image taken from Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification-A general overview of the whole **bias-mitigation classifier-build pipeline** and the **focus of our work**

In this thesis project, we are going to focus on coming up with a methodology to identify, explain and mitigate gender bias in Machine Learning data for the profession prediction from images use case. We focus on this specific angle of the problem, because following the notion that exists in Machine Learning community, the unintended bias in the data is propagated from the training data to the resulting models and finally leads to unfair applications towards certain groups. Thus a model or an application is as unbiased as the data that is fed on them. Therefore, we strongly believe that it is extremely important to tackle these aspects of the problem properly, as in a opposite case there is a high risk of harmful consequences and negative impacts towards the lives of the groups of people who receive biased, unfair and undesirable decisions.

In the rest of this chapter, we firstly identify the problem that we are going to tackle in the thesis project. After that, we mention the research motivation of the project and the reasons that we strongly believe that is vital to conduct more research in this direction. In the sequel, we transform the context and the motivation into a well defined set of research questions followed by an one-to-one mapping of the contributions that we are going to perform answering these research questions. We conclude this chapter describing a roadmap (outline) of the way that the thesis is going to be organized.

1.1. PROBLEM STATEMENT OF THE THESIS

In this work we focus on two main problems. The first one is that there might be discrimination between different groups of people with respect to their protected attributes in the aforementioned Machine Learning decision making applications. The second one is that there is lack of methods that actually interpret and explain the predictions of these Machine Learning systems which are then used to help decision making.

In our work we are focusing on a specific aspect of these decision making applications, namely in Machine Learning training data for the problem of classification. As we stated in 1, most of the Machine Learning models that are used in these decision making applications are built from human-generated data. Based on this, human biases result in a skewed and sometimes unbalanced distribution in the training data. The solution that is proposed in the literature in order to build an unbiased system is to impose an equal distribution of the training data with respect to the protected attribute of people. However,as we will show later on, this is not always true. We show that, it is not the distribution of the labels with respect to the protected attribute, but the presence of specific visual clues that leads to that bias.

Therefore, our ultimate goal is to propose a methodology that helps in detecting, reasoning upon and compensating for bias with respect to protected attributes of people in predictions of Machine Learning systems due to issues in the training data. Particularly, we adopt the use case of profession prediction from images and we focus on a specific form of data (images) and on specific protected attribute (gender) and we impose an equal distribution of the labels with respect to the gender in the training data. Intuitively, as a first step we want to observe whether there is discrimination in the predictions of a Machine Learning classification model with respect to the gender through inspecting the miss-classified data.

As a second step, we describe in a semantically rich fashion the features in the data that are likely to be related to a particular biased prediction of a Machine Learning system. In order to do so we use first an attention mechanism that gives as an output the parts of the image that influence the prediction. The goal of this step is to provide to people insights about the gender bias in the predictions of the Machine Learning classification model. However, a technical challenge that arises here is that this result is actually in a form of blobs of pixels. Thus, this form of output makes it really difficult or even impossible to provide to people interpretable explanations at scale and in an automatic way of which features in an image matter and lead towards a particular prediction in a Machine Learning classification model.

Hence, we need to find a way to end up firstly with a set of classes of pre-defined objects, helping in attaching a semantic label on top of the attention mechanism. In order to do so we propose two solutions, namely: using object detection and using crowdsourcing. The intuition behind the first solution (object detection) is to observe whether the semantic description can be attained automatically. The goal behind the second solution (crowdsourcing) is two-fold: Firstly, to understand how the intuition of people about a potential cause of gender bias actually compares with the actual reason that affects the prediction of a Machine Learning classification model. Secondly, to gain an insight of how much the intuition of people about elements of gender bias actually matches the semantic description coming from the object detection approach.

Finally, as a last step we want to propose a way to compensate for gender bias that is related with the content of the image data. Particularly, we investigate whether obfuscating the visual clues coming from the overlapping of the attention mechanism-object detection and attention mechanism-crowdsourcing improves the predictions of the classification model with respect to the gender.

1.2. RESEARCH MOTIVATION, HYPOTHESIS AND USE-CASE OF THE THESIS

To the best of our knowledge, most research tackles solely individual aspects like trying to detect and mitigate bias and does not pay any attention to explain their reasoning in a human interpretable way. Also, the main method that they use is to balance the distribution of the training data with respect to the protected attribute. However, as we show, this is not always the solution to the problem. On the contrary, we decided to study concurrently three steps (detection, semantic interpretation and mitigation of bias) to overcome these shortcoming and limitations and we perform an extensive evaluation of our method in order to verify its efficiency and effectiveness.

Therefore, we believe that our methodology for identification, reasoning upon and mitigation of gender bias of Machine Learning data, can be used into an end-to-end fashion in a Machine Learning pipeline scheme in the data-algorithmic/model fairness research community in a way that the final decisions of these systems are unbiased across all the involved people, interpretable and there are no negative impacts on human lives. Particularly, our ultimate goal is to provide to people a tool that allows them to inspect the training data at a semantic level towards understanding which particular features in the data are mostly related to the classification outcome and introduce some systematic bias on that. The reason that we focus specifically on image data is that the nature of this kind of data facilitates the inspection of their content and their analysis in a semantic level.

The main hypothesis we test is the following:

> **H:** The **presence** of **SPECIFIC visual clues** in image data that **give away** the protected attribute (e.g. **gender**), **affect** the **classification** outcome and **introduce bias** on that.

We could study different prediction tasks, as long as they involve visual bias with respect to protected attributes of people. We chose the use-case of profession prediction from images for the following reasons:

- Profession prediction from images has great application potentials in intelligent services and systems. For instance, feed of the news, products and friends requests could be dynamically suggested to users

in Social Media in an effective way by recommendation systems in case that their professions can be predicted in an automatic way.

- No study has previously been performed to study the correlation between profession prediction and gender bias that may be appear in that classification task.
- Finally, indication of unknown elements of gender bias that may reside in that kind of image data is a subjective property. Thus, something like that enables us to employ the power of the diversity that the crowdsourcing can offer.

Therefore, detecting and mitigating gender bias in Machine Learning data for the the task of profession prediction from images is a promising and interesting use case to apply our experiments.

1.3. RESEARCH QUESTIONS OF THE THESIS

Driven by our research motivation, in this thesis project we focus on studying ways of identifying, reasoning upon and mitigating gender bias in Machine Learning image training data through semantic description of attention mechanism for the specific task of profession prediction from images. Thus, our focus in this thesis project is to address the following main research question:

MAIN RESEARCH QUESTION

MRQ: How to **analyze, reason upon and fix** the **content** of Machine Learning **image training data** in order to **correct** and **reduce gender bias** in the output of the subsequent trained models?

In order to answer this main research question we split it to the following research sub-questions that need to be answered first. Also, the methods which we are going to develop in order to answer each of the research sub-questions are provided in an explicit way.

RESEARCH SUB-QUESTIONS

RSQ1: Which are the **current methods** and their **limitations** related to **bias in Machine Learning data** with respect to **protected attributes** of people?

This question aims at finding the state-of-the-art methods and their drawbacks that are related to bias with respect to protected attributes of people in Machine Learning data. In order to be able to do so, we are going to perform an in depth literature review. The main topics of the literature review that are going to be covered are definitions of bias with respect to protected attributes of people, evaluation measures of bias, bias mitigation algorithms and bias identification techniques using crowdsourcing.

RSQ2: How can we **describe** in a **semantically rich fashion at scale** the **features** in the **data** that are likely to be **related** to a particular **biased prediction** of a **Machine Learning system**?

This question aims at enabling us to make a methodology in order to find way to describe in a semantically rich fashion at scale the features in the data that are likely to be related to a particular biased prediction of a Machine Learning system in a human interpretable way.

RSQ3: How can we **compensate** for **gender bias** that is related with the **content** of the **image data** in **Machine Learning** systems?

In order to answer this question, we have to split it up into two parts, namely:

RSQ3a: How much is the **content** of the **image correlated** to the **prediction errors** with respect to the **gender**?

This question aims at enabling us to make a conclusion of whether there is a correlation of the content of an image with the prediction errors of a Machine Learning model with respect to the gender.

RSQ3b: How can we use **crowdsourcing** to help **uncover** potential **unknown elements** of **gender bias** that may reside in Machine Learning **data**?

This question aims at enabling us to make a conclusion of whether we can use crowdsourcing in an efficient way of helping us uncovering potential unknown elements of gender bias that may reside in Machine Learning data.

1.4. CONTRIBUTIONS OF THE THESIS

In this thesis project we propose a **methodology** that aims in finding ways of **detecting** and **mitigating bias** that may exist in **Machine Learning data** through **semantic description** of the **attention mechanism** for the use-case of **gender bias in profession prediction** from images. Our **methodology** consists of **three** main **steps**: Firstly, we **detect gender bias** in the data. Secondly, we use two different approaches (**object detection** and **crowdsourcing**) in order to obtain a **semantic description** of the features of the image data that matter in the classification in a way that is direct **interpretable** by people. Finally, we **mitigate** the **gender bias** through **obfuscating-blurring** the aforementioned parts of the image data. Therefore, we bring the following three main contributions:

CO1: The first contribution of the thesis is an **in-depth systematic literature review** of state-of-the-art methods and their limitations related to bias with respect to protected attributes of people in Machine Learning. We investigate existing literature on the topics of definitions and evaluation measures of bias, bias mitigation algorithms and bias identification using the crowd which enable us to highlight current limitations as well as to come up with possible directions to develop our methodology. It enables to answer the first research sub-question **(RSQ1)**.

CO2: The second contribution is the answer to the second research sub-question (**RSQ2)**. More specifically, it is the **one part** of the **methodology** that we propose of **describing** in a **semantically rich fashion at scale** the **features** in the **data** that are likely to be **related** to a particular **biased prediction** of a **Machine Learning system** in a human interpretable way.

CO3: The third contribution is the **other part** of the **methodology** that we propose and is the answer to our third research sub-question (**RSQ3)**. Particularly, of **compensating** for **gender bias** that is related with the **content** of the **image data** in **Machine Learning** systems.

1.5. OUTLINE OF THE THESIS

The thesis project is organized as follows. As a first step **(Chapter 2)** we conduct a **literature review** of the different fields concerned with our main research and sub-research questions. More specifically, we focus on (1) the different definitions of bias in Machine Learning with respect to protected attributes of people, (2) different evaluation methods to measure bias in Machine Learning, (3) several bias mitigation algorithms in Machine Learning and finally (4) ways that we can use crowdsourcing in order to identify bias in the data. The goal of this part of the thesis project is to review the state-of-the-art methods and their limitations (**RSQ1**) related to bias with respect to protected attributes of people in Machine Learning and to also to helps us answering a **part** of the second (**RSQ2**) and third research-sub questions (**RSQ3**). This actually corresponds to the first contribution (**CO1**) that we make.

In the sequel **(Chapter 3)**, we present the **use case** (profession prediction from images), **dataset** and the **classification task** to study detection, semantic interpretation and mitigation of gender bias. Then we tackle the second and third research sub-questions (**RSQ2**)+(**RSQ3**), where we describe the **methodology** that we made in order to answer these research sub-questions and corresponds to the second and third contribution (**CO2**)+(**CO3**) that we make **(Chapter 4)**. More specifically, we show the bias identification, semantic interpretation and mitigation steps that we follow. In the next chapter, we give the **evaluation** and the results of the experiments that we did in order to verify the effectiveness of our schemes **(Chapter 5)**. Finally, we discuss the overall results, **conclusions** and give suggestions for future work **(Chapter 6)**.

A roadmap of the Thesis project can be depicted in Figure 2.1.

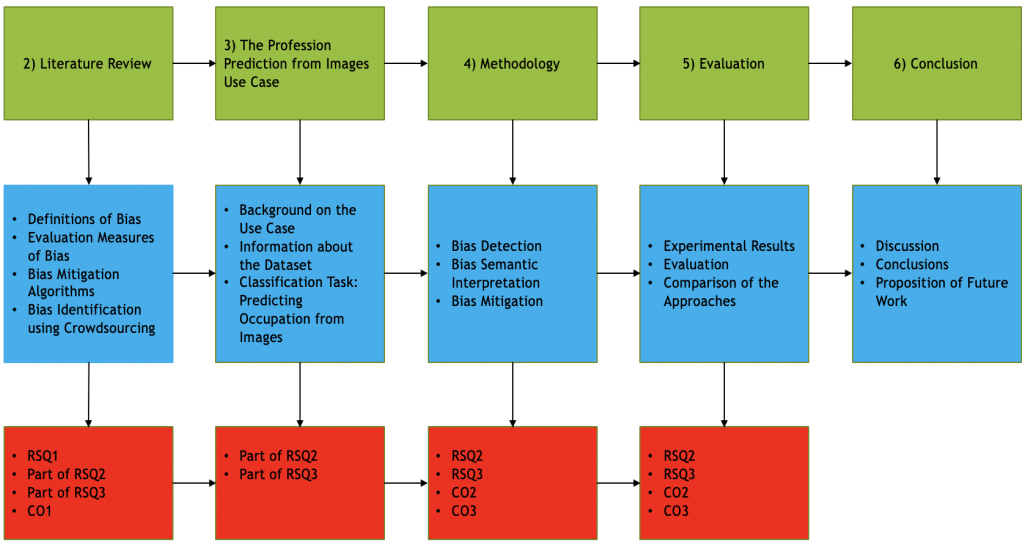


Figure 1.2: Roadmap of the Thesis Project

2

LITERATURE REVIEW

In order to be able to tackle our research questions, we are going to start by conducting an in-depth systematic literature review in the related bibliography in order to investigate the current state-of-the-art methods and their limitations that are related to bias in Machine Learning with respect to protected attributes of people. The specific parts that we are going to focus are the different definitions of bias with respect to protected attributes of people in Machine learning, the evaluations methods which are employed to measure this bias, the bias mitigation algorithms and finally approaches that use crowdsourcing as an identification technique of bias in data. This systematic literature review enables us to answer the first research sub-question **(RSQ1)** and a **part** of the second **(RSQ2)** and third research sub-questions **(RSQ3)** that we pose and actually corresponds to the first contribution **(CO1)** that we make.

Firstly, it is very important to introduce all the necessary **definitions of bias in Machine learning** with respect to **protected attributes of people (RSQ1)**. To be able to do, we perform it as a two step procedure. As a first step, we define and distinguish the different protected or sensitive attributes that exist and vary according to different contexts. As a second step, we quote all the different definitions of bias in Machine learning with respect to the protected attributes of people and give an intuitive explanation of each of them, followed by some relationships that exist between them. Thus, the goal of this part is 2-fold: firstly, to verify that our choice of gender as protected attribute is indeed valid and secondly to give a thorough and formal view of the different definitions of bias in Machine learning with respect to the protected attributes of people that exist in order to be able to choose the one that best suit in our case.

Secondly, we are interested in the **evaluation methods** that exist for the identification and **measuring of bias in Machine learning (RSQ1)**, **(RSQ2)** and **(RSQ3)**. To achieve this we are focusing on 3 layers. Firstly, we distinguish between individual and group bias with respect to protected attributes of people. After that, we bring our attention in methods and evaluation metrics that are used in **measuring group bias** in the **data**. Finally, we review the related work which is about **measuring** the **group bias** in the Machine learning **models**. Therefore the aim of this part of the literature review is to end up with the evaluation metrics which best suit to our case. More specifically, based on the related bibliography which we focus on this part, we want to end up with the dataset-bias metrics which we want to use in order to identify and calculate the gender bias on the data.

After that, we study **bias mitigation algorithms** that exist in the related bibliography and are used in order to **mitigate unwanted bias of Machine learning schemes (RSQ1)**. In order to have a clearer insight, we provide a taxonomy of these algorithms and we split them into 3 main categories (Pre-Processing, In-Processing and Post-Processing algorithms). In the first category (Pre-Processing) the algorithms that are belong to, are **applied** to the **data**. In the second category (In-Processing), the algorithms that are belong to, are applied to the corresponding **Machine learning model**. Finally, in the third category (Post-Processing) we have algorithms which are applied to the **predicted labels**. The aim of all these three categories of algorithms is to mitigate the bias of corresponding Machine learning schemes towards certain groups of people that received unfair decisions. Hence, the purpose of this part is to review the bias mitigation algorithms that are used to mitigate the bias of Machine learning schemes and to position our approach (third step of our methodology, bias mitigation step) with respect to the state of the art bias mitigation algorithms and we end up with the conclusion that there is no a single perfect bias mitigation algorithm and the choice of them is closely related to the use case and the domain of interest.

Finally, we pay our attention on related bibliography which is dealing with manners that use **crowdsourcing techniques** to **identify bias in** the **data (RSQ3)**. More specifically, we investigate two types of approaches. Firstly, methods that **explore bias** and stereotypes in the **data using the crowd**. Secondly, approaches which employ the crowd to **understand the way that humans** actually **perceive** the **bias** and **fairness** of some **attributes**. The aim of this part of the literature review is to acquire an understanding of how crowdsourcing can be used as a bias identification technique in the data.

A roadmap of the building blocks of our literature review can be depicted in Figure 2.1.

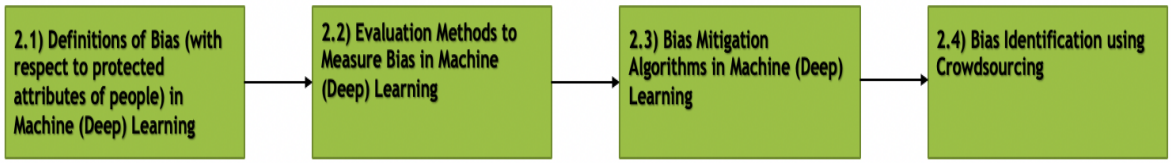


Figure 2.1: Roadmap of the building blocks of the literature review

2.1. DEFINITIONS OF BIAS IN MACHINE LEARNING WITH RESPECT TO PROTECTED ATTRIBUTES OF PEOPLE

In this part of the literature review we are going to introduce all the different definitions of protected attributes and of bias in Machine learning with respect to those attributes. It is very important to do so, in order to verify our choice of gender as protected feature, to give a thorough view of the different definitions of bias with respect to those attributes that exist and also to end up with the definition that we are going to choose in our work.

2.1.1. METHODOLOGY TO SEARCH FOR PAPERS

In order to search for different definitions of protected attributes related to context and uses cases and different definitions of bias in Machine Learning with respect to these attributes, we performed a search on Google Scholar, Scopus and ACM using the combination of the following queries and keywords: "Protected attributes of people", "Sensitive attributes of people", "Algorithmic Fairness", "Algorithmic Bias". After that we added as keywords the queries "Bias in Machine Learning", "Fairness in Machine Learning" and we selected all the papers as well as the references of the retained papers that were related to these terms. Finally we combined these queries in the form of "Definitions of fairness and bias in Machine Learning with respect to protected attributes of people" something that enabled us to find more specific papers related to our research questions. As last step we removed duplicate papers. It is worth mentioning that the vast majority of the papers that we found here were from the FAT and FAT/ML [1] conferences.

2.1.2. DIFFERENT DEFINITIONS OF PROTECTED ATTRIBUTES RELATED TO CONTEXT AND USE CASES

The Equality Act[2] is a legislation which unifies 116 separate pieces of legislation into one single Act. The purpose of this legislation is to provide a legal framework towards protecting the rights of people in achieving equal opportunities for all and promoting a fair and equal society. According to this, there are nine protected/sensitive characteristics or attributes or features of people. Namely, these are the following ones: **age**, **disability**, **gender reassignment**, **marriage or civil partnership** (only in employment) , **pregnancy and maternity**, **race**, **religion or belief**, **gender/sex** and **sexual orientation**. Consequently, in a high level, we have

[1] http://www.fatml.org
[2] https://www.legislation.gov.uk/ukpga/2010/15/contents

that in case that two individuals have exactly the same qualifications, then they should receive exactly the same treatment independently of the "values" which they have in their protected features.

Therefore, based on this regulation our choice to use gender as protected attribute is clearly verified. In the sequel of this part of the literature review we are going to examine some related bibliography on bias in Machine learning with respect to protected attributes of people to have a clearer view of how these features are used with respect to the specific context at hand and to give an intuition of some example use-cases that these issues are arisen.

In [2], [3], [4], [5], [6], [7] [8] and [9], the use case that they examine is about the COMPAS[3] algorithm in the Broward County Florida dataset[4] for the problem of assigning defendants risk scores between 1 and 10 which indicate how likely they are to commit a violent crime again while awaiting their trial. Hence, the domain here is the criminal justice and more specifically they focus on algorithms for pretrial release decisions. The protected attribute here is the race of the defendants (white and black). In [10]and [11], they focus on analyzing the ways in which people belonging to different sex or different race may experience advertising and commercial on the Web in a different manner. Therefore, in this case the protected attributes would be the gender and the race.

As another example, in [12] , they focus on the task of FICO[5] scores which are widely used in the United States in order to predict credit worthiness. The protected attribute that they have in this case is the race and they have four different categories (Asian, Black, White and Hispanic). In [13] and [14], they examine three cases, namely: the German credit dataset[6] where they classify bank account holders into credit class Good or Bad and the sensitive feature is age (similar to [15], [16], [17], [18] and [19]), the Adult income dataset[7] whether they classify the income of being larger or smaller than 50K dollars, and the protected attribute is Gender (similar to [4], [5], [16], [19], [20], [21], [22] and [23]) and the final one is a health dataset derived from the Heritage Health Prize[8] whether they classify people into two categories (whether they are going to spend any days in the hospital that year or not) and the sensitive attribute is age.

In [16] they also examined as another use case the Ricci dataset[9] that was used in order to classify firefighters into those ones who are promoted and not promoted based on the score that they had in an exam and the protected attribute was the race (white and non-white people). Moreover, in [9], [24] and [25] they employ as a use case the Stop, Question, and Frisk dataset[10] where they predict whether people on the street have or not weapons in their possession. The protected attribute that they consider is race (black, black Hispanic, white, white Hispanic and Other).

In [5] they also experiment with the Heart Dataset[11] where their purpose is to predict whether or not an individual has a heart condition and the sensitive attribute is age (middle-aged adults and seniors). Furthermore in [25], they also examine the prediction of individuals' success in law school based on some exams tests data[12] and the protected attributes are race and gender. Finally, in [26] they experiment with the North Carolina traffic stops dataset[13] where they investigate for discrimination in police searches of motorists who were stopped in North Carolina and the protected attribute is race (black, Hispanic, Asian, white).

2.1.3. DIFFERENT DEFINITIONS OF BIAS IN MACHINE LEARNING WITH RESPECT TO PROTECTED ATTRIBUTES OF PEOPLE

Before delving into the different mathematical definitions of bias in Machine Learning with respect to protected attributes of people, we think it makes sense to have a look firstly to the definition of bias and fairness in general according to the dictionary's definition. Therefore, according to Cambridge dictionary bias [14] is defined as: "the action of supporting or opposing a particular person or thing in an unfair way, because of allowing personal opinions to influence your judgment" and fairness [15] is defined as: "the quality of treating

[3] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[4] https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis
[5] https://www.federalreserve.gov/boarddocs/rptcongress/creditscore/creditscore.pdf
[6] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
[7] https://archive.ics.uci.edu/ml/datasets/adult
[8] https://www.kaggle.com/c/hhp/data
[9] https://supreme.justia.com/cases/federal/us/557/557/
[10] https://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page
[11] https://archive.ics.uci.edu/ml/datasets/heart+Disease
[12] https://eric.ed.gov/?id=ED469370
[13] https://opendatapolicing.com/nc/
[14] https://dictionary.cambridge.org/dictionary/english/bias
[15] https://dictionary.cambridge.org/dictionary/english/fairness

people equally or in a way that is right or reasonable" and according to Thesaurus bias is defined as: [16] "a particular tendency, trend, inclination, feeling, or opinion, especially one that is preconceived or unreasoned" and fairness [17] as: "Impartial and just treatment or behaviour without favouritism or discrimination". Thus, we can infer from these definitions that mitigation of bias or fairness has to do with the the state or condition of treating people equally in a way free of injustice.

Based on that, we are now ready to present the different definitions of bias as well as possible relationships between them. According to [18] there are more than twenty different definitions of bias or fairness in the related bibliography. However, not all of these definitions can be applied in every case. Thus, to have a better understanding, it is crucial to know the rationale behind each of them. One important point that we should mention is that one specific situation could considered to be unbiased or fair according to one definition but all the way around according to one other definition. However, this is something that from a mathematical point of view makes sense, as for some of them it is impossible to be satisfied concurrently as proposed in [2], [3] and [27]. Adopting the notation that is used mostly in the literature, the notations which we are going to use in this chapter can be depicted in Table 2.1.

Notation	Meaning
G	Protected or sensitive attribute
X	All additional attributes
Y	Actual label
S	Predicted probability for a certain label
d	Predicted label

Table 2.1: Notation and corresponding meaning of symbols used in definition of bias in Machine learning

Following the procedure described in [18] we split the different definitions of bias or fairness into 3 main categories, namely: **Statistical, Similarity-based** and **Causal-Reasoning** measures. A taxonomy of bias or fairness definitions in Machine Learning can be depicted in Figure 2.2.

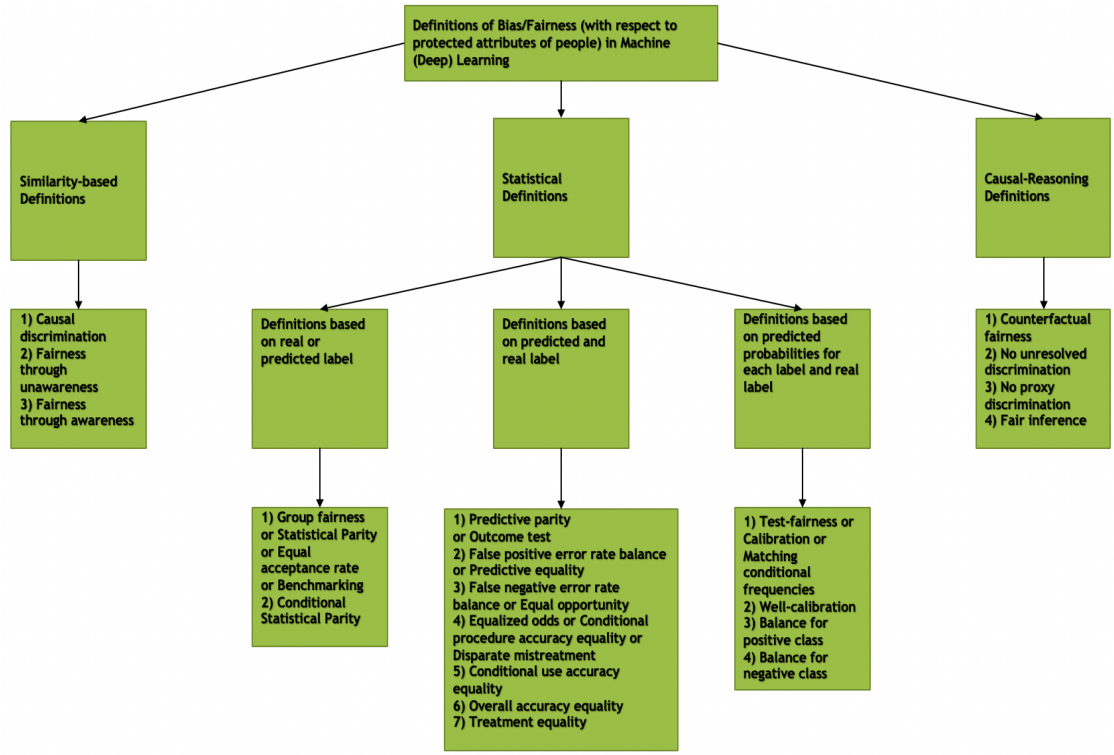


Figure 2.2: A taxonomy of different definitions of bias/fairness in Machine Learning

[16] https://www.dictionary.com/browse/bias?s=t
[17] https://www.dictionary.com/browse/fairness

Statistical measures

The first and most simple category of definitions of bias that we are going to examine is the one which is consisted of the statistical measures as is forms the basis for the other two categories (Similarity-based and Causal-Reasoning measures). Based on the fact that most of these statistical measures of bias strongly rely on some statistical metrics, we believe that makes sense to quote them for easier reference and clearer understanding. For the sake of simplicity we assume that we only have 2 classes (binary-classification) and we set one of these two classes to be the positive class (i.e. people of this class are eligible for a social welfare benefit) and the other one to be the negative one (i.e. people of this class are non-eligible for a social welfare benefit). Thus, we have the following preliminaries twelve **statistical metrics**:

1) True positive (TP): In case that both the predicted and actual label belong to the positive class.
2) False positive (FP): In case that the predicted label belongs to the positive class. and the real one belongs to the negative class.
3) False negative (FN): In case that the predicted label belongs to the negative class and the actual one belongs to the positive class.
4) True negative (TN): In case that both the predicted and real label belong to the negative class.
5) Precision or **Positive predictive value (PPV)**: It is the fraction of positive instances that were predicted correctly to belong to the positive class divided by the total number of instances that were predicted to belong to the positive class, namely: $\frac{TP}{TP+FP}$.
6) False discovery rate (FDR): It is the fraction of negative instances that were predicted incorrectly to belong to the positive class divided by the total number of instances that were predicted to belong to the positive class, namely: $\frac{FP}{TP+FP}$.
7) False omission rate (FOR): It is the fraction of the positive instances that were predicted incorrectly to belong to the negative class divided by the total number of instances that were predicted to belong to the negative class, namely: $\frac{FN}{TN+FN}$.
8) Negative predictive value (NPV): It is the fraction of negative instances that were predicted correctly to belong to the negative class divided by the total number of instances that were predicted to belong to the negative class, namely: $\frac{TN}{TN+FN}$.
9) Recall or **Sensitivity** or **True positive rate (TPR)**: It is the fraction of positive instances that were predicted correctly to belong to the positive class divided by the total number of instances that actually belong to the positive class, namely: $\frac{TP}{TP+FN}$.
10) False positive rate (FPR): It is the fraction of negative instances that were predicted incorrectly to belong to the positive class divided by the total number of instances that actually belong to the negative class, namely: $\frac{FP}{FP+TN}$.
11. False negative rate (FNR): It is the fraction of positive instances that were predicted incorrectly to belong to the negative class divided by the total number of instances that actually belong to the positive class, namely: $\frac{FN}{TP+FN}$.
12) True negative rate (TNR): It is the fraction of negative instances that were predicted correctly to belong to the negative class divided by the total number of instances that actually belong to the negative class, namely: $\frac{TN}{TN+FP}$.

Based on the aforementioned definitions of these statistical metrics, we are now ready to list a taxonomy of the **statistical definitions of bias**. In total there are thirteen different statistical definition of bias. These definitions can also be split into three sub-categories with respect to the quantities which are based on. More specifically these three sub-categories of the statistical definition are: **Definitions based on**: **Predicted label**, **Predicted and real label** and **Predicted probabilities for each label and real label**. We now present the corresponding statistical definitions of bias that belong to each of these sub-categories.

A) Definitions based on predicted or real label

In this sub-category we have two statistical definitions of bias which are based on the predicted label d for a variety of demographic distributions of people. However, we should mention that based on the fact that are simple enough, they have important limitations which are concerned by more advanced definitions which are going to be examined later.

1) Group fairness or **statistical parity** or **equal acceptance rate** or **benchmarking**

According to [22], [26] and [28], this definition is satisfied in case that individuals in both protected and

unprotected groups have equal probability to belong to the positive predicted class, namely we have that $P(d = 1|G = g_1) = P(d = 1|G = g_2)$, where g_1, g_2 represent the protected and unprotected groups respectively (i.e. men and females). Therefore, the rationale behind this definition is that the classification of people to the positive class should be independent of the group (protected or not) that belong to.

2)Conditional statistical parity

According to [2], the difference of this definition and the previous one is that the later extends the former through augmenting the set of features that are going to affect the predicted label by some other attributes which are called "legitimate attributes" (i.e. employment). This definition is said to be full-filled in case that individuals in both protected and unprotected groups have equal probability to belong to the positive predicted class controlling for a limited set of "legitimate" attributes L, namely we have that $P(d = 1|L = l, G = g_1) = P(d = 1|L = l, G = g_2)$, where l corresponds to the values of these legitimate attributes (i.e. number of years of employment).

B) Definitions based on predicted and real label

In this sub-category we have seven statistical definitions of bias which are based on both the predicted label and real label for a variety of demographic distributions of people.

1)Predictive parity or **outcome test**

According to [3] and [26], this definition is satisfied in case that both protected and unprotected groups have equal precision or PPV, namely $P(Y = 1|d = 1, G = g_1) = P(Y = 1|d = 1, G = g_2)$. This directly implies also that both protected and unprotected groups have equal FDR, namely $P(Y = 0|d = 1, G = g_1) = P(Y = 0|d = 1, G = g_2)$. The rationale behind this definition is that the fraction of the correct positive predictions has to be the same for both protected and unprotected groups.

2)False positive error rate balance or **predictive equality**

According to [2] and [3] , this definition is fulfilled in case that both protected and unprotected groups have equal FPR, namely $P(d = 1|Y = 0, G = g_1) = P(d = 1|Y = 0, G = g_2)$. This is also equivalent to $P(d = 0|Y = 0, G = g_1) = P(d = 0|Y = 0, G = g_2)$ which means that both the protected and unprotected have equal TNR. The idea here is that people in both protected and unprotected groups who actually belong to the negative class should receive similar results.

3)False negative error rate balance or **equal opportunity**.

According to [3], [12] and [25], this definition is satisfied in case that both protected and unprotected groups have equal FNR, namely $P(d = 0|Y = 1, G = g_1) = P(d = 0|Y = 1, G = g_2)$. This is equivalent to $P(d = 1|Y = 1, G = g_1) = P(d = 1|Y = 1, G = g_2)$, which means that both protected and unprotected groups have also equal TPR. The rationale here is that people in both protected and unprotected groups who actually belong to positive class should receive similar results.

4)Equalized odds or **conditional procedure accuracy equality** or **disparate mistreatment**.

This definition is a combination of the previous two definitions. According to [6], [12] and [29], it is satisfied in case that protected and unprotected groups have equal TPR and equal FPR. Thus, the mathematical relationship here is the intersection of the aforementioned two formulas, namely $P(d = 1|Y = c, G = g_1) = P(d = 1|Y = i, G = g_2)$, where $c = 0$ or $c = 1$. Intuitively, this definition implies that people who actually belong to the positive class and those ones who actually belong to the negative class must receive similar classification treatment, regardless of the group (protected or unprotected) that belong to. We have to mention here that according to [3], in case that the condition $P(Y = 1|G = g_1) \neq P(Y = 1|G = g_2)$ holds, then if the predictive parity/outcome test definition is satisfied, then this definition is impossible to be satisfied concurrently.

5)Conditional use accuracy equality

In a similar fashion with the former definition, this one also conjuncts two conditions: precision/PPV and NPV, namely $(P(Y = 1|d = 1, G = g_1) = P(Y = 1|d = 1, G = g_2)) \wedge (P(Y = 0|d = 0, G = g_1) = P(Y = 0|d = 0, G = g_2))$. The idea here [29] is that the probability of people who are predicted to belong to the positive class and actually belong to that and the probability of people who are predicted to belong to the negative class and actually belong to that should be equal for both protected and unprotected groups.

6)Overall accuracy equality

According to [29], in order for this definition to be satisfied, both protected and unprotected groups must have equal prediction accuracy (i.e. the probability of people belong to either positive or negative class to be assigned to its respective class), namely $P(d = Y, G = g_1) = P(d = Y, G = g_2)$. This means that, the TN and TP are of equal importance. The intuition here is that the probability of people who actually belong to the positive class to be correctly assigned to that class and the probability of people who actually belong to

the negative class to be correctly assigned to that class should be the same for both groups (protected and unprotected).

7)Treatment equality

According to [29], this definition takes into account the ratio of errors that a classifier has done and not its accuracy. Therefore, this definition is satisfied in case that both protected and unprotected groups have an equal ratio of FN and FP, namely $\frac{FN}{FP} \times g_1 = \frac{FN}{FP} \times g_2$.

C) Definitions based on predicted probabilities for each label and real label

In this sub-category we have four statistical definitions of bias which consider the predicted probabilities for each label and the real label.

1)Test-fairness or **calibration** or **matching conditional frequencies**

According to [3] and [12], this definition is satisfied in case that people in both protected and unprotected groups have equal probability to actually belong to the positive class for any predicted probability score S, namely $P(Y = 1|S = s, G = g_1) = P(Y = 1|S = s, G = g_2)$, where $s \in [0, 1]$. It is worth mentioning here that this definition is similar to predictive parity/outcome test. The only difference is that now in this definition the fraction of the correct positive predictions are considered $\forall s \in [0, 1]$.

2)Well-calibration

According to [27], this definition is an extension of the former definition in the sense that people in both protected and unprotected groups not only should have equal probability to actually belong to the positive class for any predicted probability score S but also this probability must equals S, namely $P(Y = 1|S = s, G = g_1) = P(Y = 1|S = s, G = g_2) = s$, where s is the predicted probability score. The idea here is that in case that a set of people have a probability s of belonging to the positive class, then s percent of those should actually belong to the positive class.

3)Balance for positive class

According to [27], this definition is fulfilled in case that people within the actual positive class from both protected and unprotected groups have an equal average predicted probability score S, namely $E(S|Y = 1, G = g_1) = E(S|Y = 1, G = g_2)$.

4)Balance for negative class

According to [27], this is a flipped version of the former definition. More specifically, this definition is satisfied in case that people within the actual negative class from both protected and unprotected groups have an equal average predicted probability score S, namely $E(S|Y = 0, G = g_1) = E(S|Y = 0, G = g_2)$.

Similarity-based measures

The main limitation that statistical definitions of bias possess is that they take into account only the sensitive attribute G and mainly ignore all the other attributes of the people who are classified. However, an approach like this conceals bias: For instance, there is a probability that the percentage of people who classified in the positive class to be the same for both the protected and unprotected groups, however the way that these people (protected and unprotected group) were chosen might be not the same according to the value of another attribute (not the sensitive one). Therefore, according to [30] for example, with respect to the definition of predictive parity/outcome test, the classifier is going to be unbiased, but there would be a discrimination in the way that people were chosen based on the group (protected or not) that belong to. Thus, the main intuition in the similarity-based measures is that they do not alienate the insensitive attributes of people. In total, there are three different similarity-based definitions of bias.

1)Causal discrimination

According to [30], this definition is satisfied in case that we have the same classification result for any two people with the same insensitive attributes X. Thus, two people (one from the protected and the other one from the unprotected group) with the same features X, are going to be classified both in the positive or both in the negative class, namely $(X_{g_1} = X_{g_2} \wedge G_{g_1} != G_{g_2}) \rightarrow d_{g_1} = d_{g_2}$, where X_{g_1} and X_{g_2} refer to the insensitive attributes, G_{g_1} and G_{g_2} refer to the sensitive attributes, and d_{g_1} and d_{g_1} refer to the classification outcome of people of protected and unprotected groups respectively.

2)Fairness through unawareness

According to [25], this definition is satisfied in case that the sensitive features are not explicitly used in the process of the classification result. Therefore, we do not use the protected attributes in the training pro-

cess of the classifier and based on that the classification outcome is independent from them and should be identical for people with the same features X from both the protected and unprotected group, namely $X_{g_1} = X_{g_2} \rightarrow d_{g_1} = d_{g_2}$.

3) Fairness through awareness

According to [28], this definition is a general and combinatory version of the aforementioned two definitions. The main intuition behind this is that similar people must have a similar classification outcome. In mathematical terms this similarity is captured through a distance metric. Thus, in order for this definition to be satisfied, the distance between the distributions of the classification outcomes for people should be at most equal to the the distance between these people. Namely, it should be that $D(M(x), M(y)) \leq k(x, y)$, where V a set of people, $k : V \times V \rightarrow R$ a distance metric between people, $M : V \rightarrow \delta A$ a mapping from a set of people to the probability distributions over the classification outcomes and D a distance metric between the distribution of the classification outcomes.

Causal reasoning measures

The final sub-category of the different definitions of bias is the the causal reasoning measures where these definitions assume a given causal graph. More specifically, it is a directed, acyclic graph (DAG), where each node represents a feature of people and edges represent the relationships between the different features. The main intuition here is that, the relationships among features and their consequence on the classification result are captured by a set of structural equations that are used as a next step to estimate the effect of the protected features through securing a tolerable level of discrepancy due to these protected features. In these graphs, we need also to introduce two new categories of attributes, proxy and resolving. An attribute is called proxy if its value can be used in order to calculate the value of another attribute. On the other hand, an attribute is called resolving in case that is influenced by a protected attribute in a non-discrepant way. In total, there are four different causal reasoning definitions of bias.

1) Counterfactual fairness

According to [25], a causal graph is counter-factually fair in case that the classification outcome d in the graph is independent from the descendants of the protected attribute G.

2) No unresolved discrimination

According to the [31], a causal graph exhibits no unresolved discrimination in case that the only path that exists from the protected attribute G to the predicted label d is through a resolving attribute.

3) No proxy discrimination

According to [31], a causal graph exhibits no proxy discrimination in case that no path exists from the protected attribute G to the predicted label d via a proxy attribute.

4) Fair inference

According to [4], in this definition the paths of a causal graph are classified into two categories, namely: legitimate or illegitimate. Therefore, a causal graph is defined as fair inferential in case that they do exist illegitimate paths from the sensitive attribute G to the predicted label d.

2.1.4. DISCUSSION

In this section, we introduced all the protected attributes according to the legislation, some use cases that employ these attributes and all the definitions of bias in Machine learning with respect to these protected features.

(RSQ1): The goal of this part was 2-fold: firstly, we verified that our choice of gender as protected attribute was indeed valid based on the legislation and we also presented some use cases that can be found in the related literature and secondly we gave a thorough and formal view of the different definitions of bias in Machine learning with respect to the protected attributes of people that exist, accompanied by the intuition behind them in order to be able to choose those ones that best suit in our case.

More specifically, we observed that there is no agreed framework in order for a classifier to be considered always as fair or not. It clearly depends on the notion of bias at the specific task and definition which we have at hand. Also, as it is stated in [18], more research is needed in order to clarify and be sure which definition could be more appropriate to a specific task. The general agreed-framework and the comment that we can also make is that although the statistical definitions are easy enough to be measured, they are in many cases insufficient [3], [27], [28] and [29]. The main cause of this problem is that the majority of the statistical metrics require a lot of available data which has a known true label. However this is not always the case. On the other side of the spectrum, similarity-based and causal reasoning metrics assume that there is the availability

of some experts (i.e. to determine the distance measure). Therefore, not only there is a difficulty of these definitions to be applied in practice (i.e. lack of experts), but also in case that even there are available experts, these measures could be eventually biased towards the opinions of the experts.

To conclude, the notion of the bias and the definition that one should adopt, clearly depends on the task at hand. Based on the aforementioned reasoning, we are going to employ the statistical measures in order to quantify the bias in the dataset as they are the ones that have been used mostly in the literature in similar cases, they are more intuitive and they do not require the use of experts or the existence of a causal graph. More specifically, we are going to use as a metric the **statistical parity** (data layer).

2.2. EVALUATION METHODS TO MEASURE BIAS IN MACHINE LEARNING

In this part of the literature review, we are going to focus on evaluation methods that are used for measuring bias in Machine learning. More specifically, we are going to perform a taxonomy between metrics and classify them into three main categories, namely: individual and group bias metrics, data metrics and model metrics. Our goal here is to present this clear taxonomy of all metrics that can be used to measure bias in Machine learning and to end up with the group evaluation metrics on data aspect which best suit to our case.

2.2.1. METHODOLOGY TO SEARCH FOR PAPERS

In order to search for different evaluation methods to measure bias in Machine learning, we performed a search on Google Scholar, Scopus and ACM using the combination of the following queries and keywords: "Algorithmic (Un)Fairness", "Algorithmic Bias", "Measuring Bias", "Measuring (Un)Fairness". After that we added as keywords the queries: "Bias in Machine Learning", "Fairness in Machine Learning", "Evaluation metrics of learning algorithms", "Metrics for Un(Fairness) in Machine learning", "Metrics for Bias in Machine learning", "Group fairness in Machine learning", "Individual fairness in Machine learning", "Data fairness/bias in Machine learning data", "Model fairness/bias in Machine learning" and we selected all the papers as well as the references of the retained papers that were related to these terms. Finally we combined these queries in the form of "Identify/Measure fairness and bias in Machine Learning data", "Measure/Identify fairness and bias in Machine Learning models", "Metrics for fairness and bias in Machine learning data and model" something that enabled us to find more specific papers related to our research questions. As a last step we removed duplicate papers. It is worth mentioning that the vast majority of the papers that we found here were again from the FAT and FAT/ML [18] conferences.

2.2.2. INDIVIDUAL VS GROUP BIAS

The first taxonomy that we should perform is that of individual and group bias. As proposed in [32], two basic frameworks have been arisen in bibliography on algorithmic discrimination in Machine learning. The first one is **individual bias/fairness**, which requires that similar individuals should be treat in a similar way independently of the values that they "possess" in their protected attributes [28]. On the other hand, **group bias/fairness** firstly partitions a population of people into groups (protected and unprotected groups) which are differentiated by the values that they "possess" in their protected attributes and after that aims for an equal treatment between these two groups [33] and [34]. Following the taxonomy that is presented here [19], in case that the task at hand is about individual bias/fairness, then the metrics that must be employed are [20]: **Euclidean distance**, **Mahalanobis distance**, **Manhattan distance**, **Mean Euclidean distance difference**, **Mean Euclidean distance ratio**, **Mean Mahalanobis distance difference**, **Mean Mahalanobis distance ratio**, **Mean Manhattan distance difference**, **Mean Manhattan distance ratio** and **Consistency** [13].

On the other hand, in case that we are interested in quantifying group bias/fairness then the measures that we should use are the following ones [21]: **Base rate**, **Disparate impact**, **Statistical parity difference** or **Mean difference**, **Number of negatives conditioned on protected attributes**, **Number of positives conditioned on protected attributes**, **Average of absolute difference in** FPR **and** TPR, **Average of difference in** FPR **and** TPR, **Number of** TP, FP, TN, FN **conditioned on protected attributes**, **Equal of opportunity difference** or **True positive difference rate**, **Accuracy**, **Error rate**, **Difference in error rates**, **Ratio of error rates**, **False discovery rate**, **Difference in false discovery rate**, **Ratio of false discovery rate**, **False negative rate**, **Difference in false negative rate**, **Ratio of false negative rate**, **False omission rate**, **Difference in false omission**

[18] http://www.fatml.org
[19] http://aif360.mybluemix.net/resources#guidance
[20] https://aif360.readthedocs.io/en/latest/modules/metrics.html#sample-distortion-metric
[21] https://aif360.readthedocs.io/en/latest/modules/metrics.html#binary-label-dataset-metric

rate, **Ratio of false omission rate**, **False positive rate**, **Difference in false positive rate**, **Ratio of false positive rate**, **Generalized false negative rate**, **Generalized false positive rate**, **Generalized true negative rate**, **Generalized true positive rate**, **Negative predictive value**, **Generalized number of** FN, **Generalized number of** FP, **Generalized number of** TN, **Generalized number of** TP, **Number of predicted negatives**, **Number of predicted positives**, PPV or **Precision**, **Selection rate**, TPR or **Recall** or **Sensitivity** and TNR or **Specificity**.

Finally, in case that the task we are dealing with is concerned with both individual and group bias/fairness and we have as a requirement to use one single measure, then we have to employ the following [22] metrics: **Between-group generalized entropy index** [35], **Between all groups generalized entropy index**, **Between-group coefficient of variation**, **Between all groups coefficient of variation**, **Between-group Theil index** , **Between all groups Theil index**, **Coefficient of variation**, **Generalized entropy index** [35] and **Theil index**. However, we could also examine simultaneously various individual or group bias/fairness metrics in case that we are not obliged to use only one single measure. The taxonomy of the individual, group and combined bias/fairness measures can be depicted in Figure 2.3.

Figure 2.3: A taxonomy of evaluation measures (**individual**, **group** and **combined**) of **bias/fairness** in Machine Learning

Since we are focusing on group bias/fairness in this work, our next step would be to split up the group bias/fairness metrics into two categories, namely: **Dataset metrics** and **Model metrics**.

2.2.3. MEASURING GROUP BIAS/FAIRNESS IN DATA

In this section, we are going to examine the group bias/fairness metrics that are used in order to quantify unwanted bias in the data. The metrics that lie in this category are a subset of the group bias/fairness metrics which we examined in the previous section. More specifically, according to [23] we have the following dataset metrics: **Base rate**, **Disparate impact**, **Statistical parity difference** or **Mean difference**, **Number of negatives conditioned on protected attributes**, **Number of positives conditioned on protected attributes**. In the sequel we will focus on measures that quantify bias in the model.

[22] https://aif360.readthedocs.io/en/latest/modules/metrics.html#classification-metric
[23] https://aif360.readthedocs.io/en/latest/modules/metrics.html#dataset-metric

2.2.4. MEASURING GROUP BIAS/FAIRNESS IN MACHINE LEARNING MODEL

In is part of the literature review, we will investigate group bias/fairness metrics that are used in order to quantify unwanted bias in the Machine learning model. The metrics that lie in this category again are a subset of the group bias/fairness metrics which we examined previously section. More specifically, according to [24] we have the following model metrics: **Statistical parity difference** or **Mean Difference**, **Average of absolute difference in** FPR **and** TPR, **Average of difference in** FPR **and** TPR, **Number of** TP, FP, TN, FN **conditioned on protected attributes**, **Equal of opportunity difference** or **True positive difference rate**, **Accuracy**, **Error rate**, **Difference in error rates**, **Ratio of error rates**, **False discovery rate**, **Difference in false discovery rate**, **Ratio of false discovery rate**, **False negative rate**, **Difference in false negative rate**, **Ratio of false negative rate**, **False omission rate**, **Difference in false omission rate**, **Ratio of false omission rate**, **False positive rate**, **Difference in false positive rate**, **Ratio of false positive rate**, **Generalized false negative rate**, **Generalized false positive rate**, **Generalized true negative rate**, **Generalized true positive rate**, **Negative predictive value**, **Generalized number of** FN, **Generalized number of** FP, **Generalized number of** TN, **Generalized number of** TP, **Number of predicted negatives**, **Number of predicted positives**, PPV or **Precision**, **Selection rate**, TPR or **Recall** or **Sensitivity** and TNR or **Specificity**. A taxonomy of the group bias/fairness measures for data and model evaluation can be delineated in Figure 2.4.

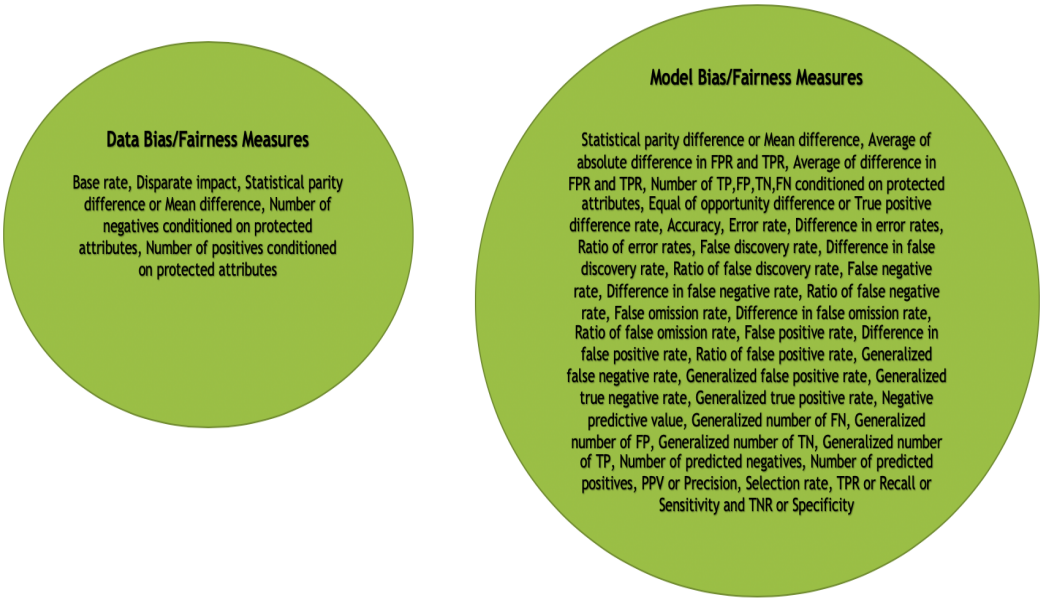


Figure 2.4: A taxonomy of evaluation measures (**data** and **model**) **bias/fairness** in Machine Learning

2.2.5. DISCUSSION

In this section, we introduced all the evaluation methods and metrics that are used for measuring bias/unfairness in Machine learning. More specifically, we performed a taxonomy between these metrics and classified them between individual and group bias/fairness metrics and data and model metrics.

(RSQ1), **(RSQ2)** and **(RSQ3)**: Related to our first sub-research question (and also establishing evaluation measures for our second and third sub-research questions), we were able in this part to identify all the different methods and evaluation metrics that are used in order to quantify bias in Machine learning. More specifically, we gave firstly a clear taxonomy and definitions of the different metrics that exist in literature in evaluating bias in individual or group layer. Based on the fact that we are focusing on this work on (group) gender bias, we examined as a next step the group metrics. In the sequel, we provided a taxonomy of these metrics with respect to the point that are used to measure bias in a Machine learning pipeline and more specifically we differentiated them into two categories data and model metrics.

Hence, the goal of this part was 2-fold: firstly, we presented the aforementioned taxonomies of the evaluation metrics of bias/fairness, which we believe is crucial in order to use them in a correct manner and to the

[24] https://aif360.readthedocs.io/en/latest/modules/metrics.html#classification-metric

right point in a Machine learning pipeline. Secondly, based on these taxonomies and explanations of each metric to end up with the group evaluation metrics on data aspect which best suit to our case.

Similarly to our discussion in the previous section, there is no unique metric that is used in order to quantify group bias/unfairness in data or model level. Thus, the metrics that one should adopt clearly depend on the task at hand and on the way that wants to present their findings. To conclude, based on the aforementioned reasoning, we are going to employ group bias/fairness data metrics in order to quantify the bias in the dataset and more specifically, we are going to examine metrics like **statistical parity difference** as that is the one that is usually used in similar cases in literature and has the less limitations.

2.3. BIAS MITIGATION ALGORITHMS IN MACHINE LEARNING

In this part of the literature review, we study **bias mitigation algorithms** that exist in the related bibliography and are used in order to **mitigate bias of Machine learning schemes**. More specifically, we are going to perform a taxonomy between them into three main categories (**Pre-Processing**, **In-Processing** and **Post-Processing algorithms**) with respect to the point (**data**, **model** or **predicted labels**) of the Machine learning pipeline that they are used. Hence, the goal of this part is to present a clear taxonomy and explanation of the bias mitigation algorithms that are used to mitigate the bias of Machine learning schemes, in order to be able to acquire a clear insight and understanding of the advantages and limitations of each of them. It is worth mentioning here that our approach (third step of our methodology, bias mitigation step) belongs to the first category of these algorithms (pre-processing algorithms) as it is used in the data. Moreover, a clear difference of our approach is that it does not manipulate the distribution of the data (e.g. oversampling, undersampling, re-weighting etc.), as such a treatment does not provide an unbiased outcome in our use case, but it changes the "nature" of the data itself through manipulating its content. A taxonomy of the bias mitigation algorithms (**Pre-Processing**, **In-Processing** and **Post-Processing algorithms**) can be seen in Figure 2.5.

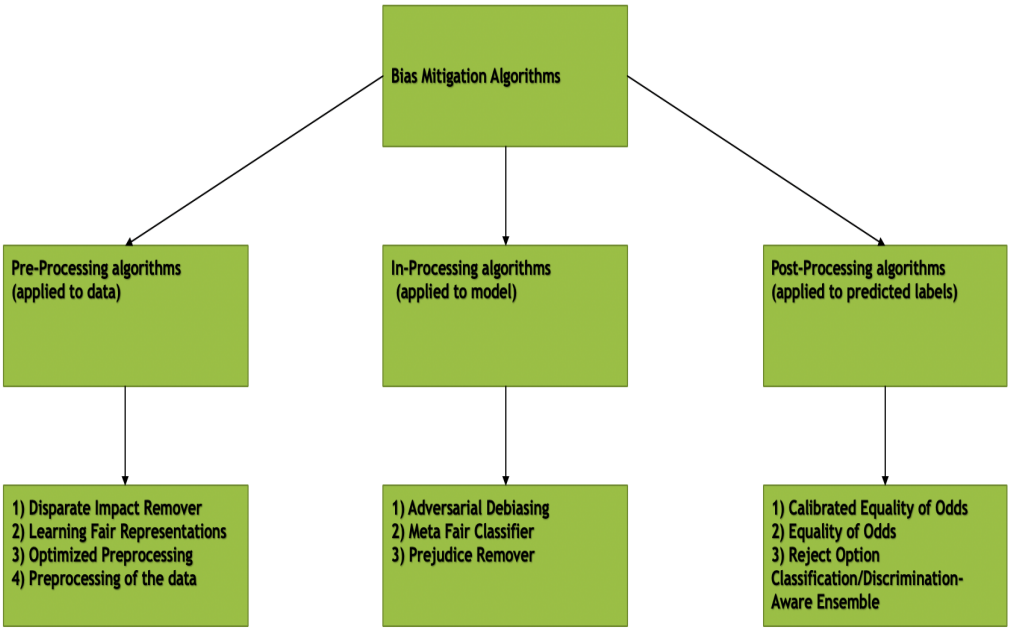


Figure 2.5: A taxonomy of the bias mitigation algorithms (**Pre-Processing**, **In-Processing** and **Post-Processing algorithms**)

A general overview of the whole **bias-mitigation classifier-build pipeline** can be depicted in Figure 2.6 .

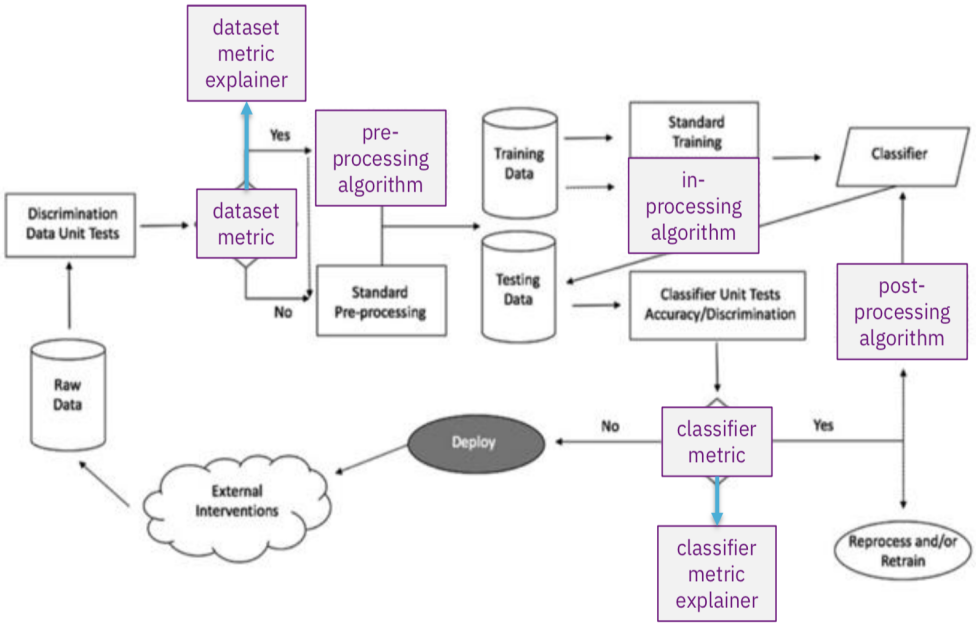


Figure 2.6: Image taken from Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification-A general overview of the whole **bias-mitigation classifier-build pipeline**

2.3.1. METHODOLOGY TO SEARCH FOR PAPERS

In order to search for different bias mitigation algorithms that are used to mitigate the bias in Machine learning, we performed a search on Google Scholar, Scopus and ACM using the combination of the following queries and keywords: "Algorithmic (Un)Fairness mitigation", "Algorithmic Bias mitigation", "Algorithmic (Un)Fairness compensation", "Algorithmic bias compensation". After that we added as keywords the queries: "Mitigate/Compensate Bias in Machine Learning", "Mitigate/Compensate unfairness in Machine Learning", "Bias mitigation algorithms in Machine learning", "Unfairness mitigation algorithms in Machine learning", "Bias mitigation algorithms for Machine learning data", "Bias mitigation algorithms for Machine learning models", "Bias mitigation algorithms for Machine learning predicted labels" and we selected all the papers as well as the references of the retained papers that were related to these terms. Finally we combined these queries in the form of "Bias mitigation algorithms to increase fairness in Machine learning", "Bias mitigation algorithms to eliminate bias in Machine learning" something that enabled us to find more specific papers related to our research questions. As a last step we removed duplicate papers.

2.3.2. BIAS MITIGATION ALGORITHMS APPLIED TO DATA

Firstly, we are going to present the first category of the bias mitigation algorithms, namely Pre-Processing algorithms that are applied to the data. In this category there are four algorithms which are used for this purpose, namely: **Disparate Impact Remover**, **Learning Fair Representations**, **Optimized Preprocessing** and **Reweighing**. Now we are going to present and analyze each of them.

Disparate Impact Remover

According to [16], they are focusing on ways that try to mitigate bias by modifying the data. The notion of bias that they adopt is that of **disparate impact** that encodes the unintentional bias which takes place in case that the classification outcome is heavily different for different groups. More specifically, they are working towards three main contributions. Firstly, they link the 80 percent classification rule of disparate impact to the balanced error rate (BER) and they try to minimize it. After that, they present a way of transforming the input dataset in order not to be able to predict the protected attribute. Their main idea here is that they change only the non-protecting attributes and preserving the protecting ones and the binary class that is going to predicted same as in the original dataset. They are doing so, in order to preserve the performance of the classification scheme in high levels in terms of evaluation measures (i.e. accuracy) in a way that they try to combat the fairness-accuracy trade-off that appears here.

Finally, they present their results on the experiments that they conducted and they compare them to similar works [13], [21] and [36]. Based on that, the conclusion that they draw is that they were able to simultaneously mask the bias that appears in the data and to preserve the information that they need in the data. However, the main limitation of their work is that they experienced with a tremendous difference in the performance of the different classification algorithms that they used. Also, the repair procedure that they suggest operates only on numerical and not on categorical attributes.

Learning Fair Representations

In [13], they propose a learning algorithm that is used for bias mitigation and fair classification in both individual and group level. More specifically, as far as the group bias/fairness is concerned, they want the number of people in a protected group who are classified into the positive class to be the same with that number of the people in the unprotected group. In the individual layer, they aim in achieving the purpose of treating similar individuals in a similar manner. Their notion of bias/fairness is that of an optimization problem in a way that they seek of finding a proper representation of the data while satisfying simultaneously two opposite goals, namely: encode the data as well as possible (preserving personal attributes of people as much as possible) and at the same time muddle part of them (deleting any personal information related to membership of people in the protected group).

Therefore, the main idea here is to map each individual from the initial input space to some probability distribution into a new (latent) representation space. The purpose of doing so is to blind all the information that can be used in order to identify the group (protected or not) that someone belongs and preserving all the other information. Finally, they present their results on the experiments section and they compare them with similar works [36], [21]. They were able to achieve better performance in terms of minimizing discrimination, maximizing the difference between accuracy and discrimination and maximizing individual bias/fairness.

Optimized Preprocessing

In [1], they present a probabilistic formulation of the dataset in order to reduce discrimination. More specifically, they introduce a convex optimization to learn a data transformation satisfying three main purposes, namely: controlling discrimination, limiting distortion in individuals and preserving utility. A pipeline of their methodology for predictive learning with discrimination prevention can be depicted in Figure 2.7.

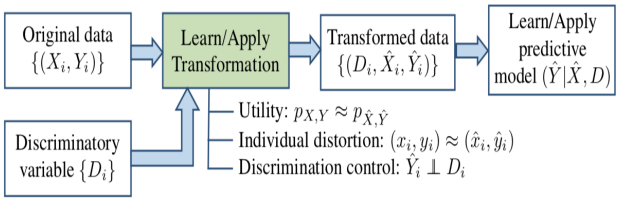


Figure 2.7: Image taken from [1]-Pipeline for predictive learning with discrimination prevention

In comparison to the work of [13] which is also considers discrimination control (or group bias/fairness), individual bias/fairness and utility, these criteria here are more visible and are declared in a more direct way. More specifically, discrimination control is defined in terms of the intermediate features in [13], where in [1] is defined in terms of the classification outcomes. Also, in the former one, classification outcomes are not taken into account by individual distortion and utility is classifier-specific.

Finally, they present their results through experimenting with two datasets, namely in the Broward County Florida dataset[25] and on Adult income[26] dataset. They were able to verify that not only all these three criteria which they proposed can be satisfied concurrently but also that the data transformation that they employed, achieved reducing the recidivism risk for the unprotected group of African-American and increasing the income for the the unprotected group of female respectively. The main limitation of their proposed framework is that lacks on theoretical characterizations.

Preprocessing of the data

In [37], they are also dealing with this Discrimination-aware classification problem, where they try to find an optimal trade-off between accuracy and non-discrimination. Again in this work they focus on ways that they can pre-process the data in order to eradicate any potential discrimination issues before building a classifier. Their concentration is on four main techniques, namely: suppression of the protected feature, changing the

[25] https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis
[26] https://archive.ics.uci.edu/ml/datasets/adult

class labels (massaging), reweighing and re-sampling of the initial data.

More specifically, using suppression (their baseline), they identify the features that have the most correlation with the protected feature and they remove both these most correlated features and the protected attribute. In massaging, they change the labels of some points on the initial dataset and in order to select these points they use a ranker (points closer to the decision border are firstly chosen to be relabeled) following the method in [36]. Since this approach seems a little bit intrusive they also use reweighing, where they assign different weights in the data points as proposed in [38]. As a last method, they are experimenting with re-sampling techniques such as oversampling and undersampling as it is not always possible to incorporate weights in the learning process. In this case, they try two different techniques, namely: Uniform Sampling (every data point has equal probability to be duplicated or removed) and Preferential Sampling, (borderline points (points which are difficult to be classified and are chosen again using a ranker) are more likely to be duplicated or removed).

Finally they present the result of their experiments on three different datasets, namely: the Census Income dataset[27], the Communities and Crimes dataset[28] in which the problem is to predict the total number of violent crimes per 100K population and the protected attribute is the race and the Dutch census dataset[29] of the year 2001 where the problem is to predict the occupation (prestigious or not) and the protected attribute is the gender. Based on that experiments they drawn the following conclusions: Removing only the protected feature from the dataset is not enough to mitigate discrimination issues (redlining effect [36]) and massaging, reweighing and preferential sampling were able to achieve a very good trade off between accuracy and non-discrimination. However, last but not least, the main limitation of their approach is that it cannot be used in dealing with cases that multiple protected attributes with multiple values exist.

2.3.3. BIAS MITIGATION ALGORITHMS APPLIED TO MACHINE LEARNING MODEL

In the sequel we are going to present the second category of the bias mitigation algorithms, namely In-Processing algorithms that are applied to the Machine learning model. In this category there are three algorithms which are used for this purpose, namely: **Adversarial Debiasing**, **Meta Fair Classifier** and **Prejudice Remover**. Now we are going to present and analyze each of them.

Adversarial Debiasing

In [23], they propose a framework that can be used in order to mitigate demographical (i.e. gender or race) bias through using a specific kind of Deep learning, adversarial learning and more specifically Generative Adversarial Networks (GANs) [39]. The main idea here is that the input to the network (they are experimenting with the Census Adult data[30] and with word embeddings [40]) , produces a prediction (whether the income exceeds 50K or completion of an analogy) and the adversary models a protected variable. Therefore their goal is to maximize the ability of making a correct classification outcome and concurrently to minimize the ability of making a correct prediction for the protected variable.

The strong elements of their methods are their ability to have accurate enough predictions that do not suffer of bias issues in the word embeddings case in comparison to the work of [40] and [41]. Moreover, they were able in achieving similar results in terms of the trade-off between accuracy-bias/fairness (equality of odds in this particular case) in the Census Adult data[31]. Last but not least, their method is compatible with a variety of notions of bias/fairness (i.e. Demographic Parity, Equality of Odds and Equality of Opportunity) and they do provide theoretical guarantees of their methodology.

The only limitation of their approach, which they also stress, is that its capability may be questionable in more complex tasks and the fact that a small change in the values of the hyperparameters will cause a sufficient enough divergence in its performance. Thus, in that sense, there might be overfitting issues related to the specific data that they tuned these hyperparameters.

Meta Fair Classifier

In [42], they propose a novel meta-algorithm for the task of classification which takes as input a variety of bias/fairness constraints with respect to multiple protected features. The main idea of their work is that they were able to invent a meta-algorithm which can be used as a solution to multiple classification problems that obey convex constraints and after that to convert bias/fairness-related problems to this category of

[27] https://archive.ics.uci.edu/ml/datasets/adult
[28] http://archive.ics.uci.edu/ml/datasets/communities+and+crime
[29] http://microdata.worldbank.org/index.php/catalog/2102/study-description
[30] https://archive.ics.uci.edu/ml/datasets/adult
[31] https://archive.ics.uci.edu/ml/datasets/adult

tasks. Thus, similar to the aforementioned directions of research they are dealing with the trade-off between bias/fairness and accuracy.

The main contribution of their work was the fact that their approach was compatible with over of ten different notions of bias/fairness (which far excels any previous approach to do so). Also, they were able to provide strong theoretical guarantees of their findings aiming in answering the problems proposed in [19]. Finally, through their experimental results were able to achieve very good results in bias/fairness with respect to a variety of different bias/fairness measures while achieving only a minor decrease in accuracy in three use cases, namely in: Census Adult[32], German credit[33] and COMPAS[34] datasets. The only limitation of their approach is that they only compare their results with three other approaches [6], and [12] [19] and there is almost a balance with respect to the performance. Therefore, they could also have include more comparisons with other similar works.

Prejudice Remover

In [43], they make two main contributions. Firstly, they provide a taxonomy of the causes of bias/unfairness in Machine learning, namely: prejudice, which in turn consists of three categories (direct, indirect and latent prejudice) underestimation and negative legacy. More specifically, they define and quantify prejudice as the dependence between protected and not protected attributes, which in turn may be direct, indirect or latent. Also, they define underestimation as the situation in which a classifier produces bias/unfair outcomes in comparison to those ones in the sample distribution, as it has not converged yet and negative legacy as the problem of bias/unfair distribution or labeling in the initial data.

In the sequel, they are focusing on the problem of indirect prejudice and for this reason they propose a regularization framework, namely a prejudice remover regularizer that confines the discriminative behavior of the classifier through ensuring independence from the protected attribute. The intuition here and in agreement with related bibliography is that they try to control the trade-off between accuracy and bias/fairness and their approach to achieve this is through tuning a regularization parameter. The main advantage of their method is that this framework is compatible with any predictive algorithm.

Finally, they compare their method with [15] in order to verify its capabilities in terms of evaluation measures through experimentation on the Census Income dataset[35]. The main limitations of their work were that their performance was inferior in comparison to [15] and also the fact that they did not perform any other comparisons with similar works or in other datasets. Last but not least, another shortcome as they stress is that their method is trapped to suboptimal solutions (local minima) due to the fact that their objective function is non-convex.

2.3.4. BIAS MITIGATION ALGORITHMS APPLIED TO PREDICTED LABELS

Finally we are going to present the third and last-one category of the bias mitigation algorithms, namely Post-Processing algorithms that are applied to the predicted labels of the Machine learning model. In this category there are three algorithms which are used for this purpose, namely: **Calibrated Equality of Odds**, **Equality of Odds** and **Reject Option Classification**. Now we are going to present and analyze each of them.

Calibrated Equality of Odds

In [5], they are dealing with the relationship between minimizing the error disparity among different groups (protected and unprotected) and maintaining the calibrated probability estimates. In agreement with related bibliography [3] and [27], they state that it is impossible to concurrently achieve calibration and satisfy the conditions of Equalized Odds. Nonetheless, they try to investigate possible relaxations of the conditions of the Equalized Odds in order to preserve calibration and end up with a sub-optimal solution in three use cases that they are experimenting with, namely: Adult Census[36], COMPAS[37] and Heart[38] datasets.

Therefore, they propose a post-processing algorithm that is applied to the predicted labels of the classifier in order to achieve the aforementioned sub-optimal solution. The main idea of their algorithm is that they keep aside predictive features for a random percentage of the initial data in order to simultaneously satisfy

[32] https://archive.ics.uci.edu/ml/datasets/adult
[33] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
[34] https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis
[35] https://archive.ics.uci.edu/ml/datasets/adult
[36] https://archive.ics.uci.edu/ml/datasets/adult
[37] https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis
[38] https://archive.ics.uci.edu/ml/datasets/heart+Disease

the conditions of the Equalized Odds and to achieve calibration.

The main limitation of their framework is that although this approach of withholding predictive features seems to be unbiased on average, eventually would be highly problematic for the unprivileged group in the sensitive cases of COMPAS[39] and heart[40] datasets. Finally, they found through their experimentation that indeed calibration and the conditions of Equalized Odds are in the vast majority of cases in-feasible to be satisfied at the same time and the best solution would be to examine each of them separately.

Equality of Odds

In [12], they propose a whole framework which can be used in order to mitigate bias. Firstly, they propose a bias/fairness measure which is used for helping against discrimination with respect to protected and unprotected groups. After that, they build a bias mitigation algorithm which is used as post-processing step to the predicted labels in order to derive an unbiased classifier that eliminates the aforementioned discrimination between the protected and unprotected groups based on the bias/fairness measure that they propose.

In the sequel they provide their experiments in the FICO[41] scores dataset through a variety of notions of bias/fairness. The main advantage of their work is the fact that they were able to provide a whole framework (define a bias/fairness measure, mitigate bias and compare it with other notions of bias/fairness). However, they neither provided experimental comparison with other similar works, nor they experimented with other datasets in order to prove the efficiency and the effectiveness of their approach.

Reject Option Classification/Discrimination-Aware Ensemble

In [44], they propose two post-processing bias-mitigation algorithms which employ the reject option that is available for probabilistic classifiers and the disagreement region of ensemble schemes of classifiers respectively. More specifically, the reject option classification (ROC) algorithm can be seen as a cost-based classification scheme, where the miss-classification cost of individuals belonging to the unprotected group is much higher in comparison to that of them belonging to the protected group. On the other hand, the Discrimination-Aware Ensemble algorithm that they propose, uses the disagreement region of ensemble schemes of classifiers to relabel individuals belonging to protected and unprotected groups in order to mitigate bias through reducing discrimination in a manner similar to [45].

Their method has three main advantages, namely: The fact that it can handle multiple protected features at the same time in comparison to [37] and [45] which are restricted to handle only one a protected feature at a time. Also, their approach is independent of the classifier that is employed in comparison to [15] and [46] which can only be used with particular classifiers (decision trees and Naive Bayes). Finally, they compare the results of their experiments with [15], [37] and [46] in two datasets (Adult Census[42] and Communities and Crimes[43]), where they show that their approach outperforms them in terms of the accuracy-bias/fairness trade-off.

2.3.5. DISCUSSION

In this section, we investigated all the **bias mitigation algorithms** that exist in the related literature for the purpose of **mitigating the bias of Machine learning schemes**. More specifically, we performed a taxonomy between these bias mitigation algorithms and classified them into three categories (**Pre-Processing**, **In-Processing** and **Post-Processing algorithms**) with respect to the position (**data**, **model** or **predicted labels**) of the Machine learning pipeline that are used. Furthermore, we positioned our approach (third step of our methodology, bias mitigation step) to the first category of these algorithms (pre-processing algorithms) as it is used in the data. As we stated, the big difference of our approach is that it does not manipulate the distribution of the data (e.g. oversampling, undersampling, re-weighting etc.), as such a treatment does not provide an unbiased outcome in our use case, but it changes the "nature" of the data itself through manipulating its content.

(RSQ1): Related to our first sub-research question, we managed in this part to analyze all the different bias mitigation algorithms which exist and are used in order to mitigate bias and increase unfairness in Machine learning. More specifically, we gave firstly a clear taxonomy of the different bias mitigation algorithms that exist in literature in mitigating bias and increasing fairness with respect to the point of the Machine learning pipeline which are employed.

[39] https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis
[40] https://archive.ics.uci.edu/ml/datasets/heart+Disease
[41] https://www.federalreserve.gov/boarddocs/rptcongress/creditscore/creditscore.pdf
[42] https://archive.ics.uci.edu/ml/datasets/adult
[43] http://archive.ics.uci.edu/ml/datasets/communities+and+crime

Therefore, the purpose of this part is to provide a taxonomy, review and analysis of the advantages and the limitations of the bias mitigation algorithms that are used to mitigate bias and increase the fairness of Machine learning schemes. The conclusion which we draw here is that there is no a single perfect bias mitigation algorithm that can be used in order to mitigate bias and increase group fairness in data or model level. Thus, the algorithms that one should employ clearly depend on the task at hand and on the domain that concerns.

2.4. BIAS IDENTIFICATION USING CROWDSOURCING

In this part of the literature review, we are dealing with methods that use **crowdsourcing techniques** in order to **identify bias in** the **data**. More specifically, we focus on two types of approaches here. Firstly, we have a look on manners that we can **employ the crowd** for the task of **exploring bias** and **stereotypes** in the **data**. Secondly, we investigate methods that try to **understand the way that humans** actually **perceive** the **bias/fairness** of some **features** and to achieve it, they **use the crowd**. Therefore, the motivation of this part of the literature review is to acquire an understanding of how crowdsourcing can be used as a bias identification technique in the data in order also for us to develop a way that would fit in our work.

2.4.1. METHODOLOGY TO SEARCH FOR PAPERS

In order to search for different ways to use crowdsourcing for bias identification in data, we performed a search on Google Scholar, Scopus and ACM using the combination of the following queries and keywords: "Algorithmic (Un)Fairness and Crowdsourcing", "Algorithmic Bias and Crowdsourcing", "(Un)Fairness and Crowd", "Bias and Crowd". After that we added as keywords the query: "Bias identification in Machine Learning" and we selected all the papers as well as the references of the retained papers that were related to these terms. Finally we combined these queries in the form of "Identify bias in Machine Learning data using crowdsourcing", "Identify bias in Machine Learning data using the crowd", something that enabled us to find more specific papers related to our research questions. As a last step we removed duplicate papers.

2.4.2. USING THE CROWD TO EXPLORE BIAS IN DATA

In this section we focus on studying related work that considers ways of using the crowd as a mean to identify potential bias that may appear in the data. Firstly we pay our attention to literature which considers the perception that the crowd has knowledge about possible bias that may exist in data, which is based probably on their demographical attributes or their personal beliefs. After that we study methods that actually use the crowd in ways that they can discover or identify themselves some "unknown" elements of bias.

In [47] four ways that crowdsourcing can be used in machine learning research are proposed, namely: data generation, evaluation and debugging of models, hybrid intelligence systems and crowdsourced behavioral experiments. Particularly, in the second way (evaluation and debugging of models), also approaches like model interpretability have been investigated. Drawing inspiration from this, we strongly believe that **crowdsourcing** can also used as a **way** of **exploring bias in the data**.

In addition, the idea of employing the crowd to identify the bias in the data can be explained by the diversity of the sample that is offered in such a case [48]. Also, having a **diversity** in the demographical features and also having an understanding of them, provides a better insight in the sense that they **view the world in a different manner** and they may seek for **different forms of bias** as mentioned in [49] and [50].

It is stated in [51] that users' opinion may be shaped and prejudiced with respect to the access that they have on information through algorithmic processes. They mention that highly ranked results in the use case of searching can influence their opinion. More specifically, they are dealing with the problem of the manipulation effect and how skewed results of search engines can shape the preferences of undecided voters in the elections. Therefore, it is crucial to ensure in a way that the results which are returned by the search engines are objective and bias-free. In case that is not possible to eradicate the whole bias that exists in data, more research should be conducted to come up with ways to be aware of the realization that the crowd has potential bias that may exist with respect to their demographic attributes.

As a step further to that, in [52] they propose a framework to measure these potential biases on Tweets searches related to political elections. Particularly, they design this framework in order to distinguish between the bias that comes from data and that one that arises from the model and also propose some definitions in order to quantify them. However, as a next step they could have use this framework not only as a mean to inform users about potential biases that may exist but also to employ them in a way in order for them to be able to compensate for these biases.

There might be also the case that these issues may arise due to the potential trustfulness that users show on the results of the search engines and which they are considered fully reliable by the them. More specifically in [53], they found through the experiments that they conducted that users had a full trust in the results of Google in their queries and they were prejudiced to choose the results that were higher in ranking even in case that they were not so relevant to their requests.

Moreover in [54], they try to understand how people's beliefs and biases that come from these beliefs impact on their decisions. The problem that they examine is how the beliefs that are inherent to people interact with potential biases that arise from the search engines (in case that they are presented results that exhibit skews) affecting the search results. The way to investigate and experiment it was via a survey that they make incorporating in yes-no questions related to people's beliefs that users were asked to answer.

In [55], they are also dealing with the problem of bias and social stereotypes that may appear on the results of search engines. More specifically, they focus on the problem of gender bias in image retrieval search results. The main idea of their work is to come up with ways of detecting gender biases in the aforementioned retrieved search results. To be able to do so, they employ crowdsourcing techniques using participants in the Crowdflower platform with a variety of demographic attributes in order to quantify the degree to which these people believe that these image search results are biased.

More specifically, they use the Ambivalent Sexism Inventory (ASI) [56] in order to figure out they way that users perceive gender bias on these image search results. The main idea of their methodology that they followed in their experiments was as follows: Firstly, people were asked to describe the retrieved images and after that they were informed about the actual query that was used to retrieve them. Finally, a comparison of the predicted description and the real query is performed in order for the objectivity of each user to be assessed, followed by the completion of the ASI questionnaire.

Finally, based on the results of their experiments, they found that individuals, who were rated more sexist (with respect to their ASI scores) were not able to identify gender biases in the image search results, as they realize these results in a different manner in comparison to non-sexist individuals.

In [57], they are dealing with the effect of terms that are biased with respect to gender in the domain of job postings and the impact that they have to job applicants. Particularly, their research is focusing on two layers, namely: Firstly, they propose algorithms which are used in order to identify and measure gender bias. Secondly, they analyze a use case where they try to infer whether the applicants are aware of these biases, to validate the results of their algorithms and finally to measure the impact of potential biases in the decisions of applicants with respect to their applications on specific positions.

On the user-study layer, they designed specific questions for workers on Amazon Mechanical Turk and for undergraduate students in order to understand the levels of understanding of gender bias that they had in the job postings related to particular keywords that were used and the possible impact that this may have for their application's decision.

Based on their experiments, they observed that users were able to identify gender bias in the job postings in a similar degree with the algorithms that they proposed. Furthermore, they found that individuals' inherent beliefs have also a significant effect (in may cases larger than the gender language biases themselves) on the decisions that they made for their applications on jobs.

In a similar fashion in [58] they stress the fact that it is vital to be able to identify and eliminate potential bias that appears in data as it is going to be amplified by the learning algorithm, once a model has been built on this data. Therefore, they propose employing the crowd in order to detect this bias in the data. Their main idea is that in case that they use people for this task who have a variety of demographical features, then it is more likely to identify various kinds of biases and stereotypes (i.e. gender or race).

Particularly, they used the crowd on Amazon's Mechanical Turk (AMT) platform through completing two questionnaires, by asking them to predict possible stereotypes in order to use and impede them from producing biases during the data collection. They strongly believe that ensuring diversity in the demographical characteristics of the crowd workers and knowing these characteristics is a crucial factor in order to be able to detect potential bias that exists in the data.

The conclusions which they drawn were that the performance of the crowd was modest, as they were able to identify only the most common stereotypes. Also they stress that in case that they wanted to achieve better results with respect to identifying more sources of biases in data, more time would be needed for that. Another limitation of their work was the fact that there was a lot of ambiguity in the design of their tasks as they tried to use the crowdworkers in a predictive manner (identify new stereotypes) in comparison to [40], where they designed a more straightforward task.

More specifically in [40], they explored gender bias and stereotypes that may appear in the use case of

word embeddings (a framework in which text data is represented as vectors) in Google News. Their goal is to come up with a methodology in order to be able to alter these word embeddings aiming in debiasing them and removing potential gender stereotypes from them.

The way that they used the crowd on the Amazon Mechanical Turk crowdsourcing platform in their work was through submitting three questionnaires in order to request from them to think about words (i.e. football) that are related with gender stereotypes, to complete analogies (i.e. doctor-man, nurse-woman) which reflect gender stereotypes and to evaluate analogies with respect to the amount of gender stereotyping that have (tall-man, short-woman).

Finally, they compare these results with the mathematical metrics that they used for this purpose (where they found an alignment) and they were able to mitigate the gender bias in the word embeddings.

Similarly in [59], they used the crowd for a slightly different task. The crowdworkers were asked about possible stereotypes which may be closely related to personality, nationality and profession of people for the purpose of animating humans in electronic games.

More particular, they used crowd workers through the Amazon Mechanical Turk platform in order to rate the probability of people conditioned on their nationality or profession to have specific personality traits. Based on the experiments that they performed, they conclude that indeed there were stereotypes related to the personalities of people associated with their nationalities and professions.

Furthermore in a quite different setting in [60], they focus on using the crowd to find the “unknown unknowns” in trained models, namely rare cases for which a model is erroneously pretty confident for its prediction through "playing a game" called “Beat the Machine”. The main idea here is to use the crowd workers to come up with test cases that would baffle the model.

Therefore, in that case there is the possibility for the authors to identify problems that they were not aware of them. They point out that these cases are precious in the sense that although they can occur rarely, they have destructive consequences. Consequently, we can infer here that a similar technique could be employed to **use the crowd for identifying potential elements of bias in the data that we are not aware of them**.

To conclude, in a slightly different context [61], they try to mitigate potential bias that exists in the data in the field of computer vision. The main focus of their work was to investigate the generalization performance of models, namely how well would be the evaluation performance of a model trained on one dataset in case that is tested in a completely different dataset. Thereafter, their idea was that in case that the assumption that a dataset is an accurate enough representation of the world holds, then a model trained on that dataset would exhibit a solid performance on similar datasets in the same domain.

However, based on their experiments they found that this assumption does not hold and the main cause of that was the bias which had been incorporated in the data through the way that had been obtained and labeled by the crowd. Although they did not come with ways of using the crowd to identify this bias in contradiction with the focus of this section of our literature review, they were able to think about innovation solutions of how this data can become bias-free and potentially these techniques may give inspiration of new ways of employing the crowd for that purpose.

2.4.3. USING THE CROWD TO UNDERSTAND HOW HUMANS PERCEIVE THE BIAS/FAIRNESS OF USING SPECIFIC ATTRIBUTES

In this section we focus on studying related work that considers ways of using the crowd through investigating methods that try to realize the manner that humans actually perceive the bias/fairness of using some features.

In [8], in contrast to related work that deals with bias/fairness in ways of making unbiased/fair decisions, they use the crowd in order to understand how they perceive bias in automatic decision making. The motivation behind their work is similar to ours. More specifically, the usage of biased algorithms that make severe decisions with important effects on human lives. The main contribution of their work is the proposal of a pipeline where they try to understand the reasons that the crowd believes that the usage of specific attributes is biased or not in the COMPAS[44] dataset.

After applying their framework, they were able to make four main findings. Firstly, the opinion of the crowd about the fairness of using a specific attribute extends beyond the discrimination. Secondly, there are disagreements on which features may be fair/unfair to use between different people. Thirdly, this lack of agreement comes from the different ways that different people estimate each of the features. Finally, different people that have a common choice of a specific feature, tend also to share a similar justification of their

[44] https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis

choice.

In a similar manner in [9], they focus on the procedural fairness (fairness of the decision making process, (means)). More specifically, they propose three new metrics for the aforementioned procedural fairness which take into account the attributes that are employed for the decision making and evaluate the opinion of the crowd with respect to the usage of these attributes. In order to verify the efficiency and the effectiveness of their methodology they perform experiments on two datasets, namely on the: COMPAS[45] and on the Stop, Question, and Frisk[46] dataset.

Through their experimentation, they were able to achieve a very good performance in terms of the trade-off between process fairness and accuracy through modelling it into some constrained sub-modular optimization problems over the set of the attributes. Finally, based on that, they drew a very interesting and unexpected conclusion: A high value of the procedural fairness may lead also to a high value (some loss of the accuracy is unavoidable by definition) in the distributive fairness (fairness of the outcome of the decision making). However, this holds only on the two aforementioned use-cases that they explored and further research on both theoretical aspect and on more use cases should be conducted for more safe conclusions.

Finally, in a slightly different context in [62] they study the interaction of the crowd with risk assessments via an experimental use-case on Amazon Mechanical Turk. They found that these interactions may produce biases towards the criminal justice. More specifically, it is shown that the decisions of the crowd were inaccurate despite the fact that advice regarding the predictions was provided. Also, the crowd failed in evaluating in a proper way the efficiency of the risk assessment's predictions and as a consequence some forms of bias were introduced. Based on there results, they urge to a need for new pipeline, which they call "algorithm-in-the-loop", aiming in helping people's decisions.

2.4.4. DISCUSSION

In this section, we reviewed methods that use the **crowd** for **bias identification in** the **data**. Especially, we focused on two types of approaches. As a first step, we had a look on ways that we can **use crowdsourcing** for **exploring bias** and **stereotypes** in the **data**. Secondly, we investigated methods that aim to **understand the way that humans** actually **understand** the **bias/fairness** in use of of some **features** in a decision-making process.

(RSQ1)+ (RSQ3): Related to our sub-research questions, we managed in this part to acquire an understanding of how crowdsourcing can be used as a bias identification technique in the data. We strongly believe that the diversity of the crowd might work as a precious source.

Therefore, the purpose of this part, except of acquiring an understanding of how crowdsourcing can be used as a bias identification technique in the data is also to invent a methodology of how we can use the crowd in our work. The conclusion which we draw here is that indeed we can use **crowdsourcing** to help us **uncovering** potential **unknown elements** of **gender bias** that may reside in our Machine learning image training **data**. They way that we are going to do so is further explained in the next chapters.

2.5. SUMMARY

With this chapter, we produced the first contribution **(CO1)** of the thesis: the extensive literature review about definitions of bias (with respect to protected attributes of people) in Machine Learning, evaluation methods to measure bias in Machine Learning, bias mitigation algorithms in Machine Learning and bias identification using crowdsourcing.

Related to **(RSQ1)** definitions and evaluation methods of bias (with respect to protected attributes of people) in Machine Learning, we found that the notion of it that one should adopt clearly depends on the task at hand. More specifically and in order to evaluate our method **(RSQ2+RSQ3)**, based on the related work and the research that we did we decided to use the definition of the statistical parity. Also we provided an extensive taxonomy of these different definitions (statistical, similarity-based and causal-reasoning) and measures (individual bias/fairness, group bias/fairness, data bias/fairness and model bias/fairness). In the sequel, related to **(RSQ1)** bias mitigation algorithms, we ended up with the conclusion that there is no a single perfect bias mitigation algorithm and the choice of them is closely related to the use case and the domain of interest. Moreover, we gave a taxonomy of these different bias mitigation algorithms (Pre-processing, In-Processing and Post-processing) and we also positioned our approach (third step of our methodology, bias mitigation step) to the first of these categories.

[45] https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis
[46] https://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page

Finally, concerning **(RSQ1**+**RSQ3)** bias identification using crowdsourcing, we discovered a lot of influencing ways of how to employ the crowd for this purpose. More specifically, based on the research that we did and in order to adapt it to our case we decided to use crowdsourcing to help us uncovering potential unknown elements of gender bias that may reside in our Machine learning image training data.

3

THE PROFESSION PREDICTION FROM IMAGES USE CASE

3.1. INTRODUCTION

In this chapter, we are interested in describing the use-case, the dataset related to this and the classification task that we study detection, semantic interpretation and mitigation of gender bias. We aim at answering a part of the second and third research sub-questions (**RSQ2**)+(**RSQ3**) and more specifically of defining a classification task in order to study detection semantic interpretation and mitigation of gender bias in Machine learning data.

Particularly, as a use-case we focus on the profession prediction from images. Especially, we are going to examine three binary classification tasks (Doctor-Nurse), (Chef-Waiter) and (Engineer-Farmer). In order to do so, we have collected an amount of image data in order to apply and evaluate our methodology. The remaining of this chapter of the Thesis report is structured as follows: We start by providing a background on the use through referring some related work on profession prediction from images and giving the motivation of our choice. After that, some information about the dataset is given such as a description of the data, example data and some challenges related to that. As a next step, we describe the classification task (predicting occupation from images) that we study. Finally, important conclusions which are drawn from this chapter are mentioned. A roadmap of this Chapter can be depicted in Figure 3.1.

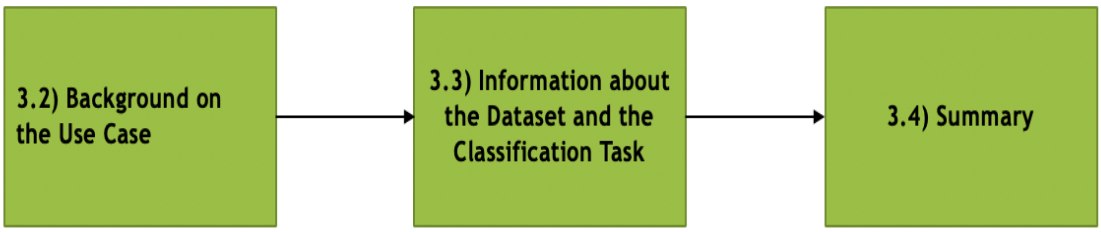


Figure 3.1: Roadmap of Chapter 03

3.2. BACKGROUND ON THE USE CASE

In this section we focus on giving some background on the task of predicting occupations from images. Notably, we start by referring some related work that deals with this task. Our motivation of choosing as a use case the profession prediction from images overlaps also with the motivation of the research works proposed in [63], [64] and [65] and it is twofold: practical application which are free of gender bias and lack of research that takes into consideration the "fairness dimension" of the topic.

Particularly, profession prediction from images has great application potentials in intelligent services and systems. Such a knowledge can contribute in a significant way in analyzing users' profile and their social circle and activities. In that sense, deeper advertising services can be developed on social media platforms and expertise networks in case that their professions can be predicted a priori in an automatic way. As an example, feed of news, propositions of products and friends requests could be dynamically suggested to users in Social Media in an effective manner through recommendation systems.

Except for that, to the best of our knowledge, no research has been previously conducted to study the correlation between profession prediction and gender bias that may be appear in that classification task. Finally, as a crucial step in our methodology is that we leverage the power of the diversity that the crowdsourcing can offer. Thus, our use-case facilitates this purpose and based on the fact that indication of unknown elements of gender bias that may reside in that kind of image data is a subjective property, we could indeed employ this case study in order to verify the effectiveness of our proposed scheme.

3.2.1. RELATED WORK ON THE USE CASE

Predicting occupation from images is a task that it was firstly studied in [63]. More specifically, the authors investigated the predicting occupation from images task through modeling the appearances of human clothing with the surrounding context. As far as the human clothing is concerned, they describe it via modeling on the aligned patches of different human body parts through semantic level patterns (hair, clothes styles etc.). The surrounding context that is used mainly depends on the background of the images. Finally, they applied their methodology in twenty occupations classes, where they found that indeed predicting occupation through modeling the appearances of human clothing with the surrounding context is a doable task.

In a similar manner in [64] the authors advocated that indeed occupation prediction from images is a doable computer vision task. In the methodology that they proposed, they started by extracting multilevel hand-crafted features and linked them with convolutional neural network features, in order to be used as image occupation descriptors. Also, they employed a multi-channel SVM (a boost strategy) to integrate features from face and body. In the evaluation of their approach they found that they can achieve a very good performance in predicting occupation from face, and promising performance with the combination of face and body information. They verified their approach into two datasets that they collected, namely: In the first one they focused only on frontal face images belonging to five different occupations. In the second dataset they combine face and body context information in order to achieve better performance and they deal with twenty-one occupations.

The third and final work (to the best of our knowledge) that deals with the task of predicting occupation from images is [65]. In this work, the authors extended the boundaries of the two aforementioned works in the sense that they deal with multiple people with arbitrary poses in a image. The basic concept of their methodology is the use of structured SVMs that employ built visual attributes and spatial configuration models. Their approach consists of three main steps, namely: Firstly, dense local clothing patches are extracted as invariant low-level features. Secondly, they propose a novel visual attribute learning method which adopts discriminative filters.

Finally, in order to learn the presence of multiple people in the photo they employ a use max-margin training procedure and in order to infer on the learned model they use a greedy forward search. Through their experimentation in fourteen occupations they concluded that predicting occupation from images of multiple people with variations in poses is a doable task and they achieved a decent performance.

It would be interesting and challenging to compare our approach with the three aforementioned methodologies, but to the best of our knowledge, none of the authors released their dataset.

3.3. INFORMATION ABOUT THE DATASET AND THE CLASSIFICATION TASK

As we mentioned in 3.1, we focus on the use-case of profession prediction from images. Especially, we focus on six professions that are split into three different classification tasks, namely: doctor/nurse, chef/waiter and engineer/farmer. The reasoning behind this choice is that we want to have classes of professions for which there are social stereotypes and gender bias. There is a significant amount of references in forums (e.g. here [1,2]) in which there is the notion that in most of the cases doctors should be men and nurses should be women.

[1] https://www.livescience.com/55134-subconscious-stereotypes-hard-to-budge.html
[2] https://journalofethics.ama-assn.org/article/gender-diversity-and-nurse-physician-relationships/2010-01

In addition, in a similar manner, there are social stereotypes (e.g. here [3,4]), that the majority of chefs are male and that of waiters are female. Moreover, there is a similar notion that most of the engineers are males [5]. Hence, based on the above reasons we are going to focus on these six occupations in order to identify whether there exists gender bias in these datasets and in case that this condition holds, to mitigate it.

Furthermore, as we stated in 1.4, the first step of our methodology is related with the bias detection. To do, we use a classification task in order to quantify our notion of gender bias. In the next sections, we are also going to present the classification task (predicting occupation from images) that we use in order to study the detection, semantic interpretation and mitigation of gender bias. Our main goal here is to identify whether there is discrimination in the predictions of a Machine Learning classification model with respect to the gender.

We are going to start this section by giving a description of the dataset. After that some example data of the datasets that we use are going to be presented, followed by identification of the challenges that are related to these data. Finally, an overview of the classification task (predicting occupation from images) that we adopt in order to quantify our notion of gender bias is going to be provided.

3.3.1. DESCRIPTION OF THE DATASET

As we said in 3.1, we are going to focus on three different classification tasks, namely: doctor/nurse, chef/waiter and engineer/farmer. Thus, we have three different datasets in order to evaluate our methodology. The first dataset in which we are going to apply our proposed scheme is a dataset that contains images of doctors and nurses.

The images that we use for the doctor class come from the IdenProf [6] dataset, which is a database that contains some identifiable professionals that were collected in order to facilitate the training of Machine learning systems to identify occupations from images. The creators of this database have also provided a comprehensive datasheet on the dataset for transparency and accountability on the collection and content of this dataset and can be accessed here [7]. Particularly, they provide information such as the ratio of the people with respect to their gender, race, the search engine (Google) that they used in order to collect their data and some other details.

However, this dataset does not contain the class of nurses. For that reason, we had to follow a similar procedure with the creators of the IdenProf [8] dataset and to obtain our images for the nurse class through performing a Google image search. Hence, following this procedure we ended up with 1000 images of doctors and nurses with an equal distribution with respect to the gender.

The second dataset in which we are going to apply our proposed scheme is a dataset that contains images of chefs and waiters. In a similar manner with the previous dataset, we used images which come from the IdenProf [9] dataset. Thus, again we ended up with 1000 images of chefs and waiters with an equal distribution with respect to the gender.

Finally, the third dataset in which we are going to apply our proposed scheme is a dataset that contains images of engineers and farmers. In a similar manner with the previous datasets, we used images which come from the IdenProf [10] dataset. Thus, again we ended up with 1000 images of engineer and farmers with an equal distribution with respect to the gender.

Nonetheless, we also had to inspect these data in order to be able to identify some main elements. We desire to have different ethnic groups and probably different uniforms for the professionals. As an example, we want for the images of doctors to have probably a stethoscope, for the images of nurses to wear short-sleeve tops, for the images of chefs to wear a toque in their heads, for the images of waiters, engineers and farmers to wear a related uniform. Following the aforementioned inspection we were indeed able to verify the presence of these elements.

An overview of the doctor/nurse dataset can be depicted in Table 3.1, of chef/waiter dataset in Table 3.2 and of engineer/farmer dataset in Table 3.3.

[3] https://www.thecaterer.com/articles/368857/male-head-waiters-earn-20-more-than-women
[4] https://www.theguardian.com/lifeandstyle/wordofmouth/2016/may/05/why-are-there-so-few-women-chefs
[5] https://www.bbc.com/news/science-environment-42655179
[6] https://github.com/OlafenwaMoses/IdenProf
[7] https://github.com/OlafenwaMoses/IdenProf/blob/master/idenprof-datasheet.pdf
[8] https://github.com/OlafenwaMoses/IdenProf
[9] https://github.com/OlafenwaMoses/IdenProf
[10] https://github.com/OlafenwaMoses/IdenProf

Doctor	**Nurse**
Total number of images: 1000	Total number of images: 1000
Number of male images: 500	Number of male images: 500
Number of female images: 500	Number of female images: 500

Table 3.1: Doctor/Nurse Dataset Information

Chef	**Waiter**
Total number of images: 1000	Total number of images: 1000
Number of male images: 500	Number of male images: 500
Number of female images: 500	Number of female images: 500

Table 3.2: Chef/Waiter Dataset Information

Engineer	**Farmer**
Total number of images: 1000	Total number of images: 1000
Number of male images: 500	Number of male images: 500
Number of female images: 500	Number of female images: 500

Table 3.3: Engineer/Farmer Dataset Information

3.3.2. EXAMPLE DATA OF THE DATASET

In the section, we are going to present some example image data of our three datasets. The reason of this is for the reader to have a better visual understanding of our data. Four example images (two female and two male) of doctors can be depicted in Figure 3.2.

Example 1 of female doctor

Example 2 of female doctor

Example 1 of male doctor

Example 2 of male doctor

Figure 3.2: Example data of the Doctor class

Similarly, four example images (two female and two male) of nurses can be seen in Figure 3.3.

Example 1 of female nurse

Example 2 of female nurse

Example 1 of male nurse

Example 2 of male nurse

Figure 3.3: Example data of the Nurse class

Four example images (two female and two male) of chefs can be delineated in Figure 3.4.

Example 1 of female chef

Example 2 of female chef

Example 1 of male chef

Example 2 of male chef

Figure 3.4: Example data of the Chef class

Four example images (two female and two male) of waiters can be found in Figure 3.5.

Example 1 of female waiter

Example 2 of female waiter

Example 1 of male waiter

Example 2 of male waiter

Figure 3.5: Example data of the Waiter class

Four example images (two female and two male) of engineers can be delineated in Figure 3.6.

Example 1 of female engineer

Example 2 of female engineer

Example 1 of male engineer

Example 2 of male engineer

Figure 3.6: Example data of the Engineer class

Finally, four example images (two female and two male) of farmers can be found in Figure 3.7.

Example 1 of female farmer

Example 2 of female farmer

Example 1 of male farmer

Example 2 of male farmer

Figure 3.7: Example data of the Farmer class

It is worth mentioning here, that the images had to be reshaped, in order to be used in a proper way in the steps of our methodology that will be presented in the next Chapter.

3.3.3. CHALLENGES RELATED TO THE DATASET

There are two main challenges that are related with our datasets. The first one has to to do with the data collection. More specifically, as we mentioned in 3.2, there are only three research works that dealt with the profession prediction task. Moreover, none of them has released their datasets. Hence, it is very difficult for us to find data. Except for the IdenProf [11] dataset, there is no other open source database that contains data related to occupations. To this end, in case that we wanted to use more data, then we had to collect them ourselves.

The second challenge that is related with our datasets, has to do with the nature of the data themselves. For example, there are some images that contain multiple subjects (e.g. doctor or nurse with patient, group of engineers, waiter with customers etc.). Also, there might be the case that the professional of interest (e.g. doctor, nurse, chef, waiter, engineer of farmer) or some specific features that lead directly to them (e.g. stethoscope, short-sleeve tops, toque etc.) to be partially occluded. In addition, there is also the case that some of the images have been taken in a side-view shot without a lot of details. Finally, the overall resolution of an image might be extremely low to deduce a safe choice. Thus, these are some of the most important challenges that we spotted and might influence in a negative way the performance of our methodology in terms of the evaluation.

3.3.4. CLASSIFICATION TASK: PREDICTING OCCUPATION FROM IMAGES

We have already discussed that we are going to focus on sic professions that are split into three different classification tasks, namely: doctor/nurse, chef/waiter and engineer/farmer. Based on that, three binary classification tasks are going to be studied: In the first one our goal is to infer the label doctor or nurse given an image, in the second one our goal is to infer the label chef or waiter given an image and in the third one our goal is to infer the label engineer or farmer given an image.

Equally important, and something that is also worth mentioning, is the fact we do care not only about the classification outcome itself (doctor/nurse, chef/waiter or engineer/farmer) but also for the confidence that is related with these classification outcomes. Meticulously, we want to have an understanding of how the probabilities that are related with a specific classification outcome vary before and after applying our proposed methodology. For instance, even if there is still a miss-prediction after applying our methodology, it is also important to monitor whether or not there is a difference in the confidence that is related with that classification outcome.

3.4. SUMMARY

In this chapter we investigated three main topics, namely: the use-case, the dataset and the classification task that we focus in order to study the detection, semantic interpretation and mitigation of gender bias.

Firstly, we provided some background information on the use-case (profession prediction from images) in order to verify that this particular use-case is indeed a valid situation to apply our methodology. The two main reasons behind such a choice are: The potential applications that such a task could have and the reduced to minimal related research work that exists on that case.

After that, we gave some details of the dataset that we employ and we verified our choice of the specific professions that we chose based on the existence of some social stereotypes and gender bias. Furthermore, a more detailed description of the dataset is provided through giving some example data points. Moreover, the two main challenges that are related with the dataset are described, namely: the difficulty with the data collection and the nature of the data.

As a final part, we discussed the classification task (predicting occupation from images) that we use in order to quantify the notion of the gender bias that we adopt. It is going to be used in the first step of our methodology that is related with the bias detection. The goal of this step is to identify whether or not exists a discrimination in the predictions of a Machine Learning classification model with respect to the gender.

To this end, and taking into account the analysis that we made in this chapter, we are going to present our methodology for bias detection, semantic interpretation and mitigation in the next chapter.

[11] https://github.com/OlafenwaMoses/IdenProf

4

Methodology for Bias Detection, Semantic Interpretation and Mitigation

4.1. Introduction

In this chapter, we describe the methodology that we propose for bias mitigation, semantic interpretation and mitigation. We aim at answering the second and third research sub-questions (**RSQ2**)+(**RSQ3**) and more specifically of semantically describe at scale the reason that a particular prediction of a Machine Learning classification model is made and finding a way to compensate for gender bias that is related with the content of the image data and actually correspond to the second and third contributions (**CO2**)+(**CO3**) that we make.

Our methodology has three main steps, namely: a bias detection step, a bias semantic interpretation step and a bias mitigation step. A pipeline of the scheme of our methodology can be seen in Figure 4.1.

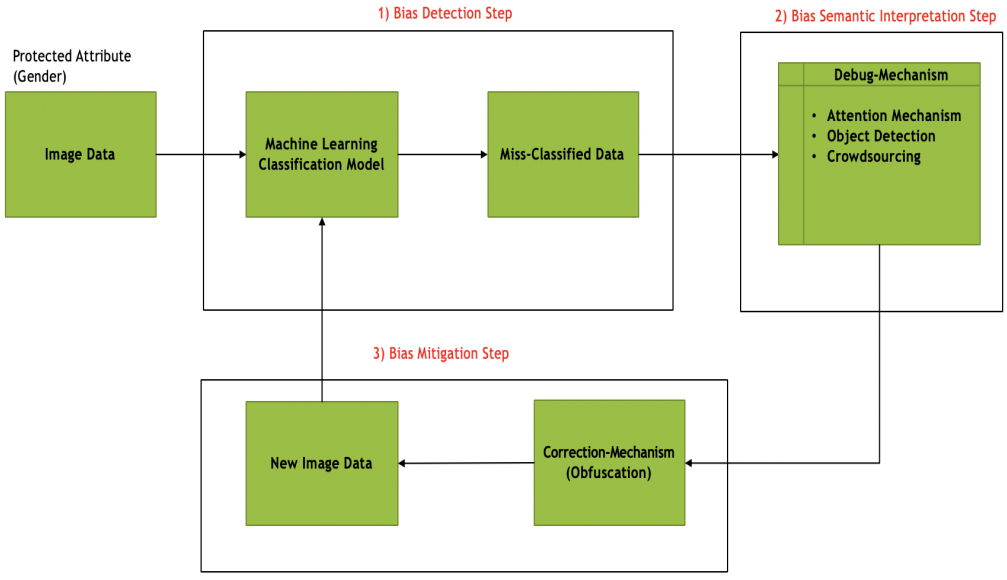


Figure 4.1: Pipeline of the scheme of our methodology

The intuition of our methodology that is depicted in Figure 4.1 is that we start with some image data and we define a protected attribute for them (we adopt the **gender** in our work). After that, we start with the first step

of our methodology (bias detection step), where we pass these images to a Machine Learning classification model and we end up with some miss-classified data. The goal of this step is to observe whether there is discrimination in the predictions of this Machine Learning classification model with respect to the gender. We continue with the second step of our methodology (bias semantic interpretation step), where we pass these miss-classified data to a debug mechanism that has three parts (attention mechanism, object detection and crowdsourcing).

The goal of this step is to semantically describe the reason of these miss-classifications in the predictions of the Machine Learning classification model. Finally, in the third step of our methodology (bias mitigation), we pass the output of the previous step to a correction mechanism (obfuscation task) and we end up with some new image data that we feed them again to the the same Machine Learning classification model. The goal of this step is to compensate for gender bias that is related with the content of the image data.

Now that we gave the intuition of our methodology, we provide more details about it. Particularly, these three steps of our methodology consist of some building blocks, namely: The bias detection step consists of one building block, a classification task. The bias semantic interpretation step consists of two building blocks, where the first of them has three parts, an attention mechanism part, an object detection part and a crowdsourcing part and the second building block has two parts which are the correlation between the attention mechanism part with the object detection part and the correlation between the attention mechanism part with the crowdsourcing part. Finally, the bias mitigation step consists of two building blocks, an obfuscation task and a re-classification task. A visual representation of the steps and their corresponding building blocks and parts of our methodology can be depicted in Figure 4.2.

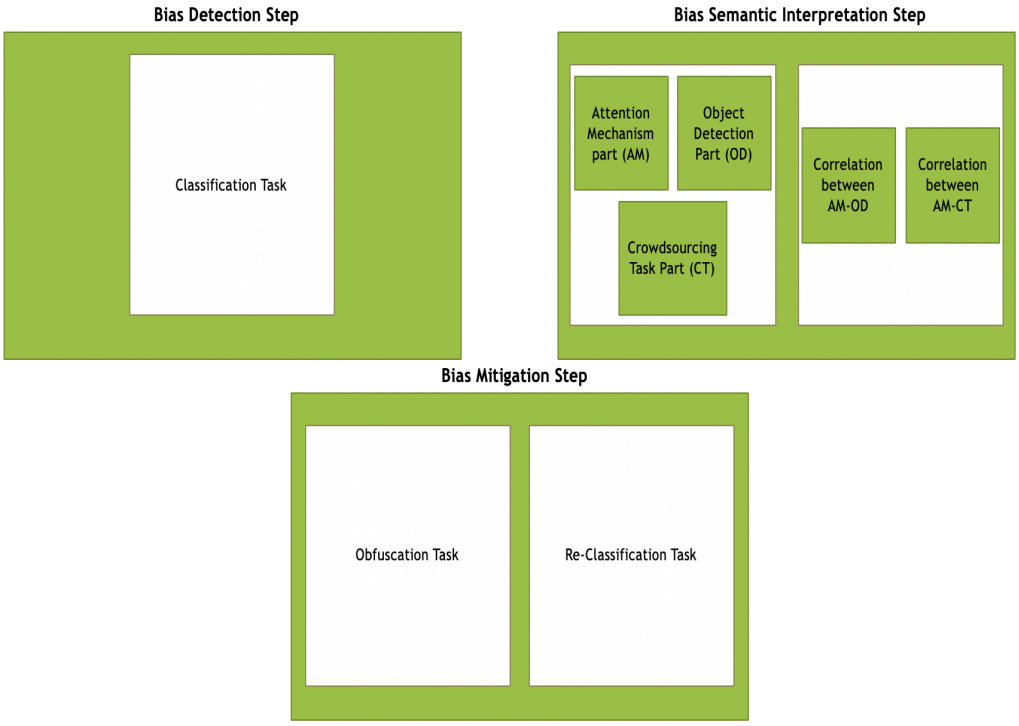


Figure 4.2: Main steps and their corresponding building blocks and parts of our methodology

The remaining of this chapter is structured as follows: We start by providing a background on the attention mechanism, on object detection and on crowdsourcing. As a next step, we describe the first step of our methodology, the bias detection step. The main idea here is to design a classification task and to identify whether there is discrimination in predictions of a Machine Learning classification model with respect to the gender. After that, the second step of our methodology, the bias semantic interpretation step is analyzed. To do so, we describe the attention mechanism, object detection and crowdsourcing task and later we investigate the correlation between the attention mechanism and the object detection and the correlation between the attention mechanism and the crowdsourcing part.

The goal here is to semantically describe at scale the reason that a particular prediction of a Machine

Learning classification model is made. The third step of our methodology, the bias mitigation step comes thereinafter. The intuition here is to obfuscate the outcomes of the previous step and to perform again the classification task in order to compensate for gender bias. Finally, important conclusions which are drawn from this chapter are mentioned. A roadmap of this chapter can be seen in Figure 4.3.

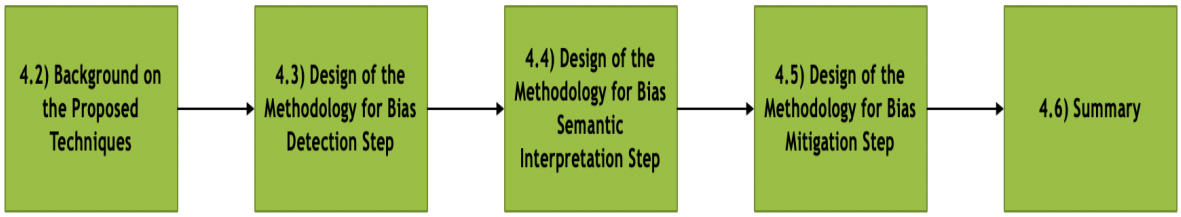


Figure 4.3: Roadmap of Chapter 04

4.2. BACKGROUND ON THE PROPOSED TECHNIQUES

In this section we focus on giving some background on the attention mechanism, on object detection and on crowdsourcing. More specifically, we start by explaining what the purpose of an attention mechanism is. After that, the task of object detection is described. Finally, we present how crowdsourcing can be used in our case based on relevant research work. In order to do so, we refer some related work that deals with these topics.

4.2.1. BACKGROUND ON ATTENTION MECHANISM

Explaining the output of involved classification models like deep neural networks is one of the main challenges that adversaries of these models can pose. However, in case that the area of interest involves image data (like in our work), a good direction in trying to explain the predictions of a particular model is to focus solely on pixels that strongly affect the final classification outcome. As it is stated in [66], a good way to start this methodology is the gradient of the class score function with respect to input image data.

The intuition here is that the aforementioned gradient acts as a sensitivity or saliency map that is typically visualized as a heatmap. Thus, an attention mechanism actually reveals in which features a classification model puts more of its attention in order to make a prediction. In case that we have as a classification model, a deep neural network, the attention mechanism is applied on the very last convolution layer of the network. Therefore, as an an example in a heatmap, these features that matter (that actually are some pixels or blobs of pixels) are colored with a white color and the rest of them are colored with a black color.

There is quite some research that deals with the problem of explaining the prediction of a classification model [66], [67], [68], [69], [70], [71], [72] and [73]. More specifically, the authors in [66] propose a technique which is called SMOOTH-GRAD, and is about understanding how a the network performs a specific classification. The main idea of this technique is that it takes as an input an image data point, called the image of interest, it generates a numbers of sample similar images through adding noise to that image and finally takes the average of the resulting sensitivity or saliency maps for each sampled image. Also they claim that adding noise at training time which is a regularization technique that is proposed in [74] offers an additional de-noising effect on the sensitivity or saliency maps.

Based on the very good results that they achieve in the experimentations that they provide in [66], as they compare their approach with the aforementioned similar research works in terms of the reduction in the noise that the resulting sensitivity or saliency map is going to have and on the goal that we want to achieve, we are going to employ their technique in our bias semantic interpretation step for the attention mechanism part.

However as we stated in 4.1, the one part of our goal is related with the second research sub-question (**RSQ2**) and is to attach a semantic rich description on the reason that a particular prediction of a Machine Learning classification model is made. The main limitation in [66], is that the explanation that is provided in this technique is in form of pixels or blob of pixels. Thereafter, it is not directly interpretable at scale by peo-

ple. This leads us to use also object detection and crowdsourcing, where some pre-defined classes of objects exist (and therefore to be used to attach a semantic label on top of the attention mechanism) and to compute the correlation with the attention mechanism, as we are going to explain in the next sections of this chapter.

Finally, this aforementioned correlation, help us in identifying and obfuscating (it is going to be discussed in the next sections of this chapter) semantically meaningful visual elements that appear in our Machine Learning data and may facilitate our procedure towards compensating for gender bias that is related with the content of the image data and is related with our third research sub-question (**RSQ3**) as we stated in 4.1.

4.2.2. BACKGROUND ON OBJECT DETECTION

Object detection is a task in computer vision which deals with detecting semantic objects that belong to some certain class out of some pre-defined classes. Notably, it is a classification and localization task, as there is a name of a class (classification task) and the position in the image (coordinates of a bounding box) of that specific object as an output.

In general, methods that tackle this object detection task fall into two categories, namely: Machine Learning-based approaches or Deep Learning-based approaches. The main difference of these two categories is the fact that for the Machine Learning approaches, it is mandatory to firstly pre-define some features and after that to do the classification. On the other side of the spectrum, Deep Learning approaches are able to do an end-to-end object detection without defining firstly some features, and the core of them are based on Convolutional Neural Networks (CNNs).

The main three approaches that belong to the first category (Machine Learning approaches) are [75], [76] (which are based on Haar-like features), [77] (which is based on Scale-invariant feature transform (SIFT)) and [78] (which is based on Histograms of oriented gradients (HOG)). In the second category (Deep Learning approaches), we can detect three main different approaches, namely: [79], [80] and [81] (which are based on region proposals), [82] (which is based on discretization of the output space of bounding boxes into a set of default boxes) and finally [83], [84] and [85] (which is based on the fact that they frame object detection as a regression problem to spatially separated bounding boxes and associated class probabilities).

As we said before, in YOLO (You Only Look Once) [83] they frame object detection as a regression problem to spatially separated bounding boxes and associated class probabilities. The main idea here (and the big difference with the aforementioned Deep Learning object detection approaches) is that they break down the images into grid cells. As a next step, the YOLO algorithm associates the middle point of the bounding box to the respective grid cell containing it. However, a direct problem that derives from this is having multiple objects in the same grid cell. In order to tackle this, the authors proposed the use of anchor boxes. The intuition of the anchor boxes is that they add one more dimension to the output labels through defining a priori a number of anchor boxes, something that enables the detection of multiple objects even if they belong in the same grid cell and therefore one object is assigned to each anchor box. Finally, YOLO can be used to perform the object detection task.

In our approach we decided to use as an object detection algorithm the YOLO algorithm (we used a Deep Learning approach as they offer an end-to-end object detection pipeline) [83], for four main reasons, namely: Firstly, in comparison to the other Deep Learning object detection approaches, it is able to predict bounding boxes and class probabilities directly given the input images in one evaluation as the whole detection pipeline is one single network. Therefore, it is extremely fast. Secondly, based on the fact that this algorithm sees the whole image during the training and testing time it implicitly encodes contextual information about classes and for that reason has a significant lower rate of False Positives (FP).

Third, it is able to learn generalizable representations of objects and for that reason it is not prone to overfitting. Finally, the last reason that we chose to go with YOLO is due to the ease that is offers during the implementation in the sense that is pre-trained on multiple different datasets and thereafter is facilitates the experiments with a lot of different classes in comparison to other object detection approaches.

Hence, based on the fact that the object detection task offers a set of classes of pre-defined objects (and on the fact that as we stated in 4.1 we want to attach a semantic label on top of the attention mechanism that is related with our second research sub-question (**RSQ2**) we are going to make a correlation between the object detection and the attention mechanism that will be explained in the next sections of this chapter.

Furthermore, identifying and obfuscating (it is going to be discussed in the next sections of this chapter) semantically meaningful visual elements that appear in our Machine Learning data may facilitate our procedure towards compensating for gender bias that is related with the content of the image data and is related with the first part of our third research sub-question (**RSQ3a**) as we stated in 4.1.

4.2.3. BACKGROUND ON CROWDSOURCING

In the Literature Review that we provided in 2.4.2 we concluded that crowdsourcing can be used potentially as a mean to identify potential bias that may appear in the data. Notable, we stated that the crowd has knowledge about possible bias that may exist in data, that is based probably on their demographical attributes or their personal beliefs. Therefore, it can be inferred that the crowd can be employed in a way in order to discover or identify themselves some "unknown" elements of gender bias that may reside in Machine Learning data that is related with the second part of our third research sub-question (**RSQ3b**).

During the Literature Review that we provided in 2.4.2, we actually identified fourteen related research works that use crowdsourcing in order to explore bias in the data [40], [48], [49], [50], [51], [52], [53], [54], [55], [57], [58], [59], [60] and [61]. However, our work is more similar to that proposed in [58]. As we mentioned in 2.4.2 they stress the fact that it is vital to be able to identify and eliminate potential bias that appears in data as it is going to be amplified by the learning algorithm, once a model has been built on this data. Therefore, they propose employing the crowd in order to detect this bias in the data.

Their main idea here is that in case that they use people for this task who have a variety of demographical features, then it is more likely to identify various kinds of biases and stereotypes (i.e. gender or race). Hence, they used the crowd on Amazon's Mechanical Turk (AMT) platform through completing two questionnaires, by asking them to predict possible stereotypes in order to use and impede them from producing biases during the data collection.

Drawing inspiration from their work, we decided to employ crowdsourcing in a way that people detect elements of gender bias that may reside in Machine Learning data through drawing a bounding box around such an element and writing its name. Particularly, these elements of gender bias are visual clues that give away directly the gender (e.g. face, tie, painted nails etc.).

Thereafter, the reasoning behind this choice is three-fold. Firstly, a crowdsourcing task like this offers a set of classes of objects and as we stated in 4.1 we want to attach a semantic label on top of the attention mechanism that is related with our second research sub-question (**RSQ2**). Secondly, these classes of objects are not pre-defined like in an object detection task and for that reason we could be able to end up with semantic objects that cannot be identified by an object detection approach.

Finally, identifying and obfuscating (it is going to be discussed in the next sections of this chapter) elements of gender bias that appear in our Machine Learning data may facilitate our procedure towards compensating for gender bias that is related with the content of the image data and is related with the second part of our third research sub-question (**RSQ3b**) as we stated in 4.1. Based on these reasons, we are going to make a correlation between the crowdsourcing task and the attention mechanism that will be explained in the next sections of this chapter.

4.3. DESIGN OF THE METHODOLOGY FOR BIAS DETECTION STEP

As we stated in 4.1, the methodology that we propose consists of three main steps. In this section, we are going to present and analyze the first step of our scheme, namely the bias detection step, which consists of one building block, a classification task. Our main goal here is to identify whether there is discrimination in the predictions of a Machine Learning classification model with respect to the gender. We are going to start by giving a quick overview of this bias detection step and after that we are going to provide more details of this first step of our methodology.

4.3.1. DESCRIPTION OF THE BIAS DETECTION STEP

In this first step of our methodology, we focus on how to detect potential gender bias in the predictions of a Machine Learning classification model. To do so, we use a classification task. Therefore we aim in understanding whether there is discrimination in the predictions of this Machine Learning classification model with respect to the gender.

Moreover, as we said in 3.3.4, it is a matter of an equal importance the fact that we do care not only about the classification outcome itself but also for the confidence that is related with these predictions the Machine Learning classification model. Particularly, we want to have an understanding of how the probabilities that are related with a specific prediction vary before and after applying our proposed methodology. For example, even if there is still a miss-prediction after applying our methodology, it is also significant to monitor whether or not there is a difference in the confidence probabilities that are related with that prediction.

4.3.2. DESCRIPTION OF THE CLASSIFICATION TASK

Now, that we gave a quick overview of this bias detection step, the analytic pipeline that we propose for the classification task that lies inside this bias detection step of our methodology is going to be presented. More specifically, an overview of our pipeline for this classification task can be delineated in Figure 4.4.

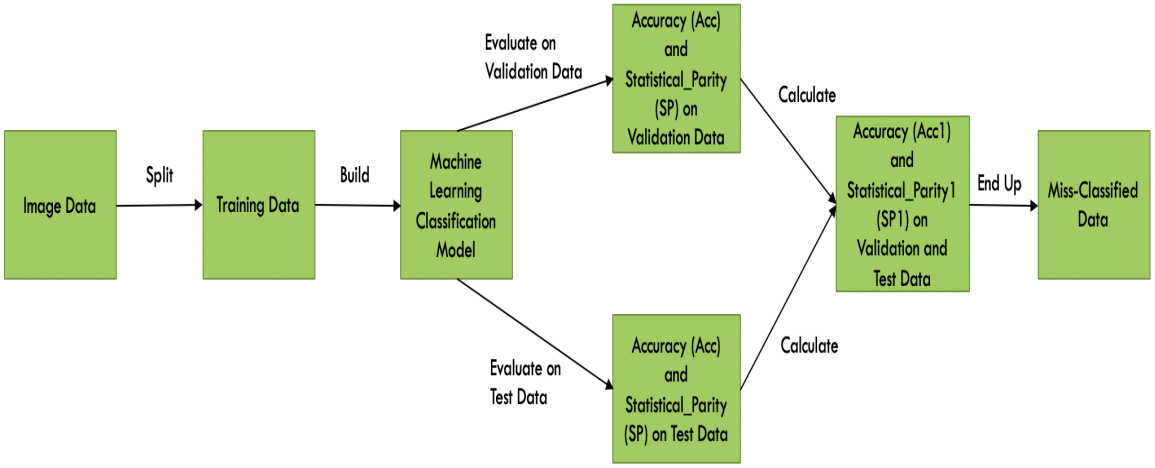


Figure 4.4: Visual overview of our pipeline for the first step for bias detection of our methodology, namely: the classification task

The pipeline that we propose for the classification task of the first step (bias detection) of our methodology consists of five main steps. As a first step with start with some image data. Next to that, we split this dataset of images into three disjoint sets: training, validation and test set. In the second step, based on that training set we build a Machine Learning classification model that is trained with the training data. The third step consists of the evaluation of this Machine Learning model on both the validation and on the test.

We use two measures in order to evaluate this model. The reason that we chose these two specific metrics in order to evaluate our approach is that they actually complement each other. More specifically, the first one provides a good indication about the general performance of the approach and the second one evaluates the approach with respect to a "fairness dimension" (evaluation with respect to gender).

The first measure that we we employ is the accuracy which is given by the formula: $Accuracy = (TP + TN)/(TP + TN + FP + FN)$, where: True positive (TP): In this case both the predicted and actual label belong to the positive class (we assume that one class is the positive and the second is the negative one), False positive (FP): In this case the predicted label belongs to the positive class and the real one belongs to the negative class, False negative (FN): In this case the predicted label belongs to the negative class and the actual one belongs to the positive class, True negative (TN): In this case both the predicted and real label belong to the negative class. Intuitively, accuracy is the number of the correct decisions of our model divided by the total number of the test examples.

Nevertheless, our goal is not only to make an accurate model but also a model that exhibits a good performance in all classes and for all groups of people. For that reason and based on the literature review in that we studied 2.1.3, statistical parity is adopted as our second evaluation measure.

Statistical parity in general is given by the formula: $P(d = 1|G = g_1) = P(d = 1|G = g_2)$, where $d = 1$ represents one class (i.e. doctor), G represents the protected attribute (i.e. gender) and g_1, g_2 represent the protected and unprotected groups respectively (i.e. men and women). Thus, in our case (and let say for the class of doctor) this formula takes the form: $P(d = doctor|G = male) = P(d = doctor|G = female)$.

Therefore, the intuition here is that the classification of people to one class should be independent of the group (protected or not) that belong to and this definition is satisfied in case that individuals in both protected and unprotected groups have equal probability to belong to this predicted class. Thus, our goal here, is to have a model that has similar performance for both male and female.

After that, in the fourth step we are going to calculate the average of the accuracy and statistical parity on

validation and test. Finally, in the fifth step, we are going to end up with the miss-classified images, that we are going to feed next to the second step of our methodology that is going to be presented in the next sections of this chapter.

Thereafter, after completing this first step (bias detection step) of our methodology, we are able to draw two important conclusions based on the two metrics that we use for evaluation, namely: Firstly, to monitor whether our model achieves a good prediction performance based on the value of the accuracy. Secondly, to identify whether there is discrimination in predictions of the Machine Learning classification model with respect to the gender based on the value of the statistical parity.

4.4. DESIGN OF THE METHODOLOGY FOR BIAS SEMANTIC INTERPRETATION STEP

As we mentioned in 4.1, the methodology that we propose consists of three main step. In this section, we are going to present and analyze the second step of our scheme, namely the bias semantic interpretation steps, which consists of two building blocks, where the first of them has three parts, an attention mechanism part, an object detection part and a crowdsourcing part and the second building block has two parts which are the correlation between the attention mechanism part with the object detection part and the correlation between the attention mechanism part with the crowdsourcing part. The goal here is to semantically describe at scale the reason that a particular prediction of a Machine Learning classification model is made.

We are going to start by giving a quick overview of this bias semantic interpretation step and after that we are going to provide more details of this second step of our methodology through analytically describe the two building blocks and their corresponding parts that constitute this second step of our methodology.

4.4.1. DESCRIPTION OF THE BIAS SEMANTIC INTERPRETATION STEP

In this second step of our methodology, we focus on how to semantically describe at scale the reason that a particular prediction of a Machine Learning classification model is made. To do so, we split this step into two building blocks. The first of these building blocks consists of three main parts. We start with the first of these parts that is the attention mechanism. The intuition here is that the output of this part would be some features in the image that a Machine Learning classification model picks up in order to make its predictions. However, our goal is to semantically describe the reason that such a particular prediction is made.

In order to achieve this, we use firstly the second and the third part of this first building block, namely the object detection and crowdsourcing parts, in order to end up with some classes of objects. After that, we proceed in the second building block, where we employ the two parts that constitute this block, namely the correlation between the attention mechanism with the object detection and the correlation between the attention mechanism with the crowdsourcing task. Following this procedure, we are able to semantically describe the reason that a particular prediction of Machine Learning classification model is made as we analytically describe in the next sections.

4.4.2. DESCRIPTION OF THE ATTENTION MECHANISM TASK

In this section we focus on describing the first part of the first building block of this second step for semantic interpretation of bias of our methodology, namely the attention mechanism. The intuition here is that the output of this part would be some features in the image in which a Machine Learning classification model pays more of its attention in order to make its predictions. As we stated in 4.2.1, as an example in a heatmap, these features that matter for the prediction of the Machine Learning classification model (that actually are some pixels or blobs of pixels) are colored with a white color and the rest of them are colored with a black color.

Hence, given as an input an image in a Deep Learning classification model and applying an attention mechanism (e.g. SMOOTH-GRAD) [66]) in the neurons of the last convolutional layer of the deep neural network, we end up with this kind of heatmap. For instance, it can be depicted in Figure 4.5 (left) an image of our dataset (doctor class) that is fed in a Deep Learning classification model and in Figure 4.5 (right) the resulting output after applying an attention mechanism (SMOOTH-GRAD [66]) in the neurons of the last convolutional layer of the deep neural network given this image as an input.

Input image (doctor class)

Output of the Attention Mechanism

Figure 4.5: Input image and output of the Attention Mechanism

Through this Figure 4.5, we can understand that the pixels that have a more bright color are these ones that really affect the classification outcome. Hence, we could say (in a semantic level) that the face or the tie are the features or the objects in which the model pays more of its attention in order to make its predictions. It is also worth mentioning here, that that the model has learned to look at the person's face or at some features that give away directly the gender (e.g. tie) to distinguish doctors from nurses, hence learning a gender stereotype. Thus, this attention mechanism can also be used for bias identification in the data.

However, a direct problem that arises here is that this result is actually in a form of blobs of pixels. Thus, this form of output makes it really difficult or even impossible to provide to people interpretable explanations at scale and in an automatic way of which features in an image matter and lead towards a particular prediction in a Machine Learning model. Based on that, we propose a methodology (step 2-bias semantic interpretation) in order to attach a semantic label on top of the attention mechanism, in the sense that we want to provide a semantically rich description on the contents of the images that actually affect a given prediction. This methodology corresponds to the second contribution (**C02**) of our work and it is related with the second research sub-question (**RSQ2**) that we pose of semantically describe at scale the reason that a particular prediction of a Machine Learning classification model is made.

In order to do so, we are going to use object detection and crowdsourcing, where some pre-defined classes of objects exist (and therefore to be used to attach a semantic label on top of the attention mechanism) and to compute the correlation with the attention mechanism, as we are going to explain in the next sections of this chapter. Nonetheless, there is also one more thing that we need to do before being able to calculate these aforementioned correlations between attention mechanism-object detection and attention mechanism-crowdsourcing.

As we already mentioned in 4.2.2 and in 4.2.3, the output of an object detection algorithm is one (or more) bounding box and the same applies for the crowdsourcing task that we designed, as people detect elements of gender bias that may reside in Machine Learning data through drawing bounding boxes around them. Thereafter, to be able to perform the desired correlation or overlapping between the attention mechanism with the object detection and the attention mechanism with the crowdsourcing, we need firstly to invent a way to draw in an automatic manner bounding boxes around each blob of pixels of the features that matter for the prediction of the Machine Learning classification model in the resulting output of the attention mechanism. A visual description of the problem that we need to solve can be seen in Figure 4.6.

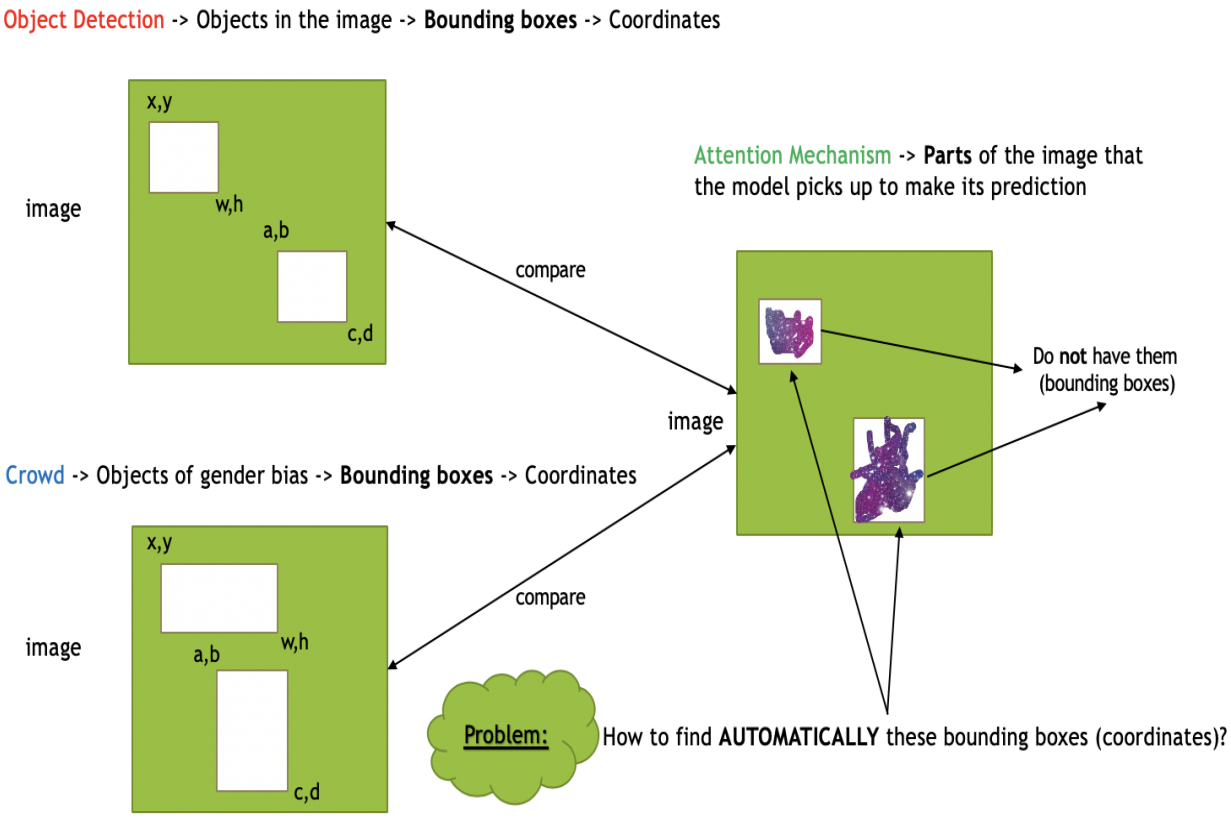


Figure 4.6: Visual Description of the problem of finding bounding boxes for each part in the image that the Machine Learning classification model picks up to make its prediction

In order to be able to tackle this problem, we are going to employ an method that is called Connected Component Analysis or Connected Component Labeling [86] and [87]. The idea here is that this method is used to detect connected regions in images and so it facilitates the blob extraction that is performed on the resulting image through integrating a thresholding step. The inspiration of using this method comes from the research work proposed in [88] where they use Connected Component Analysis or Connected Component Labeling for detection and classication of tumor cells from bone magnetic resonance imagery data.

In our work, we are going to use this method in order to draw in an automatic way bounding boxes around each blob of pixels of the features that matter for the prediction of the Machine Learning classification model in the resulting output of the attention mechanism. Hence, we frame this problem as detection of multiple brightest spots in an image. This technique consists of five main steps that are as follows:

1. We convert the resulted image of the output of the attention mechanism to a gray-scale and we smooth it (e.g. blurring). We do this in order to reduce the high frequency noise. Thus, in that way, brightest regions would be more bright and less bright regions would be more dim.

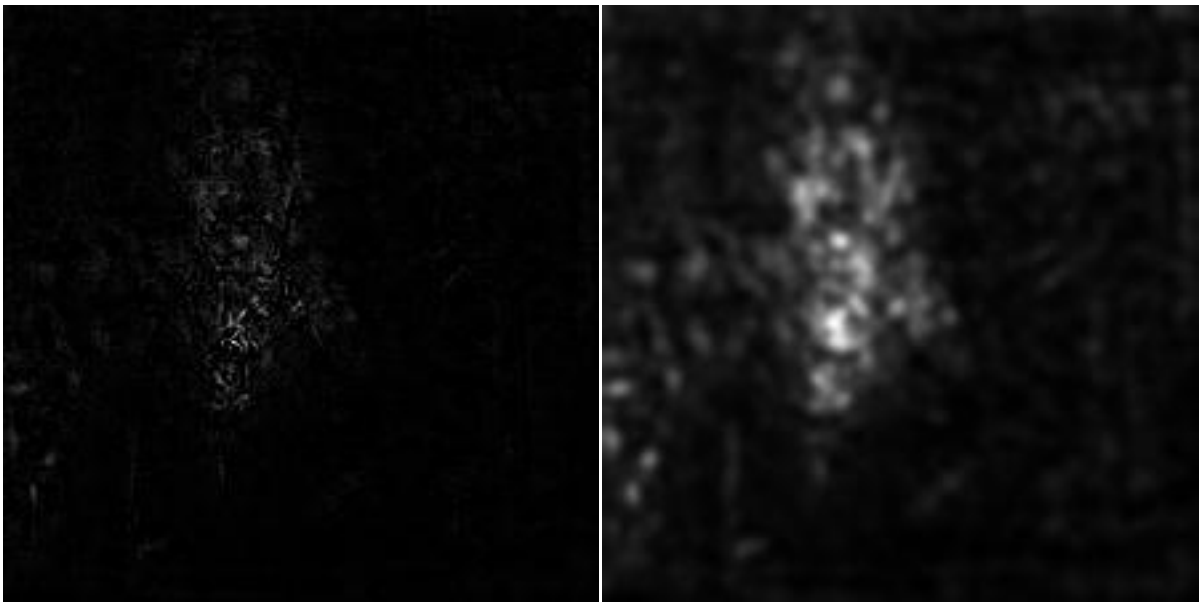
2. We reveal the brightest regions in the resulting blurred image through apply thresholding. More specifically, we say that: Any pixel value $p >= pre-definedthresh$ set it to white and pixel values $p < pre-definedthresh$ are set to black.

3. We clean up the noise (i.e. small blobs) that appears in this image by performing a series of erosions and dilations.

4. We filter out any leftover "noisy" regions through performing a connected-component analysis. Therefore, we end up with a mask that contains only the larger blobs in the image (which are also the brightest ones)

5. Finally, we detect the contours in the mask and a bounding box is automatically drawn for each of them.

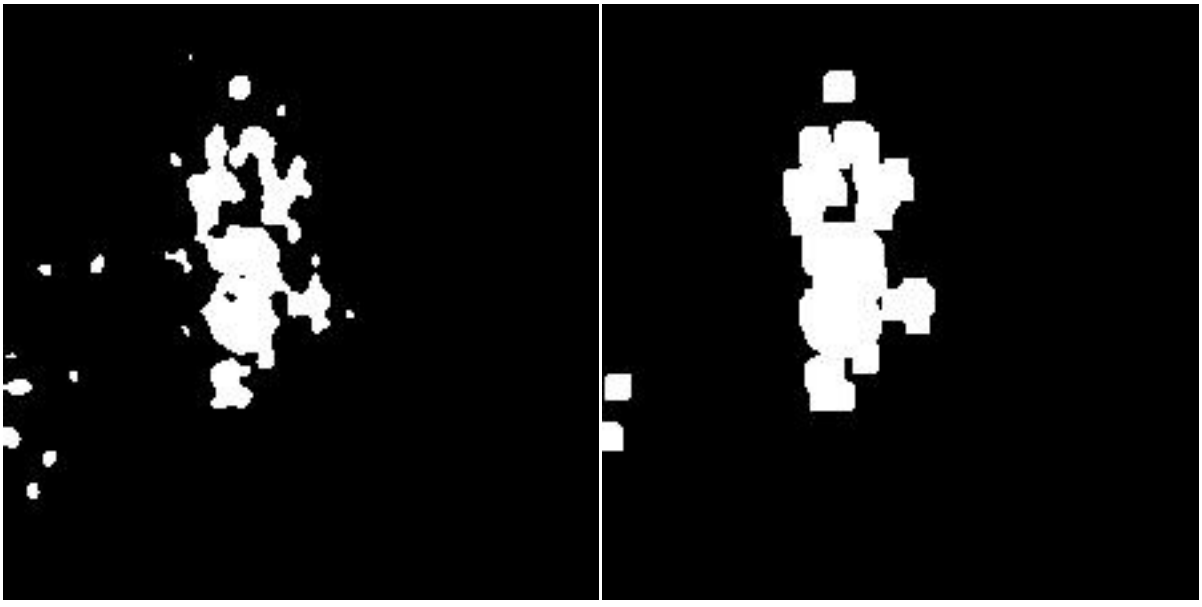
A visual step by step example of this procedure that is based on the input image and on the output of the attention mechanism that we provided in Figure 4.5 can be delineated in Figures 4.7, 4.8 and 4.9.



Gray-scaled image

Filtered Image

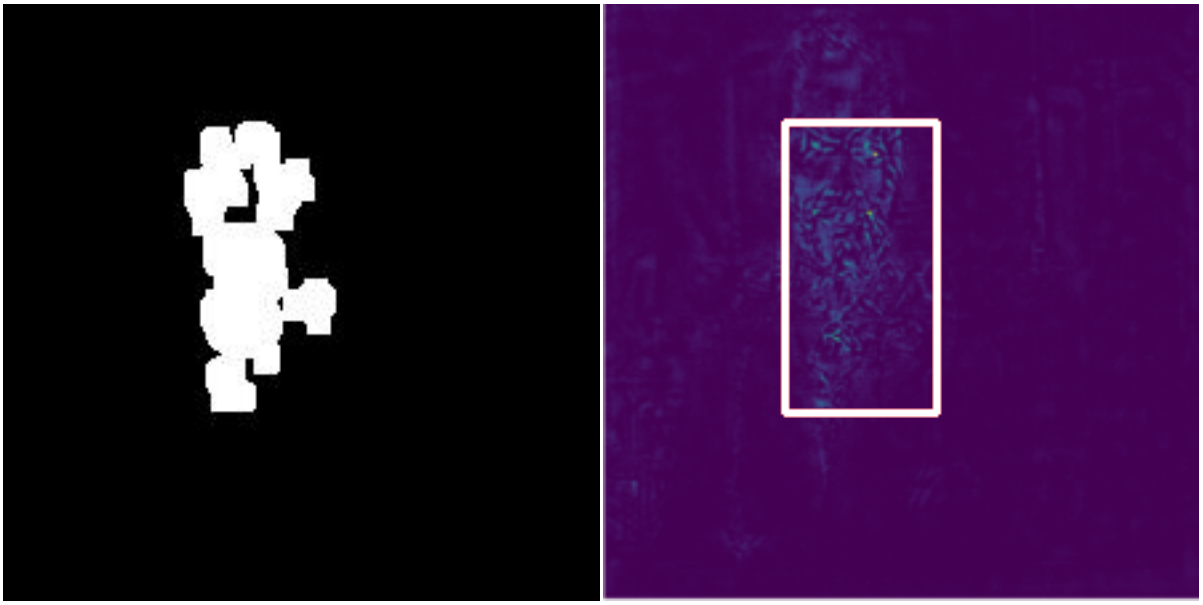
Figure 4.7: Step 1) of the procedure



Resulted image after thresholding

Resulted Image after erosions and dilations

Figure 4.8: Steps 2)-3) of the procedure



Detected mask

Final outcome of the procedure

Figure 4.9: Steps 4)-5) of the procedure

Thereafter, after applying this procedure we end up with one or more bounding boxes for each part in each image that the Machine Learning classification model picks up in order to make its prediction. A visual overview of our pipeline for this first part of the first building block of the second step for semantic interpretation of bias of our methodology, namely: the attention mechanism and the aforementioned procedure of drawing these bounding boxes that we just described can be seen in Figure 4.10.

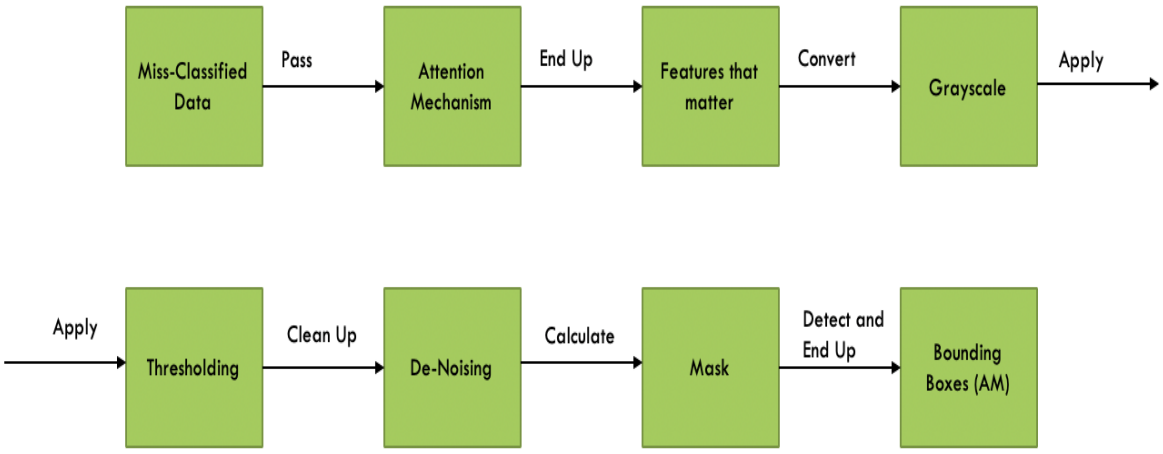


Figure 4.10: Visual overview of our pipeline for the first part of the first building block of the second step for semantic interpretation of bias of our methodology, namely: the attention mechanism and the procedure of drawing bounding boxes

Regarding this pipeline in Figure 4.10, we want to stress three important points. Firstly, the first block of this pipeline, namely the miss-classified data, are the ones that came out from the last block of the pipeline of the classification task in Figure 4.4. Therefore, we pass in the attention mechanism the validation and the test data for which our model gave a wrong prediction, as we want to understand the reason of these miss-predictions.

Secondly, the second and the third block, namely the attention mechanism and the features that matter would be the output of the attention mechanism that is in the form that is described in Figure 4.5. Finally, the other five blocks, namely: gray-scale, thresholding, de-noising, mask and bounding boxes (AM) are the five respective steps of the technique that we described above of detecting multiple brightest spots in the image through applying the Connected Component analysis method. The output of this, would be in the form that is described in Figure 4.9.

So, after implementing this first part of the first building block of this second step for semantic interpretation of bias of our methodology, namely the attention mechanism and the aforementioned procedure of drawing these bounding boxes, we are able to perform the desired correlation or overlapping between the attention mechanism with the object detection and the attention mechanism with the crowdsourcing. Before doing so, we are going firstly to present the second and the third part of the first building block of this second step for semantic interpretation of bias of our methodology, namely the object detection and the crowdsourcing task in the next sections of this chapter.

4.4.3. DESCRIPTION OF THE OBJECT DETECTION TASK

In this section we focus on describing the second part of the first building block of the second step for semantic interpretation of bias of our methodology, namely the object detection part. The intuition here is that the output of this part would be a semantic description of the content of the images. As we stated in 4.2.2, object detection is a classification and localization task. There is a name of a class (classification task) and the position in the image (coordinates of a bounding box) of that specific object as an output.

In addition, as we mentioned in 4.4, our goal in this second step (bias semantic interpretation of bias) of our methodology is to semantically describe at scale, the reason that a particular prediction of a Machine Learning classification model is made. Furthermore, in the next step (third) of our methodology (bias mitigation), our goal is to identify and obfuscate semantically meaningful visual clues that exist in the image data in order to compensate for potential gender bias that may appear in the data.

Therefore, object detection can "serve" us as a mean in order to end up with some classes of semantic objects in which we can build on towards implementing step two and three of our methodology. A visual overview of our pipeline for this second part of the first building block of the second step for semantic interpretation of bias of our methodology, namely: the object detection part that we just described can be depicted in Figure 4.11.

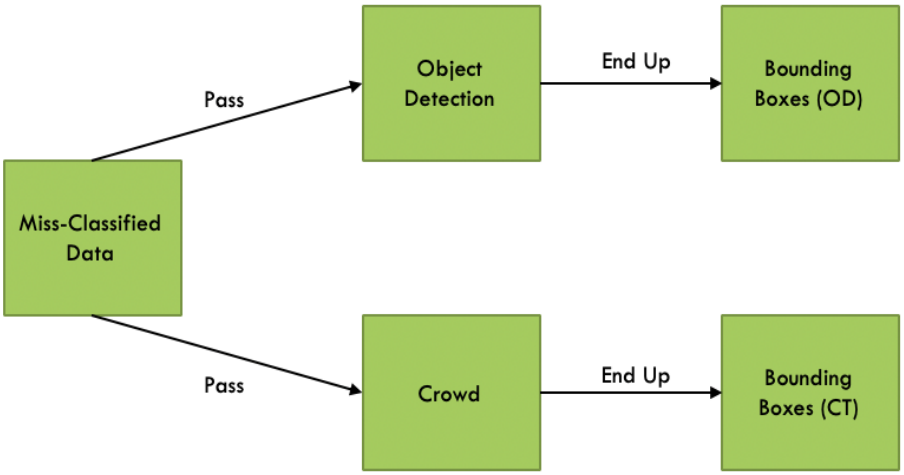


Figure 4.11: Visual overview of our pipeline for the second part of the first building block of the second step for semantic interpretation of bias of our methodology, namely the object detection part

Regarding this pipeline in Figure 4.11, we want to stress three important points. Firstly, the first block of this pipeline, namely the miss-classified data, are the same ones that they are also fed to the attention mechanism in the previous pipeline in Figure 4.10. Secondly, they are also the same ones that came out from the last

block of the pipeline of the classification task in Figure 4.4. Therefore, we pass in the object detection part the validation and the test data for which our model gave a wrong prediction, as we want to understand the reason of these miss-predictions.

Finally, the output of this part (bounding boxes (OD)) would be a semantic description of the content of the images. More specifically, it would contain the name and the position of the objects that appear in the images. An example output of this part with taking as an input the image in Figure 4.5(left) can be depicted in Figure 4.12. We need to emphasize also, that the two lower boxes (crowd and bounding boxes (CT)) are going to be analyzed in the next section as are related with the third part of the first building block of this second step for semantic interpretation of bias of our methodology, namely: the crowdsourcing task.

Figure 4.12: Output of the object detection part

Hence, after implementing this second part of this first building block of the second step for semantic interpretation of bias of our methodology, namely the object detection part we are able to perform the desired correlation or overlapping between the attention mechanism with the object detection and the attention mechanism with the crowdsourcing. Before doing so, we are going firstly to present the third part of the first building block of this second step for semantic interpretation of bias of our methodology, namely the crowdsourcing task in the next section of this chapter.

4.4.4. DESCRIPTION OF THE CROWDSOURCING TASK

In this section we focus on describing the third part of the first building block of the second step for semantic interpretation of bias of our methodology, namely the crowdsourcing part. The intuition here as we mentioned in 4.2.3 is that we designed a crowdsourcing task in which we ask people to detect elements of gender bias that may reside in Machine Learning data through drawing a bounding box around such an element and writing its name. More specifically, these elements of gender bias are visual clues that give away directly the gender (e.g. face, tie, painted nails etc.).

Therefore, the output of this part would be a again semantic description of the content of the images as in 4.4.3. However, the big difference of this crowdsourcing part with the object detection part is that, this semantic description concerns elements that people consider to introduce bias to the data and there are not just elements that constitute the image. An example output of this part with taking as an input the image in Figure 4.5(left) can be depicted in Figure 4.13.

Figure 4.13: Output of the crowdsourcing part

In addition, as we mentioned in 4.4, our goal in this second step (bias semantic interpretation of bias) of our methodology is to semantically describe at scale, the reason that a particular prediction of a Machine Learning classification model is made. Furthermore, in the next step (third) of our methodology (bias mitigation), our goal is to identify and obfuscate semantically meaningful visual clues that exist in the image data in order to compensate for potential gender bias that may appear. However as we stated in 4.2.3, using this crowdsourcing part, we are able not only to identify and obfuscate semantically meaningful visual clues, but also elements of gender bias.

Thereafter, our motivation of implementing this crowdsourcing part is actually two-fold: Firstly, to have an understanding of how the intuition of people about a potential cause of gender bias or unfairness actually compares with the actual reason that affects a prediction of a Machine Learning classification model (to be discussed in the next sections, correlation between attention mechanism and crowdsourcing). Secondly, to gain an insight of how much the intuition of people for elements of gender bias actually matches the semantic description of the images that comes from the object detection part.

An overview of this crowdsourcing task that we designed and submitted in Figure Eight platform can be delineated in Figure 4.14. In this Figure (4.14), the description and the instructions that were sent to crowdworkers of Figure Eight platform can be depicted. Also, we can see an example image, in which people have to draw bounding boxes around elements that give away the gender and a text-box where people include the name of these objects.

So, similar to the object detection part, this crowdsourcing task can "serve" us as a mean in order to end up with some classes of semantic objects in which we can build on towards implementing step two and three of our methodology. A visual overview of our pipeline for this third part of the first building block of the second step for semantic interpretation of bias of our methodology, namely: the crowdsourcing task part that we just described can be depicted in Figure 4.11 (same pipeline as object detection part).

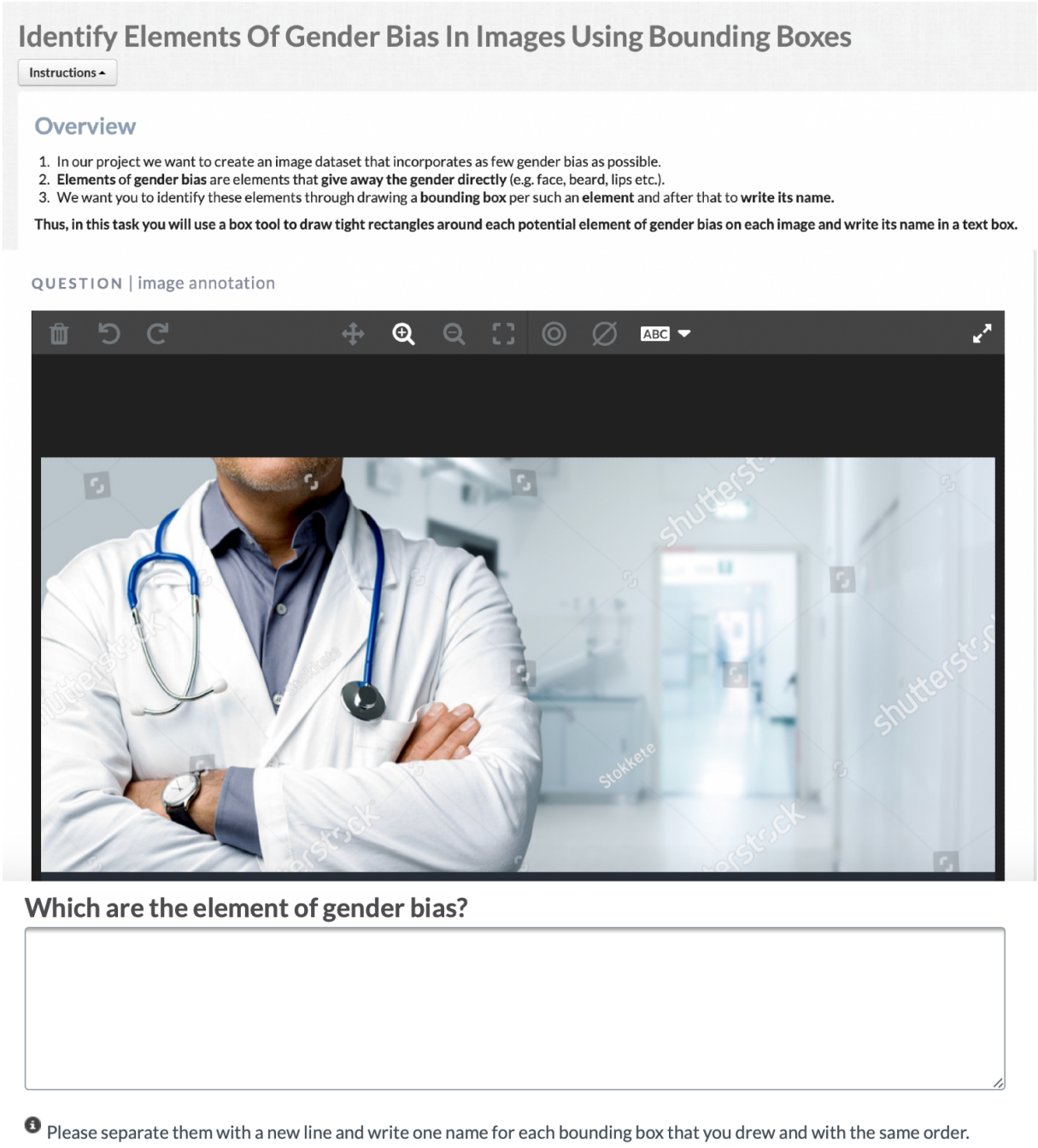


Figure 4.14: Crowdsourcing Task

Regarding the pipeline in Figure 4.11, we want to stress two important points related to this section. Firstly, the procedure is exactly the same as the one that we described in 4.4.3. More specifically, we pass to the crowd the same miss-classified data (that came from the classification task that was described in Figure 4.4), as the ones that we fed to the attention mechanism (Figure 4.10) and to the object detection (Figure 4.11).

Thereafter, we pass in the crowdsourcing task part the validation and the test data for which our model gave a wrong prediction, as we want to understand the reason of these miss-predictions. Secondly, the output of this part (bounding boxes (CT)) would be a semantic description of the content of the image of elements of gender bias. More specifically, it would contain the name and the position of the objects that appear in the images and for which people believe that they introduce bias in the data.

Hence, after implementing this third part of the first building block of this second step for semantic interpretation of bias of our methodology, namely the crowdsourcing task part we are able to perform the de-

sired correlation or overlapping between the attention mechanism with the object detection and the attention mechanism with the crowdsourcing. We are going to analyze these two correlations in the next section through describing the two parts of the second building block of this second step for semantic interpretation of bias of our methodology.

4.4.5. CORRELATION BETWEEN ATTENTION MECHANISM-OBJECT DETECTION AND ATTENTION MECHANISM-CROWDSOURCING

In this section we focus on describing the first and second part of the second building block of the second step for semantic interpretation of bias of our methodology, namely correlation or overlapping between the attention mechanism with the object detection and the attention mechanism with the crowdsourcing. The intuition here is that we want to find a way in order to compute the overlapping of the objects that come from the attention mechanism (features in which a Machine Learning classification model pays most of its attention towards a specific prediction) and the objects that come from the object detection task or the the crowdsourcing task.
Hence, we want to end up with a list of semantic objects (that come from the object detection task or the crowdsourcing task) that matter towards the classification outcome. In case that we have this list of objects, then we can answer our second and third sub-research questions (**RSQ2+RSQ3**) and more specifically to semantically describe at scale the reason that a particular prediction of a Machine Learning classification model is made (it is analyzed in this section) and compensate for gender bias that is related with the content of the image data (it is analyzing in the next section).
Nonetheless, we need firstly to a define measure in order to quantify this overlapping between the objects that come from the attention mechanism, object detection and crowdsourcing task. Based on the fact, that all the objects that come from these three sources are in form of bounding boxes, we decided to adopt as a measure for this quantification the Intersection over Union (IoU) metric.
We also take into consideration as an additional motivation of our choice, that Intersection over Union (IoU) is an evaluation metric which is used to measure the accuracy of an object detector on a particular dataset. Therefore, it is an indicator of overlapping between bounding boxes and for these reasons, it is a perfect candidate for our case. A visual definition of Intersection over Union (IoU) can be depicted in Figure 4.15.

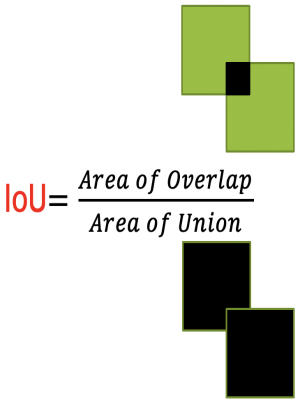


Figure 4.15: Definition of Intersection over Union (IoU)

Thus, examining this equation in Figure 4.15, we can notice that Intersection over Union (IoU) is a quite simple ratio. Specifically, in the numerator, the area of overlap between the two bounding boxes is computed. Similarly, the denominator is the area of their union, or in other words, the area which is encompassed by the two bounding boxes. After that, taking the fraction of the area of overlap dividing by the area of union yields the Intersection over Union (IoU) score. Now, that we have an appropriate metric to measure this overlapping, we are ready to solve the problem of identifying semantic objects that affect the classification outcome. A visual overview of the problem that we frame can be delineated in Figure 4.16.

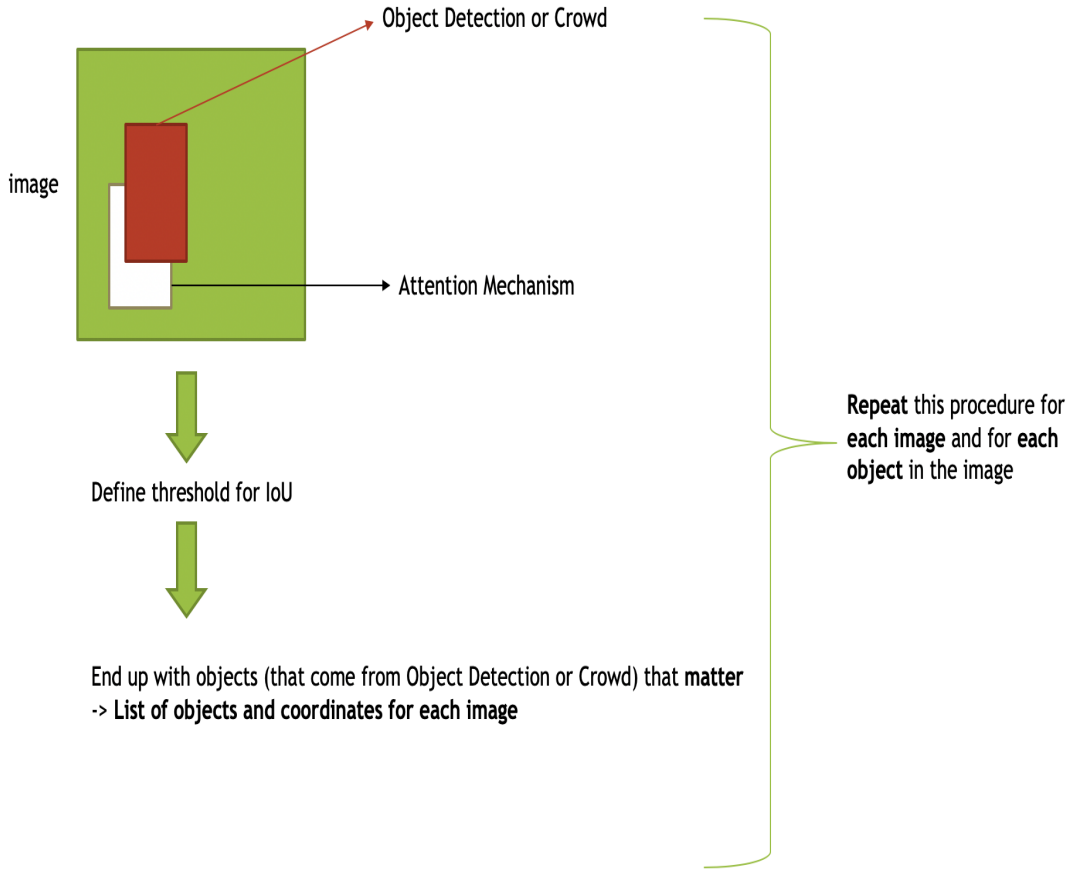


Figure 4.16: Problem of identifying semantic objects that affect the classification outcome

Regarding the problem that we depict in Figure 4.16 we can make the following comments: Firstly, we take as an input the bounding boxes that come from the attention mechanism, object detection and crowdsourcing task. After that, we compute the Intersection over Union (IoU) score for each object that comes from the object detection or crowdsourcing task with all the objects that come from the attention mechanism for each image data point.

Finally, we end up with a list of semantic objects (that come from the object detection task or the crowdsourcing task), that have an Intersection over Union (IoU) score which is larger than a pre-defined threshold and which are the objects that actually matter towards the classification outcome. It is worth mentioning here, that an Intersection over Union (IoU) score larger than 0.5 is considered a good prediction for the object detectors. Therefore, we adopt the same value in our work.

A visual overview of our pipeline for this first and second part of the second building block of this second step for semantic interpretation of bias of our methodology, namely: the correlation or overlapping between the attention mechanism with the object detection and the attention mechanism with the crowdsourcing that we just described can be found in Figure 4.17.

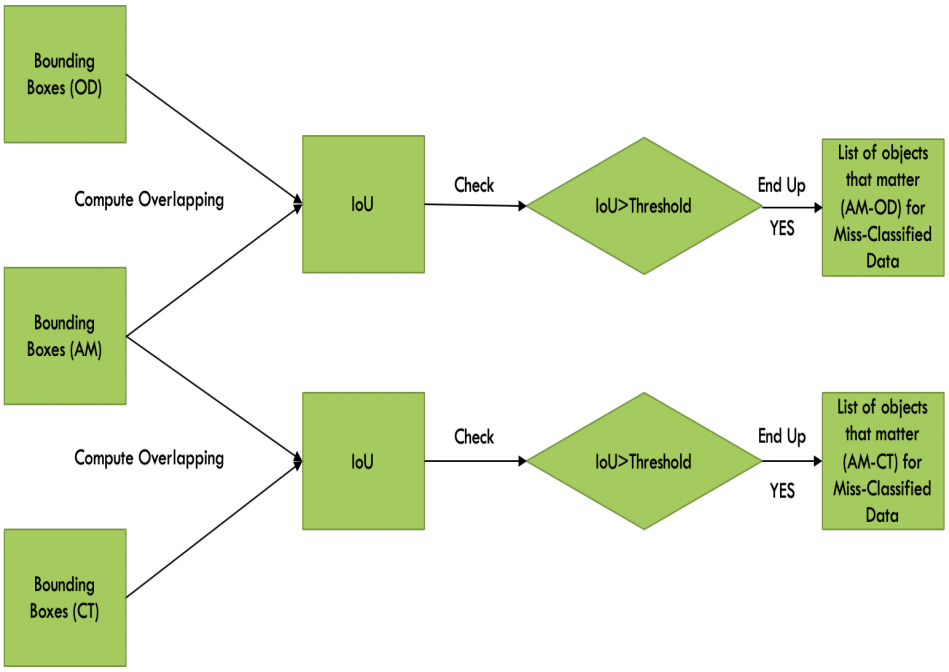


Figure 4.17: Visual overview of our pipeline for the first and second part of the second building block of the second step for semantic interpretation of bias of our methodology, namely: the correlation or overlapping between the attention mechanism with the object detection and the attention mechanism with the crowdsourcing

Regarding this pipeline in Figure 4.17, we want to stress three important points. Firstly, the first three blocks of this pipeline, namely the Bounding Boxes (AM, OD and CT) are the ones that come as an output from the attention mechanism, object detection and crowdsourcing tasks that were discussed previously in pipelines in Figures 4.10 and 4.11.

Secondly, we calculate the Intersection over Union (IoU) score for each object that comes from the object detection or crowdsourcing task with all the objects that come from the attention mechanism. Finally, in case that this score is larger than a pre-defined threshold (we adopt the value of 0.5 for this threshold), we pass these objects to the next block, which is a list of semantic objects (that come from the object detection task or the crowdsourcing task), that actually matter towards the classification outcome for the miss-classified data that come from the pipeline that was described in Figure 4.4.

As an example of this part, let suppose that we take as an input the images of Figures 4.5(left), 4.12 and 4.13. Then we are going to compute the IoU score between the two objects (person and tie, Figure 4.12) that were detected by the object detection algorithm and the two objects (face and tie, Figure 4.13) that were detected by the crowd with the bounding boxes that we found in the attention mechanism (Figure 4.5). These objects that have and IoU score larger than 0.5 are passed to the list of semantic objects and are the ones that actually matter towards the classification outcome (e.g. face, tie etc.).

Consequently, after implementing these two parts of the second building block of this second step for semantic interpretation of bias of our methodology, namely the correlation or overlapping between the attention mechanism with the object detection and the attention mechanism with the crowdsourcing, we are able to answer the second research sub-question (**RSQ2**) that we pose, of semantically describe the reason that a particular prediction is made by a Machine Learning classification model and actually corresponds to the second contribution (**CO2**) that we make.

As a next and final step, we are going to feed this list of semantic objects that matter towards the classification outcome into the next (third) step for bias mitigation of our methodology in the next section.

4.5. DESIGN OF THE METHODOLOGY FOR BIAS MITIGATION STEP

As we mentioned in 4.1, the methodology that we propose consists of three main steps. In this section, we are going to present and analyze the third and final step of our scheme, namely the bias mitigation step, which consists of two blocks: An obfuscation task and a Re-classification task. The goal here is to propose a way to compensate for gender bias that is related with the content of the image data and aims in answering the third research sub-question (**RSQ3**) that we pose.

We are going to start by giving a quick overview of this bias mitigation step and after that we are going to provide more details of this third and final step of our methodology through analytically describe the two building blocks that constitute it.

4.5.1. DESCRIPTION OF THE BIAS MITIGATION STEP

In this third and final step of our methodology, we focus on how to compensate for gender bias that is related with the content of the image data. To do so, we split this step into two building blocks, where each of them consists of one part. We start with the first building block which is an obfuscation task. The intuition here is that we want to give as an input to this task the list of semantic objects that matter towards the classification outcome with which we ended up in the previous section in 4.4.5 and to obfuscate them. Thus, the output here would be some new image data that will be the same as our initial data but some of their parts would be obfuscated.

The second building block is a re-classification task. The intuition here is that we want to feed these aforementioned obfuscated new image data to our initial Machine Learning classification model and to evaluate its performance again (with respect to the metrics that we employ, accuracy and statistical parity) in order to observe whether the predictions of this model have been improved and the gender bias has been reduced.

4.5.2. DESCRIPTION OF THE OBFUSCATION TASK

In this section we focus on describing the first building block of this third step for bias mitigation of our methodology, namely the obfuscation task. The intuition here as we mentioned in 4.5.1 is that we want to obfuscate some objects in the image data that we have in order to end up with some new image. More specifically, we start with the list of semantic objects that matter towards the classification outcome with which we ended up in 4.4.5. After that we obfuscate these objects in the initial miss-classified image data that come from the pipeline that was described in Figure 4.4. An example output of this part with taking as an input the image in Figure 4.5(left) can be depicted in Figure 4.18.

Figure 4.18: Output of the obfuscation part

Doing so, we end up with some new image data that will be the same as our initial data but some of

their parts would be obfuscated. Consequently, as we mentioned in 4.1, our goal in this third step (bias mitigation) of our methodology is to compensate for gender bias that is related with the content of the image data. Before being able to achieve this goal and present a visual overview of our pipeline for this third step of our methodology, we analyze firstly in the next section, the second building block of this step, namely the re-classification task.

4.5.3. DESCRIPTION OF THE RE-CLASSIFICATION TASK

In this section we focus on describing the second building block of this third step for bias mitigation of our methodology, namely the re-classification task. The intuition here as we mentioned in 4.5.1 is that we want to feed the new image data that come from the first building block of this step that was described in 4.5.2 to our initial Machine Learning classification model.

After that, we evaluate its performance again (with respect to the metrics that we employ, accuracy and statistical parity) in order to observe whether the predictions of this model have been improved and the gender bias has been reduced. Therefore, as we mentioned in 4.1, our goal in this third step (bias mitigation) of our methodology is to compensate for gender bias that is related with the content of the image data by combining this building block and the one that we described in 4.5.2.

A visual overview of our pipeline for the two building blocks of this third and final step for bias mitigation of our methodology, namely: the obfuscation and the re-classification task that we just described can be depicted in Figure 4.19.

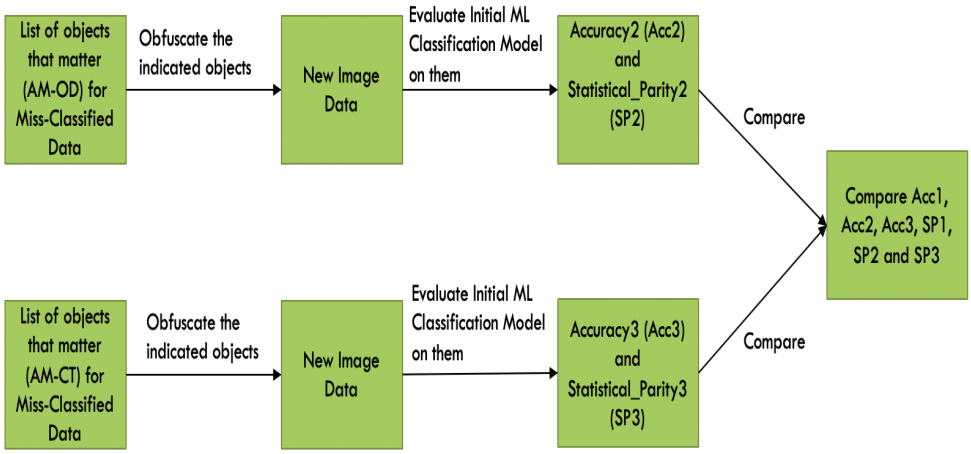


Figure 4.19: Visual overview of our pipeline for the two building blocks of the third and final step for bias mitigation of our methodology, namely: the obfuscation and the re-classification task

Regarding this pipeline in Figure 4.19, we want to stress three important points. Firstly, the first two blocks of this pipeline, namely the list of objects that matter for miss-classified data (AM-CT and AM-OD) are the ones that come as an output from the correlation of attention mechanism-object detection and attention mechanism-crowdsourcing task that was discussed previously in the pipeline in Figure 4.17.

Secondly, we obfuscate these semantic objects that belong to these lists and we end up with some new image data which are the initial miss-classified image data that come from the pipeline that was described in Figure 4.4 with some of of their objects being obfuscated.

Finally, we feed these new image data to our initial Machine Learning classification model and we evaluate its performance again (with respect to the metrics that we employ, accuracy and statistical parity) in order to observe whether the predictions of this model have been improved and the gender bias has been reduced.

Hence, after implementing these two building blocks of the third step for bias mitigation of our methodology, namely the obfuscation and the re-classification task, we are able to answer the third research sub-question (**RSQ3**) that we pose, compensating for gender bias that is related with the content of the image data and actually corresponds to the third contribution (**CO3**) that we make.

4.6. SUMMARY

In this chapter, we described the methodology that we propose for bias detection, semantic interpretation and mitigation and actually corresponds to the second and third contributions (**CO2**)+(**CO3**) that we make. The goal of this chapter was to provide the methodology that we follow in order to answer the second and third research sub-questions (**RSQ2**)+(**RSQ3**) and more specifically of semantically describe at scale the reason that a particular prediction of a Machine Learning classification model is made and compensating for gender bias that is related with the content of the image data.

Firstly, we analyzed the structure of our methodology. We stated that our method consists of three main steps, namely: a bias detection step, a bias semantic interpretation step and a bias mitigation step. After that we presented the structure of each of these three steps. We mentioned that these three steps consist of some building blocks and corresponding parts.

More specifically, the bias detection steps consists of one building block, a classification task. The bias semantic interpretation step consists of two building blocks, where the first of them has three parts, an attention mechanism part, an object detection part and a crowdsourcing part and the second building block has two parts which are the correlation between the attention mechanism part with the object detection part and the correlation between the attention mechanism part with the crowdsourcing part. Finally, the bias mitigation step consists of two building blocks, an obfuscation task and a re-classification task.

On the following sections we analyzed with details these building blocks and their corresponding parts for each of the three steps of our methodology. Particularly, we gave some background information on the attention mechanism, on object detection and on crowdsourcing. As a next step, we described our methodology for bias detection. The main idea here was to design a classification task and to identify whether there is discrimination in predictions of a Machine Learning classification model with respect to the gender.

After that, the bias semantic interpretation step was analyzed. To do so, we described the attention mechanism, object detection and crowdsourcing task and later we investigated the correlation between the attention mechanism and the object detection and the correlation between the attention mechanism and the crowdsourcing part. The goal here was to semantically describe at scale the reason that a particular prediction of a Machine Learning classification model is made.

As a last step, we decomposed the bias mitigation step. The intuition here was to obfuscate the outcomes of the bias semantic interpretation step and to perform again the classification task in order to compensate for gender bias that is related with the content of the image data.

To this end, and taking into account our methodology that we analyzed in this chapter, we are going to evaluate it in the next chapter. More specifically, we are going to present the experiments that we performed for the use-case of profession prediction from images. Especially, we evaluate the bias detection step and the semantic interpretation and mitigation of bias step (two approaches, attention mechanism-object detection (approach 1) and attention mechanism-crowdsourcing (approach 2)) in terms of the evaluation metrics that we adopt.

Particularly, we measure accuracy (good indicator of performance) and statistical parity (good indicator of bias) before and after applying our methodology towards compensating for gender bias that is related with the content of the image data. Finally, we compare these two approaches, with respect to the semantic description of the reason that a particular prediction of a Machine Learning classification model is made, that they end up.

5

EXPERIMENTATION AND EVALUATION OF THE PROPOSED METHODOLOGY

5.1. INTRODUCTION

In this chapter, we are interested in evaluating the methodology that we propose for bias detection, semantic interpretation and mitigation in chapter 4. Particularly, we want to have an understanding of how our methodology performs towards answering our second and third research sub-questions (**RSQ2**)+(**RSQ3**) and more specifically of semantically describe at scale the reason that a particular prediction of a Machine Learning classification model is made and compensating for gender bias that is related with the content of the image.

Meticulously, we present the experiments that we performed for the use-case of profession prediction from images (doctor/nurse, chef/waiter and engineer/farmer). Especially, we evaluate the bias detection step and the semantic interpretation and mitigation of bias step (two approaches, attention mechanism-object detection (approach 1) and attention mechanism-crowdsourcing (approach 2)) in terms of the evaluation metrics that we adopt.

Notably, we measure accuracy (good indicator of performance) and statistical parity (good indicator of bias) before and after applying our methodology towards compensating for gender bias that is related with the content of the image data. Finally, we compare these two approaches, with respect to the semantic description of the reason that a particular prediction of a Machine Learning classification model is made, that they end up.

The remaining of this chapter of the Thesis report is organized as follows: We start by providing the evaluation of the bias detection step, where we want to observe whether there is discrimination in the predictions of the Machine Learning classification model with respect to the gender and is based on the pipeline of the Figure 4.4. As a next step, we evaluate the second and the third step of our methodology, the bias semantic interpretation step and the bias mitigation step. Our goal here is two-fold: Firstly, to semantically describe the reason that a particular prediction of a Machine Learning classification model is made.

Secondly, to observe whether the application of our methodology is able to compensate for gender bias that is related with the content of the image. In order to achieve these goals, we evaluate two different approaches, namely: The first one uses the correlation or overlapping between the attention mechanism and the object detection (Approach 1) and the second one uses the correlation or overlapping between the attention mechanism and the crowdsourcing task (Approach 2) and are based on the pipelines of the Figures 4.10, 4.11, 4.17 and 4.19.

After that, we provide a comparison of the results of these two approaches ((Approach 1)-(Approach 2)) with respect to their performance and differences. Finally, important conclusions which are drawn from this chapter are mentioned. A roadmap of this chapter can be seen in Figure 5.1.

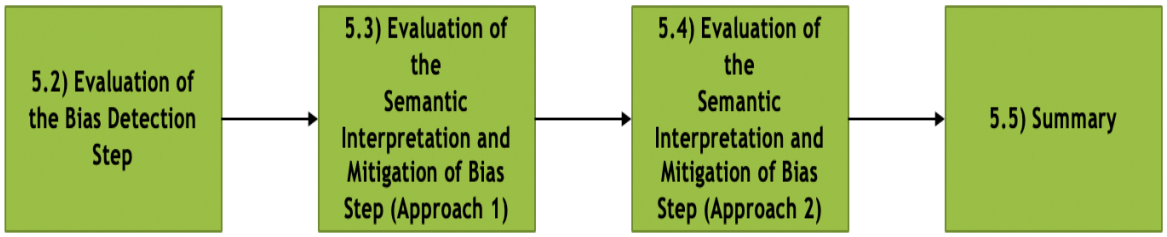


Figure 5.1: Roadmap of Chapter 05

5.2. EVALUATION OF THE BIAS DETECTION STEP

As we mentioned in 4.1, the methodology that we propose consists of three main steps. In this section, we provide the evaluation of the first step, the bias detection step, where we want to observe whether there is discrimination in the predictions of the Machine Learning classification model with respect to the gender and is based on the pipeline of the Figure 4.4. As we stated in 4.3, this step consist of one building block, a classification task. Therefore, we use this classification task for evaluating this step. The metrics that we adopt for this reason are accuracy (good indicator of performance) and statistical parity (good indicator of bias). We start this section by providing the implementation details of the evaluation of the bias detection step. After that, the results are given and we end up with a discussion of these results and with important conclusions that are drawn.

5.2.1. IMPLEMENTATION DETAILS

In this section, we provide the implementation details of the evaluation of the bias detection step. As we stated in 3.1, the use-case of profession prediction from images is adopted. More specifically we focus on three classifications tasks, doctor/nurse, chef/waiter and engineer/farmer. Our implementation follows the order of the pipeline in Figure 4.4. The first block in this pipeline is Image Data. The data that we use for these three classification tasks (doctor/nurse, chef/waiter and engineer/farmer) are the ones that are described in Tables 3.1 and 3.2.

Based on the same pipeline, as a next step we split these data into training, validation and test set. Particularly, we split the initial data into 80%-20% train and test set and after that we split the train data into 80%-20% train and validation set. We use the train data to build our classification model, the validation data for tuning our hyperparameters and for evaluation and the test data for evaluation. Based on that, the information about our datasets (doctor/nurse, chef/waiter and engineer/farmer) for these classification tasks can be depicted in Tables 5.1, 5.2 and 5.3.

Total number of data	Size of Training Set	Size of Validation Set	Size of Test Set
1000 (500 males and 500 females)	640	160	200

Table 5.1: Doctor/Nurse Dataset for the Classification Task

Total number of data	Size of Training Set	Size of Validation Set	Size of Test Set
1000 (500 males and 500 females)	640	160	200

Table 5.2: Chef/Waiter Dataset for the Classification Task

Total number of data	Size of Training Set	Size of Validation Set	Size of Test Set
1000 (500 males and 500 females)	640	160	200

Table 5.3: Engineer/Farmer Dataset for the Classification Task

The next step of our pipeline is to build a Machine Learning classification model based on the training data. Firstly, we need to mention that we use a Deep Learning approach, mostly based on the promise that has shown recently in image processing tasks. Thereafter, our classification model is a convolutional neural network [89]. More specifically we chose a pre-trained ResNet (residual network) Model [90] trained on the ImageNet dataset. The reason that we chose a pre-trained network is due to the huge amount of timing and computing power that is needed in order to train a deep neural network from scratch. Also, we chose ResNet (residual network) model due to its decent performance in similar image processing tasks.

Based on the small number of data that we have available, we perform also some data augmentation configurations like rotate, rescale, and flip of the data in order to avoid overfitting issues. As we used a pre-trained model, the next step of our procedure is to employ a Transfer Learning approach, where we use the convolutional layers for feature extraction. The idea here is that even though the pre-trained model is trained on a different task than our task at hand, it provides a useful starting point as the features that are learned while training on the old task are also useful for our task.

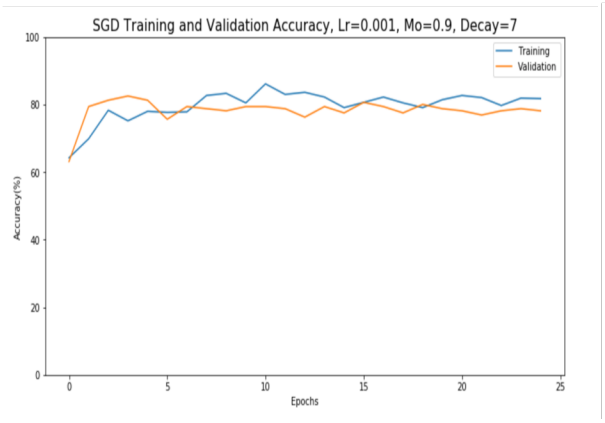
However, in order for this pre-trained model to be fully useful to us, we need one extra step. This step is the addition of a classifier on top of the convolutional base. More specifically, we add a fully connected layer that is followed by a softmax layer with 2 outputs and is based on our data. Finally, this softmax layer outputs the probability distribution over each possible class label and then the images are classified according to the most probable class (doctor/nurse, chef/waiter and engineer/farmer). Now that we have built our classification model, we use the validation set in order to tune its hyperparameters. The values that we ended up can be depicted in Table 5.4.

Hyperparameters	**Value**
Learning Rate	10^{-3}
Optimizer	SGD
Multiplicative factor of learning rate decay	0.1
Period of learning rate decay	7
Dropout of the FC layers	0.5
Activation Function	ReLU
Loss Function	Binary Cross Entropy
Momentum	0.9

Table 5.4: Values of the Hyperparameters

Regarding this Table 5.4, we can see that the learning rate (determines to what extent newly acquired information overrides old information) ends up with a value of 10^{-3}. The best optimizer (iterative method for optimizing the objective function) is SGD (Stochastic Gradient Decent). The multiplicative factor of learning rate decay (adjustment of the learning rate during training, reducing the learning rate according to a predefined schedule) is 0.1. The period of learning rate decay is 7 (decay of the learning rate every 7 epochs). The dropout (regularization technique for reducing overfitting, dropping out percentage of the neurons) of the fully connected layers is 0.5. The activation function (node that defines the output of that node) is ReLU. The loss function (represents some "cost" that is related with an action) is a Binary Cross Entropy. Finally, the momentum (helps in accelerating gradients vectors in the right directions) has a value of 0.9.

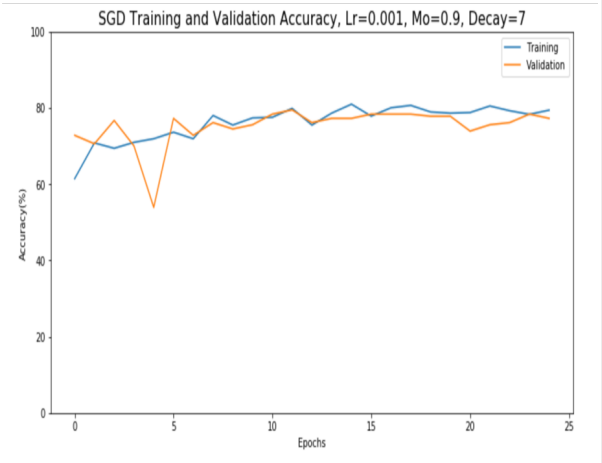
Moreover, in Figures 5.2, 5.3, 5.4 we can depict the training and validation accuracy-loss curves for the doctor-nurse classification task, for the chef/waiter classification task and for the engineer/farmer classification task respectively. From these Figures we can see that our model does not overfit at all, as the there is almost no gap between the accuracy-loss curves for the training and the validation set for all datasets (doctor/nurse, chef/waiter and engineer/farmer).



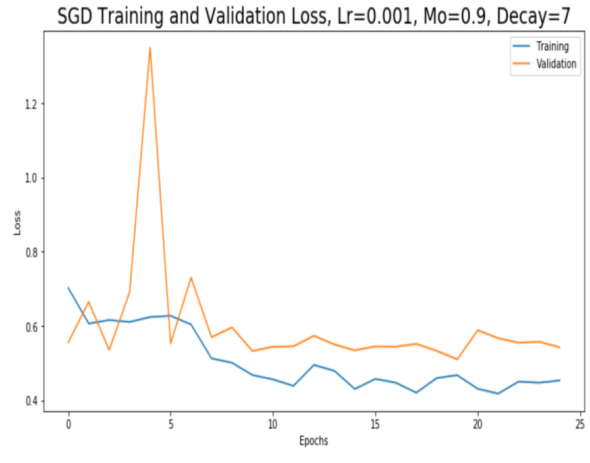
Training and validation accuracy curve

Training and validation loss curve

Figure 5.2: Training and validation accuracy-loss curves for the doctor-nurse classification task

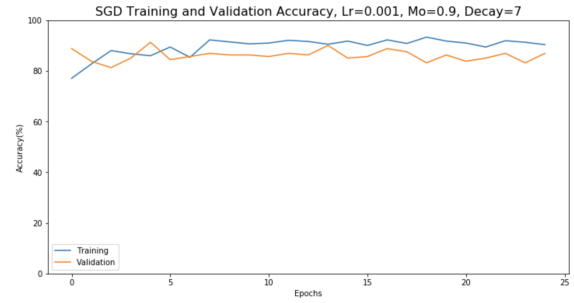


Training and validation accuracy curve

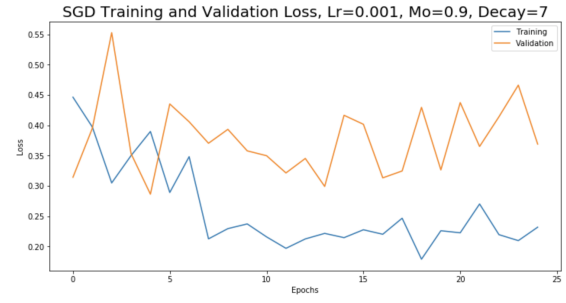


Training and validation loss curve

Figure 5.3: Training and validation accuracy-loss curves for the chef-waiter classification task



Training and validation accuracy curve



Training and validation loss curve

Figure 5.4: Training and validation accuracy-loss curves for the engineer-farmer classification task

According, to our pipeline in Figure 4.4, now that we have built our classifications model, the next step is to evaluate it based on the metrics that we use (accuracy (good indicator of performance) and statistical parity (good indicator of bias)) and to end up with the miss-classified data. The goal of this part is to observe whether there is discrimination in the predictions of our classification model with respect to the gender. We present

the results on our datasets (doctor/nurse, chef/waiter and engineer/farmer), followed by a discussion in the next section.

5.2.2. RESULTS AND DISCUSSION

In this section, we provide the results followed by a discussion that is related with the bias detection step. As we stated in 5.2.1, and according to our pipeline in Figure 4.4 for this step, now that our classifications model is built, the next box in this pipeline is to evaluate it based on the metrics that we adopt (accuracy (good indicator of performance) and statistical parity (good indicator of bias)) and to end up with the miss-classified data.

Our goal in this section is to observe whether there is discrimination in the predictions of our classification model with respect to the gender. We are now ready to present these results on our datasets (doctor/nurse, chef/waiter and engineer/farmer). We start the with the evaluation of the doctor/nurse dataset. It can be delineated in Table 5.5 the prediction performance of our classification model per class (doctor/nurse) on validation and test set.

Validation Set	**Test Set**
Doctor Accuracy: 80%	Doctor Accuracy: 80%
Nurse Accuracy: 88.8%	Nurse Accuracy: 81%

Table 5.5: Prediction performance per class on validation and test set (doctor-nurse dataset)

According to Table 5.5 we can stress three important points. Firstly, we see that our model in general achieves a quite good performance. Overall, it has an accuracy of at least 80% for each class on both validation and test set. However, we can say here that based on the fact that we have only two classes, this performance is not perfect.

Secondly, we observe that there is a quite large difference in the prediction performance in validation set for the class of doctor and nurse. Particularly, our model has better performance on nurse class in comparison to doctor class.

Finally, based on this Table 5.5, we cannot understand how our model performs across the two genders (male and female), that is the goal of this section to observe whether there is discrimination in the predictions of our classification model with respect to the gender. We also conjecture, that this difference between the doctor and nurse class may be due to difference in performance across the gender.

Therefore, we break down the performance of the classification model per gender in order to verify our hypothesis. It can be depicted in Table 5.6 the prediction performance of our classification model per gender (male/female) on validation set for the doctor and nurse class.

Doctor Class	**Nurse Class**
Male Accuracy: 86.6%	Male Accuracy: 86.6%
Female Accuracy: 73.3%	Female Accuracy: 91.1%

Table 5.6: Prediction performance per gender (male/female) on validation set for the doctor and nurse class

According to Table 5.6 we can stress two important points. Firstly and most importantly, we see that there is a significant gender bias in the predictions in doctor class (13.3% difference in accuracy). This difference in accuracy between the two class is the definition of the statistical parity. Based on that, we can infer that there is discrimination in the predictions of our classification model with respect to the gender for the doctor class in the validation set. Thus, there are a lot of female images that belong to the doctor class and the classification for them was the nurse class.

Secondly, we observe that there is no big difference (4.5% difference in accuracy) in the predictions of our classification model with respect to the gender for the nurse class in the validation set. Now we proceed with the evaluation of our model in the test set. It can be seen in Table 5.7 the prediction performance of our classification model per gender (male/female) on test set for the doctor and nurse class.

Doctor Class	**Nurse Class**
Male Accuracy: 88%	Male Accuracy: 84%
Female Accuracy: 72%	Female Accuracy: 78%

Table 5.7: Prediction performance per gender (male/female) on test set for the doctor and nurse class

According to Table 5.7 we can stress two important points. Firstly and most importantly, we see again that there is a significant gender bias in the predictions in doctor class (16% difference in accuracy). Therefore, the value of the statistical parity is 16. We need to emphasize here, the in order for a classification model to be bias neutral, we want for the statistical parity to be as close as possible to 0. Based on that, we can infer that there is discrimination in the predictions of our classification model with respect to the gender for the doctor class in the test set. Thus, again like in the validation set, there are a lot of female images that belong to the doctor class and the classification for them was the nurse class.

Secondly, we observe that there is no big difference (6% difference in accuracy) in the predictions of our classification model with respect to the gender for the nurse class in the test set. Now, in order to have a better overview, we proceed with the evaluation of our model in both the validation and test set by taking the average of these parts. It can be delineated in Table 5.8 the prediction performance of our classification model per gender (male/female) on validation and test set for the doctor and nurse class.

Doctor Class	**Nurse Class**
Male Accuracy: **87.4%**	**Male** Accuracy: **85.3%**
Female Accuracy: **72.6%**	**Female** Accuracy: **84.5%**

Table 5.8: Prediction performance per gender (male/female) on validation and test set for the doctor and nurse class

According to Table 5.8 we can stress two important points (similar to Tables 5.6-5.7. Firstly and most importantly, we see again that there is a significant gender bias in the predictions in doctor class (14.8% difference in accuracy). Therefore, the value of the statistical parity is 14.8, which is such a large value. Hence, we can infer again that there is significant discrimination in the predictions of our classification model with respect to the gender for the doctor class in the validation and test set.

Secondly, we observe that there is a negligible difference (0.8% difference in accuracy) in the predictions of our classification model with respect to the gender for the nurse class in the validation and test set. Thereafter, the conclusion that we draw for the first step of our methodology (bias detection step) for the doctor-nurse dataset is that there is a significant gender bias in the predictions in doctor class, even though there is an equal distribution among the doctors with respect to the gender in the training set.

Now, we present a similar evaluation of the first step of our methodology (bias detection step) of the chef/waiter dataset. It can be delineated in Table 5.9 the prediction performance of our classification model per class (chef/waiter) on validation and test set.

Validation Set	**Test Set**
Chef Accuracy: 74.4%	Chef Accuracy: 71%
Waiter Accuracy: 84.4%	Waiter Accuracy: 86%

Table 5.9: Prediction performance per class on validation and test set (chef-waiter dataset)

According to Table 5.9 we can stress three important points. Firstly, we see that our model in general achieves a quite good performance. Overall, it has an accuracy of at least 70% for each class on both validation and test set. However, we can say here that based on the fact that we have only two classes, this performance is not perfect.

Secondly, we observe that there is a quite large difference in the prediction performance in validation and test set for the class of chef and waiter. Particularly, our model has better performance on waiter class in comparison to chef class.

Finally, based on this Table 5.9, we cannot understand how our model performs across the two genders (male and female), that is the goal of this section to observe whether there is discrimination in the predictions of our classification model with respect to the gender. We also conjecture, that this difference between the chef and waiter class may be due to difference in performance across the gender.

Therefore and similar to the previous dataset (doctor/nurse), we break down the performance of the classification model per gender in order to verify our hypothesis. It can be depicted in Table 5.10 the prediction performance of our classification model per gender (male/female) on validation set for the chef and waiter class.

Chef Class	**Waiter Class**
Male Accuracy: 88.8%	Male Accuracy: 88.8%
Female Accuracy: 60%	Female Accuracy: 80%

Table 5.10: Prediction performance per gender (male/female) on validation set for the chef and waiter class

According to Table 5.10 we can stress two important points. Firstly and most importantly, we see that there is a huge gender bias in the predictions in chef class (28.8% difference in accuracy). Thus, the value of the statistical parity here is 28.8. Based on that, we can infer that there is discrimination in the predictions of our classification model with respect to the gender for the chef class in the validation set. Thus, there are a lot of female images that belong to the chef class and the classification for them was the waiter class.

Secondly, we observe that there is no such a big difference (8.8% difference in accuracy) in the predictions of our classification model with respect to the gender for the waiter class in the validation set. Now we proceed with the evaluation of our model in the test set. It can be seen in Table 5.11 the prediction performance of our classification model per gender (male/female) on test set for the chef and waiter class.

Chef Class	**Waiter Class**
Male Accuracy: 84%	Male Accuracy: 82%
Female Accuracy: 58%	Female Accuracy: 90%

Table 5.11: Prediction performance per gender (male/female) on test set for the chef and waiter class

According to Table 5.11 we can stress two important points. Firstly and most importantly, we see again that there is a huge gender bias in the predictions in chef class (26% difference in accuracy). Therefore, the value of the statistical parity is 26. Based on that, we can infer that there is discrimination in the predictions of our classification model with respect to the gender for the chef class in the test set. Thus, again like in the validation set, there are a lot of female images that belong to the chef class and the classification for them was the waiter class.

Secondly, we observe that there is no such a big difference (8% difference in accuracy) in the predictions of our classification model with respect to the gender for the waiter class in the test set. Now, in order to have a better overview, we proceed with the evaluation of our model in both the validation and test set by taking the average of these parts. It can be delineated in Table 5.12 the prediction performance of our classification model per gender (male/female) on validation and test set for the chef and waiter class.

Chef Class	**Waiter Class**
Male Accuracy: **86.3%**	**Male** Accuracy: **85.4%**
Female Accuracy: **58.9%**	**Female** Accuracy: **85%**

Table 5.12: Prediction performance per gender (male/female) on validation and test set for the chef and waiter class

According to Table 5.12 we can stress two important points (similar to Tables 5.10-5.11. Firstly and most importantly, we see again that there is a huge gender bias in the predictions in chef class (27.4% difference in accuracy). Therefore, the value of the statistical parity is 27.4, which is such a large value. Hence, we can infer again that there is significant discrimination in the predictions of our classification model with respect to the gender for the chef class in the validation and test set.

Secondly, we observe that there is a negligible difference (0.4% difference in accuracy) in the predictions of our classification model with respect to the gender for the waiter class in the validation and test set. Thereafter, the conclusion that we draw for the first step of our methodology (bias detection step) for the chef-waiter dataset is that there is a huge gender bias in the predictions in chef class, even though there is an equal distribution among the chefs with respect to the gender in the training set.

Now and before applying the second and third step of our methodology, we present a similar evaluation of the first step of our methodology (bias detection step) of the engineer/farmer dataset. It can be delineated in

Table 5.13 the prediction performance of our classification model per class (engineer/farmer) on validation and test set.

Validation Set	**Test Set**
Engineer Accuracy: 91.3%	Engineer Accuracy: 93%
Farmer Accuracy: 91.3%	Farmer Accuracy: 99%

Table 5.13: Prediction performance per class on validation and test set (engineer/farmer dataset)

According to Table 5.13 we can stress three important points. Firstly, we see that our model in general achieves a very good performance. Overall, it has an accuracy of at least 91.3% for each class on both validation and test set. However, we can say here that based on the fact that we have only two classes, this performance is not perfect and can be further improved.

Secondly, we observe that there is a quite large difference in the prediction performance in test set for the class of engineer. Particularly, our model has better performance on farmer class in comparison to engineer class in the test set.

Finally, based on this Table 5.13, we cannot understand how our model performs across the two genders (male and female), that is the goal of this section to observe whether there is discrimination in the predictions of our classification model with respect to the gender. We also conjecture, that this difference between the engineer and farmer class in test set may be due to difference in performance across the gender.

Therefore and similar to the previous datasets (doctor/nurse and chef/waiter), we break down the performance of the classification model per gender in order to verify our hypothesis. It can be depicted in Table 5.14 the prediction performance of our classification model per gender (male/female) on validation set for the engineer and farmer class.

Engineer Class	**Farmer Class**
Male Accuracy: 95%	Male Accuracy: 90%
Female Accuracy: 87.5%	Female Accuracy: 92.5%

Table 5.14: Prediction performance per gender (male/female) on validation set for the engineer and farmer class

According to Table 5.14 we can stress two important points. Firstly and most importantly, we see that there is a large gender bias in the predictions in engineer class (7.5% difference in accuracy). Thus, the value of the statistical parity here is 7.5. Based on that, we can infer that there is discrimination in the predictions of our classification model with respect to the gender for the engineer class in the validation set. Thus, there are a lot of female images that belong to the engineer class and the classification for them was the farmer class.

Secondly, we observe that there is a minor difference (2.5% difference in accuracy) in the predictions of our classification model with respect to the gender for the farmer class in the validation set. Now we proceed with the evaluation of our model in the test set. It can be seen in Table 5.15 the prediction performance of our classification model per gender (male/female) on test set for the engineer and farmer class.

Engineer Class	**Farmer Class**
Male Accuracy: 96%	Male Accuracy: 99%
Female Accuracy: 90%	Female Accuracy: 100%

Table 5.15: Prediction performance per gender (male/female) on test set for the engineer and farmer class

According to Table 5.15 we can stress two important points. Firstly and most importantly, we see again that there is a large gender bias in the predictions in engineer class (6% difference in accuracy). Therefore, the value of the statistical parity is 6. Based on that, we can infer that there is discrimination in the predictions of our classification model with respect to the gender for the engineer class in the test set. Thus, again like in the validation set, there are a lot of female images that belong to the engineer class and the classification for them was the farmer class.

Secondly, we observe that there is only a minor difference (1% difference in accuracy) in the predictions of our classification model with respect to the gender for the farmer class in the test set. Now, in order to have a better overview, we proceed with the evaluation of our model in both the validation and test set by taking

the average of these parts. It can be delineated in Table 5.16 the prediction performance of our classification model per gender (male/female) on validation and test set for the engineer and farmer class.

Engineer Class	**Farmer Class**
Male Accuracy: **95.5%**	**Male** Accuracy: **94.4%**
Female Accuracy: **88.8%**	**Female** Accuracy: **96.6%**

Table 5.16: Prediction performance per gender (male/female) on validation and test set for the engineer and farmer class

According to Table 5.16 we can stress two important points (similar to Tables 5.14-5.15. Firstly and most importantly, we see again that there is a large gender bias in the predictions in engineer class (6.7% difference in accuracy). Therefore, the value of the statistical parity is 6.7, which is quite a large value. Hence, we can infer again that there is significant discrimination in the predictions of our classification model with respect to the gender for the engineer class in the validation and test set.

Secondly, we observe that there is a minor difference (2.2% difference in accuracy) in the predictions of our classification model with respect to the gender for the farmer class in the validation and test set. Thereafter, the conclusion that we draw for the first step of our methodology (bias detection step) for the engineer-farmer dataset is that there is a large gender bias in the predictions in engineer class, even though there is an equal distribution among the engineers with respect to the gender in the training set.

In the next sections of this Chapter, and more specifically in the second step of our methodology (bias semantic interpretation step), we want to verify our hypothesis that is stated in 1.2 that this discrimination comes from the presence of semantically meaningful visual clues that appear in the image data and give away the gender in a way that introduce bias on them. Finally, in the third and last step of our methodology (bias mitigation step), we want to compensate for this gender bias by obfuscating these visual clues.

5.2.3. CONCLUSIONS

In this section, we evaluated the first step of our methodology, the bias detection step. We presented the experiments that we performed for our use-case of profession prediction from images (doctor/nurse, chef/waiter and engineer/farmer) in order to observe whether there is discrimination in the predictions of the Machine Learning classification model with respect to the gender. We were able to draw three important conclusions.

Firstly and most importantly, we saw that there was a huge gender bias in the predictions in doctor class (doctor/nurse dataset) with a 14.8% difference in accuracy, in chef class (chef/waiter dataset) with a 27.4% difference in accuracy and in engineer class (engineer/farmer dataset) with a 6.7% difference in accuracy. Hence, we inferred that there was a significant discrimination in the predictions of our classification model with respect to the gender for the doctor, chef and the engineer class in the validation and test sets, even though there was an equal distribution among the doctors, chefs and engineers with respect to the gender in the training set.

Secondly, we observed that there was a negligible difference (0.8% difference in accuracy), (0.4% difference in accuracy) and (2.2% difference in accuracy) in the predictions of our classification model with respect to the gender for the nurse, waiter and farmer class in the validation and test sets respectively.

Finally, we saw that our classification model achieved a good but not perfect performance in terms of the accuracy based on the fact that we have only two classes (at least 80% for doctor/nurse dataset, at least 70% for chef/waiter dataset and at least 91.3% for engineer/farmer dataset).

In the next sections of this Chapter, and more specifically in the second (bias semantic interpretation step) and third (bias mitigation step) steps of our methodology, we try to do three important modifications. Firstly, we want to understand the reason of this discrimination in the predictions of our classification model with respect to the gender through semantically describe it. Secondly, we try to compensate for this gender bias. Finally, we aim in increasing the general performance (accuracy) of our classification model among all the classes doctor, nurse, chef, waiter, engineer and farmer for both genders.

5.3. EVALUATION OF THE SEMANTIC INTERPRETATION AND MITIGATION OF BIAS STEP (APPROACH 1, CORRELATION ATTENTION MECHANISM-OBJECT DETECTION)

As we mentioned in 4.1, the methodology that we propose consists of three main steps. In this section, we provide the evaluation of the second (bias semantic interpretation step) and third (bias mitigation step) steps

of our methodology (approach 1, correlation attention mechanism-object detection) building on the results which we found in 5.2. More specifically, in the second one (bias semantic interpretation step) we want to verify our hypothesis that is stated in 1.2, that the discrimination in the predictions of our classification model with respect to the gender comes from the presence of semantically meaningful visual clues that appear in the image data, which give away the gender in a way that introduce bias on them. Finally, in the third one (bias mitigation step), we want to compensate for this gender bias by obfuscating these visual clues. The evaluation of these steps is based on the pipelines of the Figures 4.10, 4.11, 4.17 (step two) and 4.19 (step three). We start this section by providing the implementation details of the evaluation of these two steps. After that, the results are given and we end up with a discussion of these results and with important conclusions that are drawn.

5.3.1. IMPLEMENTATION DETAILS

In this section, we provide the implementation details of the evaluation of the bias semantic interpretation and bias mitigation steps of our methodology (approach 1, correlation attention mechanism-object detection). Our implementation follows the order of the pipelines 4.10, 4.11, 4.17 (step two) and 4.19 (step three).

Firstly, we start with the implementation details of the pipeline regarding the Figure 4.10. Here, we start with the miss-classified data of our classification model that we ended up in the previous section 5.2. As a next step, we pass these data to an attention mechanism in order to end up with features (blob of pixels) in images that matter towards the classification outcome. As we stated in 4.4.2, we use in our experiments as an attention mechanism the SMOOTH-GRAD method [66].

There are two hyperparameters here that we have to choose a value for them: σ, the noise level or standard deviation of the Gaussian perturbations (noise that we add in the input image) and n, which corresponds to the number of samples to average over. Following the same procedure with the authors in [66] we ended up with the values 10 and 100 respectively.

The next boxes in the pipeline of Figure 4.10, namely: gray-scale, thresholding, de-noising, mask and bounding boxes (AM) as mentioned in 4.4.2 are the five respective steps of the technique that we described of detecting multiple brightest spots in the image through applying the Connected Component analysis method [86] and [87]. Based on experimentations, we found that the optimal values were: For the first step of this method (smoothing), a Gaussian filter with radius of 11. For the second step (thresholding), a threshold value of 5×10^{-4}. Thus any pixel value $p >= thresh$ is set to white and pixel values $p < thresh$ are set to black.

For the third step (de-noising), the value of iterations of erosions and dilations was set to 4. For the fourth step (mask), we wanted to define a value for the number of pixels in each component which is considered to be sufficiently large in order to add this component to the mask of "large blobs". The value that was chosen here was 200. Finally, in the fifth step (bounding boxes (AM)) we did not have to define any value, as the procedure of obtaining the contours based on the masks is quite straightforward.

Now, we continue with the implementation details of the pipeline regarding the Figure 4.11. Here, we start again with the miss-classified data of our classification model that we ended up in the previous section 5.2. In our first approach (will be discussed in this section), we pass these data to an object detection algorithm. In our second approach (will be discussed in the next section), we pass these data to a crowdsourcing task. Regarding the first approach (object detection), as we mentioned in 4.2.2, we decided to use YOLO (version 3) [85] as our object detection algorithm.

Therefore we employed a Pre-trained YOLO Detector. It it is worth mentioning here that we made a batch of experiments that we performed through using weights from pre-training on four different datasets, namely: ImageNet, COCO, Pascal VOC, Open Images dataset and combination of these four. The best results (will be presented in the next section) were obtained from weights coming from COCO dataset that has 80 different classes of objects. Moreover, we should mention here that we used as a threshold of confidence (we detect and display objects that have confidence equal or larger to the value of this threshold of confidence) the value of 0.25 as suggested by the authors of this algorithm [85].

Regarding the pipeline in Figure 4.17, as far as its implementation details are concerned, the only thing that we have to mention as we also stated in 4.4.5, is that we adopt the value of 0.5 for the Intersection over Union (IoU) score as it is also considered a good score for the evaluation of the object detectors. Thus, based on the output (bounding boxes) of the pipelines in Figures 4.10 and 4.11, we end up with a list of objects that matter for the classification for which the Intersection over Union (IoU) score of their corresponding bounding boxes is at least 0.5.

Finally, regarding the pipeline in Figure 4.19, we start with this list of objects that matter for the classification that was the output of the pipeline in Figure 4.17. After that, we obfuscate these objects that belong to this list, and we pass these new image data (same as the initial ones, but their objects that matter for the

classification are obfuscated) to the same Machine Learning classification model that was used in pipeline in Figure 4.4. Finally, we compute the evaluation metrics that we adopt (accuracy and statistical parity) and compare the initial results that we found in 5.2.2 with the results after applying this methodology. We should also stress here, that for this obfuscation task, we ended up (after experimentation) with a Gaussian filter of radius with a value of 20 in order to blur these objects that matter towards the classification outcome.

Consequently, the goal of this part is two-fold: Firstly, to verify our hypothesis that is stated in 1.2, that the discrimination in the predictions of our classification model with respect to the gender comes from the presence of semantically meaningful visual clues that appear in the image data and which give away the gender in a way that introduce bias on them. Secondly, to compensate for this gender bias by obfuscating these visual clues. In the next section, we present the results of our methodology (first approach using object detection) on our datasets (doctor/nurse, chef/waiter and engineer/farmer), followed by a discussion.

5.3.2. RESULTS AND DISCUSSION

In this section, we provide the results followed by a discussion that are related with the bias semantic interpretation and mitigation steps. As we stated in 5.3.1, and according to our pipeline in Figures 4.17 and 4.19 for these steps, we want to end up with a list of objects that matter towards the classification outcome and to compare the initial results that we found in 5.2.2 with the results after evaluating our first approach that we described in 5.3.1.

Hence, our goal in this section is two-fold: Firstly, to semantically describe the reason that a particular prediction of a Machine Learning classification model is made. Secondly, to compensate for gender bias that is related with the content of the image data. We are now ready to present these results on our datasets (doctor/nurse, chef/waiter and engineer/farmer). Firstly, we start the with the evaluation of the bias semantic interpretation step in the doctor/nurse dataset. It can be delineated in Figure 5.5 the number of occurrences of objects that matter towards classification for the missclassified images of the doctor class.

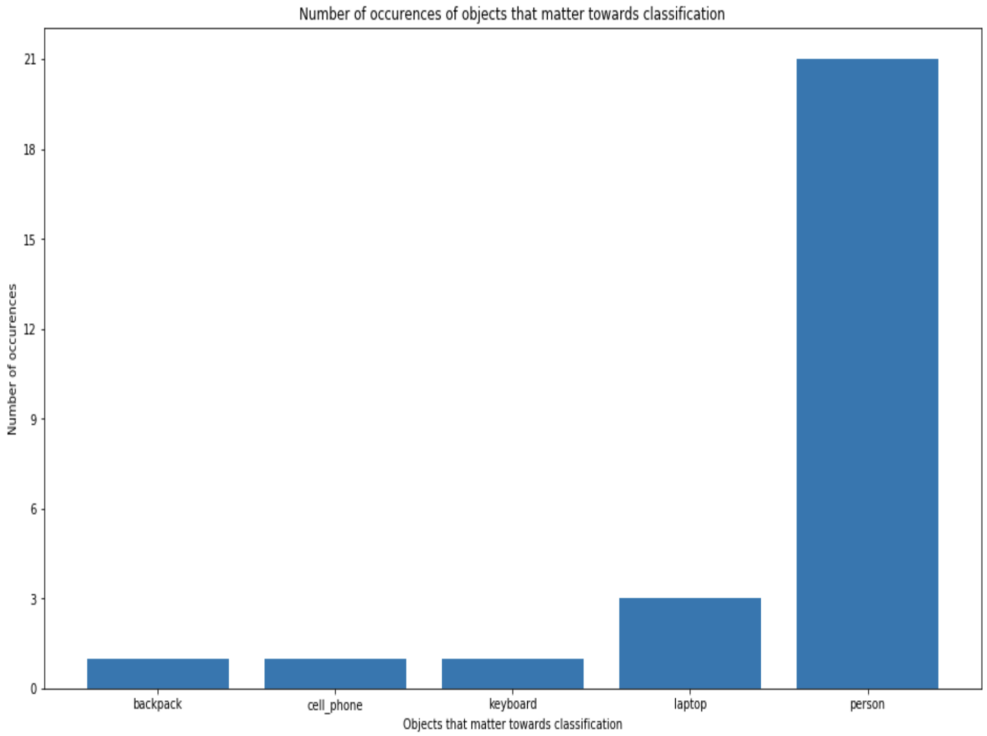


Figure 5.5: Number of occurrences of objects that matter towards classification for the doctor class

Based on Figure 5.5, we can observe that the primary reason of a classification of a person as a doctor is the presence of the face or body (class person). Therefore, the model does not look at all in the presence of a stethoscope but it looks in the face or the body (class person). Thus, the model learns to act in a gender discriminative way. Finally, it looks also in objects like laptop, keyboard, cell phone and backpack which may give a clue that the person in the image is a doctor (e.g. in a office setup), but this is not a clear indication of

that.

It can be seen in Figure 5.6 the number of occurrences of objects that matter towards classification for the missclassified images of the nurse class.

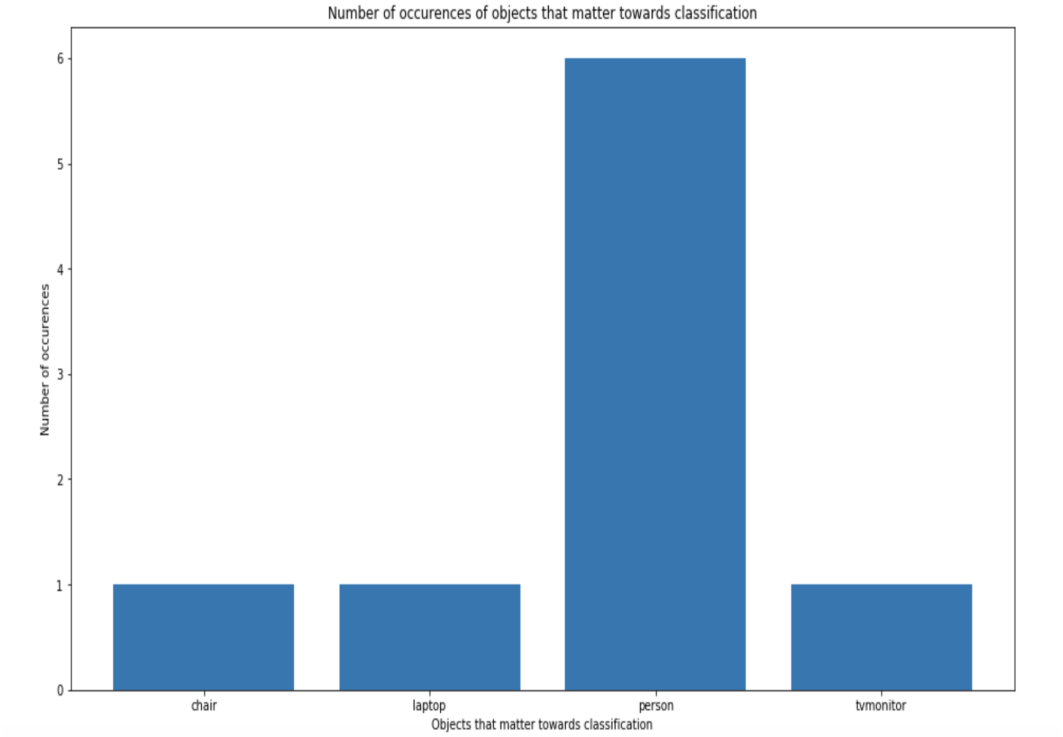


Figure 5.6: Number of occurrences of objects that matter towards classification for the nurse class

Based on Figure 5.6, we can observe that again the primary reason of a classification of a person as a nurse is the presence of the face or body (class person). However we see now that our model takes into account the presence of a person only in 6 images and not in 21 like in the doctor class. Therefore, the model does not act in a gender-bias way as in the doctor class, something that it was also verified in Table 5.8, where we saw that there was no gender bias in the predictions of the model in the nurse class. Finally, it looks also in objects like laptop, chair, and TV monitor which may give a clue that the person in the image is a nurse (e.g. in a office setup), but this is again not a clear indication of that.

Now, we present the evaluation of the bias semantic interpretation step in the chef/waiter dataset. It can be depicted in Figures 5.7 and 5.8 the number of occurrences of objects that matter towards classification for the missclassified images of the chef class.

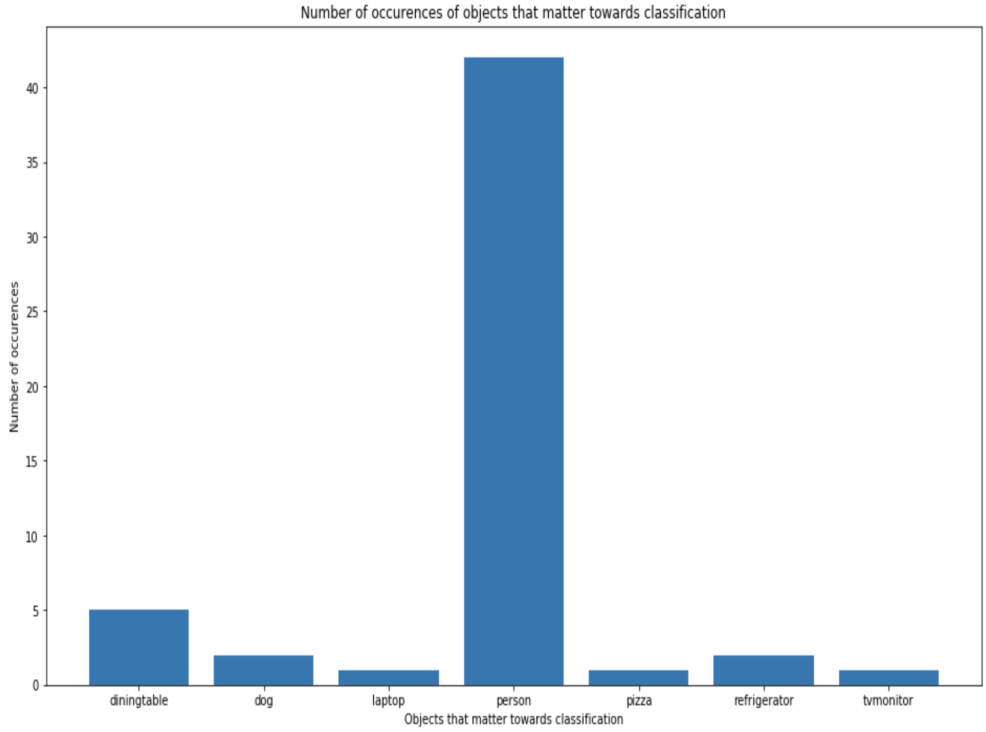


Figure 5.7: Number of occurrences of objects that matter towards classification for the chef class (Part1)

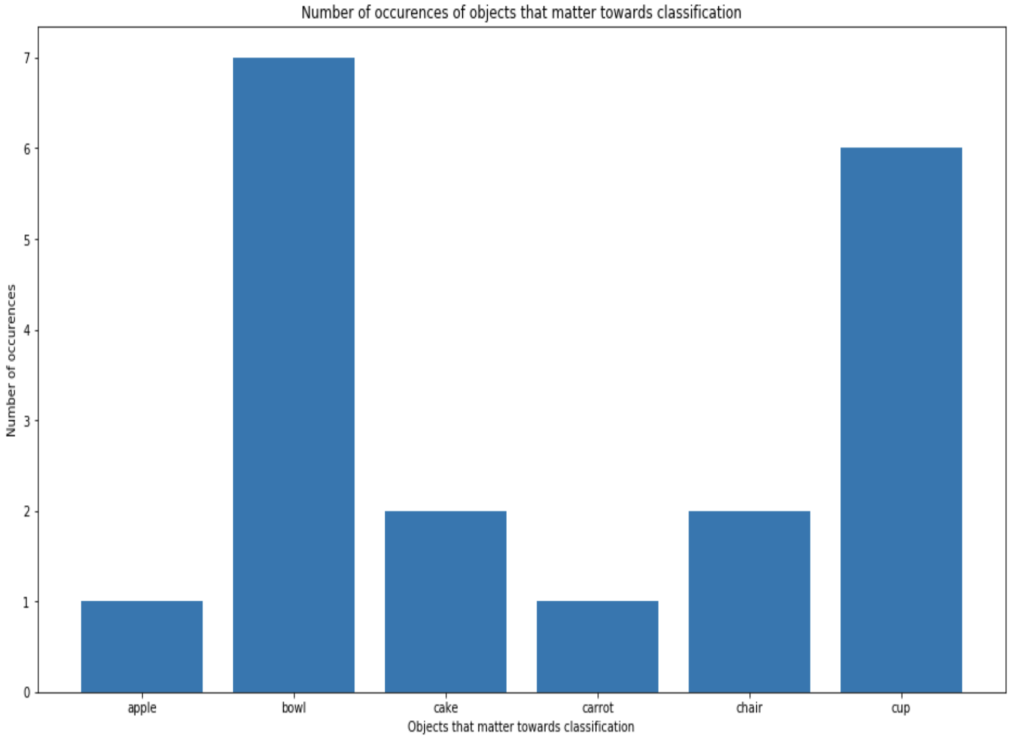


Figure 5.8: Number of occurrences of objects that matter towards classification for the chef class (Part2)

Based on Figures 5.7 and 5.8, we can observe that the primary reason of a classification of a person as a chef is the presence of the face or body (class person). Therefore, the model does not look at all in the presence of a uniform but it looks in the face or the body (class person). Thus, the model learns to act in a gender

discriminative way like in the doctor class. Finally, it looks also in objects (among others) like bowl, cup, carrot, cake, apple and dining table, pizza and refrigerator which give a clue that the person in the image is a chef (e.g. in a restaurant setup), but this is not a clear indication of that and it happens only occasionally.

It can be seen in Figure 5.9 the number of occurrences of objects that matter towards classification for the missclassified images of the waiter class.

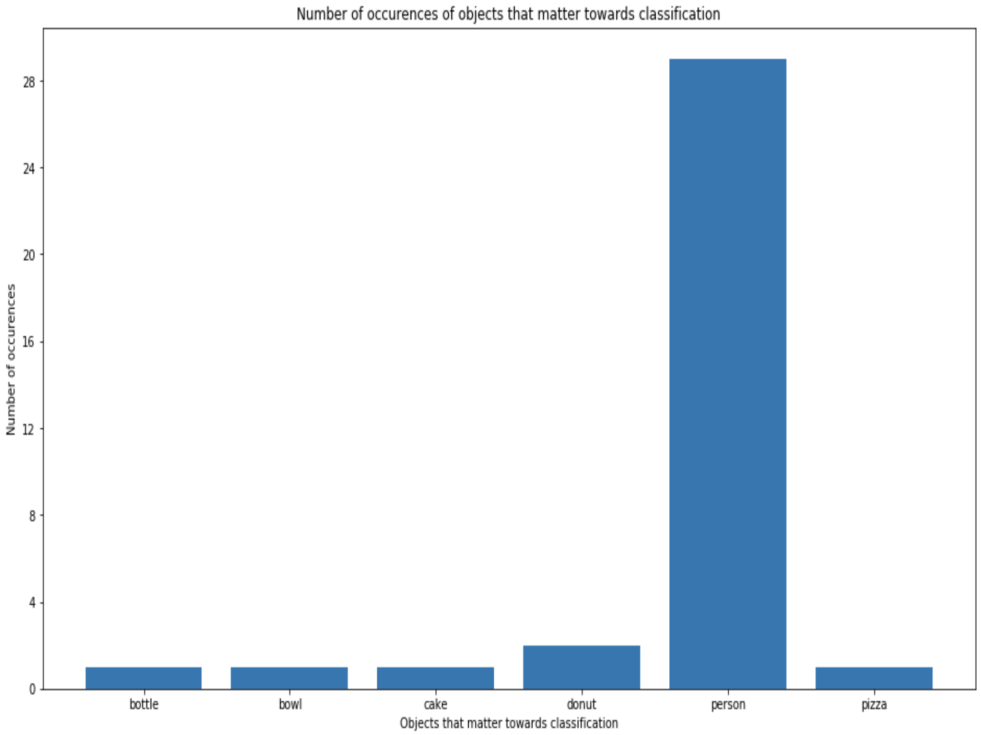


Figure 5.9: Number of occurrences of objects that matter towards classification for the waiter class

Based on Figure 5.9, we can observe that again the primary reason of a classification of a person as a waiter is the presence of the face or body (class person). However we see now that our model takes into account the presence of a person in a smaller number of images (28 in comparison to 43 in chef class). Therefore, the model does not act so much in a gender-bias way as in the chef class, something that it was also verified in Table 5.12, where we saw that there was no gender bias in the predictions of the model in the waiter class. Finally, it looks also in objects like bottle, bowl, cake, donut and pizza which may give a clue that the person in the image is a waiter (e.g. in a restaurant setup), but this is again not a clear indication of that.

Finally, we present the evaluation of the bias semantic interpretation step in the engineer/farmer dataset. It can be depicted in Figure 5.10 the number of occurrences of objects that matter towards classification for the missclassified images of the engineer class.

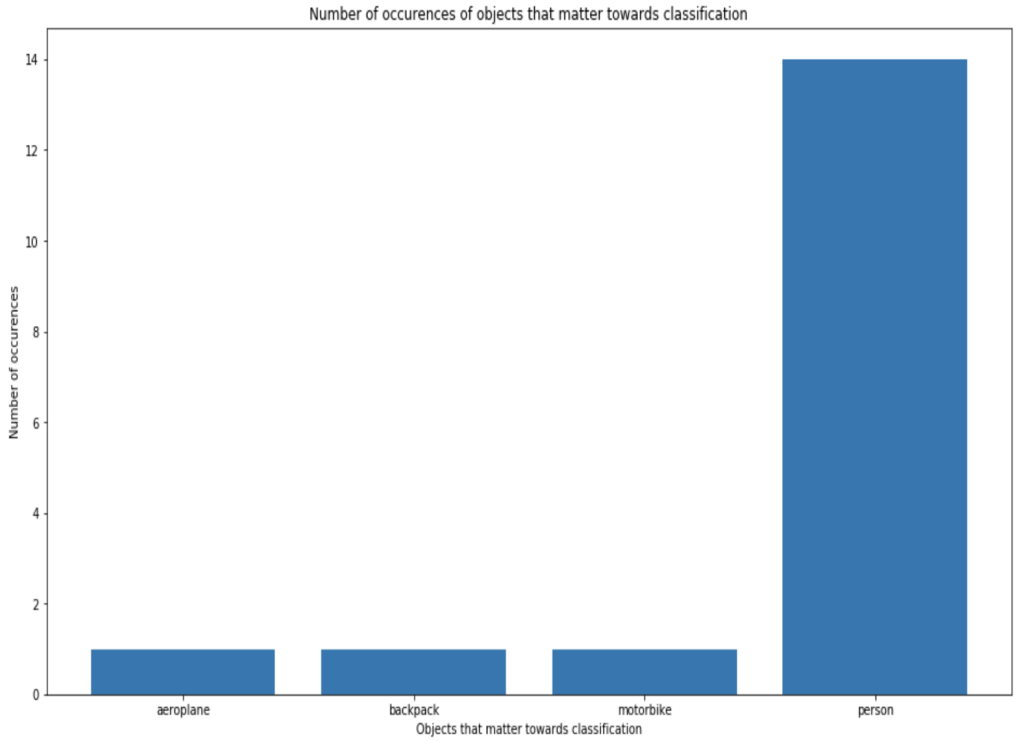


Figure 5.10: Number of occurrences of objects that matter towards classification for the engineer class

Based on Figure 5.10 , we can observe that the primary reason of a classification of a person as a engineer is the presence of the face or body (class person). Therefore, the model does not look at all in the presence of something that is related with the profession of an engineer (e.g. a uniform) but it looks in the face or the body (class person). Thus, the model learns to act in a gender discriminative way like in the doctor and chef class. Finally, it looks also in objects like airplane, backpack and motorbike which give a clue that the person in the image is a engineer (e.g. in a construction site), but this is not a clear indication of that and it happens only occasionally.

It can be seen in Figure 5.11 the number of occurrences of objects that matter towards classification for the missclassified images of the farmer class.

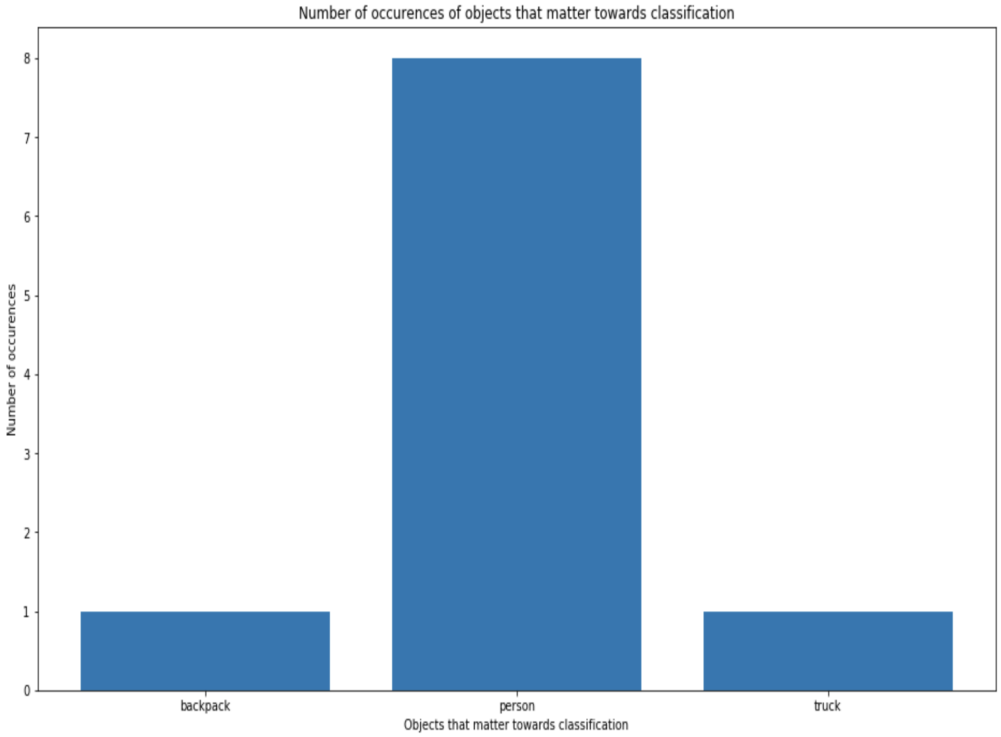


Figure 5.11: Number of occurrences of objects that matter towards classification for the farmer class

Based on Figure 5.11, we can observe that again the primary reason of a classification of a person as a farmer is the presence of the face or body (class person). However we see now that our model takes into account the presence of a person in a smaller number of images (8 in comparison to 14 in engineer class). Therefore, the model does not act so much in a gender-bias way as in the engineer class, something that it was also verified in Table 5.16, where we saw that there was no gender bias in the predictions of the model in the farmer class. Finally, it looks also in objects like backpack and truck which may give a clue that the person in the image is a farmer (e.g. in a farm setup), but this is again not a clear indication of that.

Now, we continue with the evaluation of the third and last step of our methodology, the bias mitigation step in the doctor/nurse dataset. It can be delineated in Table 5.17 the prediction performance of our classification model per class (doctor/nurse) on validation and test set after applying our methodology (approach 1).

Validation Set	**Test Set**
Doctor Accuracy: 88.8%	Doctor Accuracy: 88%
Nurse Accuracy: 91.1%	Nurse Accuracy: 83%

Table 5.17: Prediction performance per class on validation and test set (doctor-nurse dataset) (approach 1)

According to Table 5.17 we can stress two important points. Firstly, we see that our model now achieves a better performance. Overall, it has an accuracy now in the doctor class of at least 8% more than before applying our methodology on both validation and test set. Also, it has an accuracy in the nurse class of at least 2% more than before applying our methodology on both validation and test set.

Secondly, we observe that there is a very small difference now in the prediction performance in validation set for the class of doctor and nurse. Particularly, this difference is only 2.3%, in comparison to 8.8% that was before.

It can be depicted in Table 5.18 the prediction performance of our classification model after applying our methodology (approach 1) per gender (male/female) on validation set for the doctor and nurse class.

Doctor Class	**Nurse Class**
Male Accuracy: 93.3%	Male Accuracy: 88.8%
Female Accuracy: 84.4%	Female Accuracy: 93.3%

Table 5.18: Prediction performance per gender (male/female) on validation set for the doctor and nurse class (approach 1)

According to Table 5.18 we can stress three important points. Firstly and most importantly, we see that there is a significant decrease in gender bias in the predictions in doctor class (8.9% difference in accuracy, in comparison to 13.3% that it was before). Based on that, we can infer that there is smaller discrimination in the predictions of our classification model with respect to the gender for the doctor class in the validation set after applying our method.

Secondly, we observe that still there is no big difference (4.5% difference in accuracy) in the predictions of our classification model with respect to the gender for the nurse class in the validation set. Finally, we can also mention that there is a significant increase in accuracy of at least 2.2% for both classes and for both male and female.

It can be seen in Table 5.19 the prediction performance of our classification model after applying our methodology (approach 1) per gender (male/female) on test set for the doctor and nurse class.

Doctor Class	**Nurse Class**
Male Accuracy: 92%	Male Accuracy: 86%
Female Accuracy: 84%	Female Accuracy: 80%

Table 5.19: Prediction performance per gender (male/female) on test set for the doctor and nurse class (approach 1)

According to Table 5.19 we can stress three important points. Firstly and most importantly, we see that there is a significant decrease in gender bias in the predictions in doctor class (8% difference in accuracy, in comparison to 16% that it was before). Based on that, we can infer that there is smaller discrimination in the predictions of our classification model with respect to the gender for the doctor class in the test set after applying our method.

Secondly, we observe that still there is no big difference (6% difference in accuracy) in the predictions of our classification model with respect to the gender for the nurse class in the test set. Finally, we can also mention that there is a significant increase in accuracy of at least 2% for both classes and for both male and female.

Finally, it can be shown in Table 5.20 the prediction performance of our classification model after applying our methodology (approach 1) per gender (male/female) on validation and test set for the doctor and nurse class.

Doctor Class	**Nurse Class**
Male Accuracy: **92.7%**	**Male** Accuracy: **87.3%**
Female Accuracy: **84.2%**	**Female** Accuracy: **86.7%**

Table 5.20: Prediction performance per gender (male/female) on validation and test set for the doctor and nurse class (approach 1)

According to Table 5.20 we can stress three important points (similar to Tables 5.18-5.19. Firstly and most importantly, we see that there is a significant decrease in gender bias in the predictions in doctor class (8.5% difference in accuracy, in comparison to 14.8% that it was before). Based on that, we can infer that there is smaller discrimination in the predictions of our classification model with respect to the gender for the doctor class in both validation and test set after applying our method.

Secondly, we observe that still there is almost no difference (0.6% difference in accuracy) in the predictions of our classification model with respect to the gender for the nurse class in both validation and test set. Finally, we can also mention that there is a significant increase in accuracy of at least 2% for the nurse class for both male and female and at least 5.3% for the doctor class for both genders.

Thereafter, the conclusion that we draw after applying our methodology (approach 1) for the doctor-nurse dataset is that there is a significant mitigation of gender bias in the predictions in doctor class and a significant increase in accuracy in both classes for both male and female.

Now we present a similar analysis of the results after applying our methodology (approach 1) for the

chef/waiter dataset. It can be delineated in Table 5.21 the prediction performance of our classification model after applying our methodology (approach 1) per class (chef/waiter) on validation and test set.

Validation Set	**Test Set**
Chef Accuracy: 85.6%	Chef Accuracy: 82%
Waiter Accuracy: 88.9%	Waiter Accuracy: 90%

Table 5.21: Prediction performance per class on validation and test set (chef-waiter dataset) (approach 1)

According to Table 5.21 we can stress two important points. Firstly, we see that our model now achieves a better performance. Overall, it has an accuracy now in the chef class of at least 11% more than before applying our methodology on both validation and test set. Also, it has an accuracy in the waiter class of at least 4% more than before applying our methodology on both validation and test set.

Secondly, we observe that there is a very small difference now in the prediction performance in validation and test set for the class of chef and waiter. Particularly, this difference is only 3.3%, in comparison to 10% that was before for the validation set and 8%, in comparison to 15% for the test set that was before.

It can be depicted in Table 5.22 the prediction performance of our classification model after applying our methodology (approach 1) per gender (male/female) on validation set for the chef and waiter class.

Chef Class	**Waiter Class**
Male Accuracy: 93.3%	Male Accuracy: 93.3%
Female Accuracy: 77.8%	Female Accuracy: 84.4%

Table 5.22: Prediction performance per gender (male/female) on validation set for the chef and waiter class (approach 1)

According to Figure Table 5.22 we can stress three important points. Firstly and most importantly, we see that there is a significant decrease in gender bias in the predictions in chef class (15.5% difference in accuracy, in comparison to 28.8% that it was before). Based on that, we can infer that there is smaller discrimination in the predictions of our classification model with respect to the gender for the chef class in the validation set after applying our method.

Secondly, we observe that still there is no big difference (8.9% difference in accuracy) in the predictions of our classification model with respect to the gender for the waiter class in the validation set. Finally, we can also mention that there is a significant increase in accuracy of at least 4.4% for both classes and for both male and female.

It can be seen in Table 5.23 the prediction performance of our classification model after applying our methodology (approach 1) per gender (male/female) on test set for the chef and waiter class.

Chef Class	**Waiter Class**
Male Accuracy: 90%	Male Accuracy: 86%
Female Accuracy: 74%	Female Accuracy: 94%

Table 5.23: Prediction performance per gender (male/female) on test set for the chef and waiter class (approach 1)

According to Table 5.23 we can stress three important points. Firstly and most importantly, we see that there is a significant decrease in gender bias in the predictions in chef class (16% difference in accuracy, in comparison to 26% that it was before). Based on that, we can infer that there is smaller discrimination in the predictions of our classification model with respect to the gender for the chef class in the test set after applying our method.

Secondly, we observe that still there is no big difference (8% difference in accuracy) in the predictions of our classification model with respect to the gender for the waiter class in the test set. Finally, we can also mention that there is a significant increase in accuracy of at least 4% for both classes and for both male and female.

Finally, it can be shown in Table 5.24 the prediction performance of our classification model after applying our methodology (approach 1) per gender (male/female) on validation and test set for the chef and waiter class.

Chef Class	**Waiter Class**
Male Accuracy: **91.6%**	**Male** Accuracy: **89.5%**
Female Accuracy: **75.8%**	**Female** Accuracy: **89.3%**

Table 5.24: Prediction performance per gender (male/female) on validation and test set for the chef and waiter class (approach 1)

According to Table 5.24 we can stress three important points (similar to Tables 5.22-5.23. Firstly and most importantly, we see that there is a significant decrease in gender bias in the predictions in chef class (15.8% difference in accuracy, in comparison to 27.4% that it was before). Based on that, we can infer that there is smaller discrimination in the predictions of our classification model with respect to the gender for the chef class in both validation and test set after applying our method.

Secondly, we observe that still there is almost no difference (0.2% difference in accuracy) in the predictions of our classification model with respect to the gender for the waiter class in both validation and test set. Finally, we can also mention that there is a significant increase in accuracy of at least 4.1% for the waiter class for both male and female and at least 5.3% for the chef class for both genders.

Thereafter, the conclusion that we draw after applying our methodology (approach 1) for the chef-waiter dataset (and similar to the doctor/nurse dataset) is that there is a significant mitigation of gender bias in the predictions in chef class and a significant increase in accuracy in both classes for both male and female.

Finally we present a similar analysis of the results after applying our methodology (approach 1) for the engineer/farmer dataset. It can be delineated in Table 5.25 the prediction performance of our classification model after applying our methodology (approach 1) per class (engineer/farmer) on validation and test set.

Validation Set	**Test Set**
Engineer Accuracy: 93.8%	Engineer Accuracy: 95%
Farmer Accuracy: 97.5%	Farmer Accuracy: 100%

Table 5.25: Prediction performance per class on validation and test set (engineer-farmer dataset) (approach 1)

According to Table 5.25 we can stress two important points. Firstly, we see that our model now achieves a better performance. Overall, it has an accuracy now in the engineer class of at least 2% more than before applying our methodology on both validation and test set. Also, it has an accuracy in the farmer class of at least 1% more than before applying our methodology on both validation and test set.

Secondly, we observe that there is a smaller difference now in the prediction performance in validation and test set for the class of engineer and farmer with an improvement of 1%.

It can be depicted in Table 5.26 the prediction performance of our classification model after applying our methodology (approach 1) per gender (male/female) on validation set for the engineer and farmer class.

Engineer Class	**Farmer Class**
Male Accuracy: 95%	Male Accuracy: 97.5%
Female Accuracy: 92.5%	Female Accuracy: 97.5%

Table 5.26: Prediction performance per gender (male/female) on validation set for the engineer and farmer class (approach 1)

According to Table 5.26 we can stress three important points. Firstly and most importantly, we see that there is a significant decrease in gender bias in the predictions in engineer class (2.5% difference in accuracy, in comparison to 7.5% that it was before). Based on that, we can infer that there is smaller discrimination in the predictions of our classification model with respect to the gender for the engineer class in the validation set after applying our method.

Secondly, we observe that there is no difference (0% difference in accuracy) in the predictions of our classification model with respect to the gender for the farmer class in the validation set. Finally, we can also mention that there is a significant increase in accuracy of at least 5% for both classes and for both male and female (engineer class for male gender is excluded, as there is no difference).

It can be seen in Table 5.27 the prediction performance of our classification model after applying our methodology (approach 1) per gender (male/female) on test set for the engineer and farmer class.

Engineer Class	**Farmer Class**
Male Accuracy: 96%	Male Accuracy: 100%
Female Accuracy: 94%	Female Accuracy: 100%

Table 5.27: Prediction performance per gender (male/female) on test set for the engineer and farmer class (approach 1)

According to Table 5.27 we can stress three important points. Firstly and most importantly, we see that there is a significant decrease in gender bias in the predictions in engineer class (2% difference in accuracy, in comparison to 6% that it was before). Based on that, we can infer that there is smaller discrimination in the predictions of our classification model with respect to the gender for the engineer class in the test set after applying our method.

Secondly, we observe that there is no difference (0% difference in accuracy) in the predictions of our classification model with respect to the gender for the farmer class in the test set. Finally, we can also mention that there is an increase in accuracy of 4% in engineer class for female gender and 1% in farmer class for male gender.

Finally, it can be shown in Table 5.28 the prediction performance of our classification model after applying our methodology (approach 1) per gender (male/female) on validation and test set for the engineer and farmer class.

Engineer Class	**Farmer Class**
Male Accuracy: **95.5%**	**Male** Accuracy: **98.8%**
Female Accuracy: **93.3%**	**Female** Accuracy: **98.8%**

Table 5.28: Prediction performance per gender (male/female) on validation and test set for the engineer and farmer class (approach 1)

According to Table 5.28 we can stress three important points (similar to Tables 5.26-5.27). Firstly and most importantly, we see that there is a significant decrease in gender bias in the predictions in engineer class (2.2% difference in accuracy, in comparison to 6.7% that it was before). Based on that, we can infer that there is smaller discrimination in the predictions of our classification model with respect to the gender for the engineer class in both validation and test set after applying our method.

Secondly, we observe that still there is no difference (0% difference in accuracy) in the predictions of our classification model with respect to the gender for the farmer class in both validation and test set. Finally, we can also mention that there is a significant increase in accuracy of 4.4% in the engineer class for female gender and at least 2.2% in the farmer class for both genders.

Thereafter, the conclusion that we draw after applying our methodology (approach 1) for the engineer-farmer dataset (and similar to the doctor/nurse and chef/waiter datasets) is that there is a significant mitigation of gender bias in the predictions in engineer class and a significant increase in accuracy in both classes for both male and female.

Finally, we would like to stress that for all datasets, after applying our methodology (approach 1), the probabilities that are related with a specific classification outcome were improved in at least 60% of the cases. For instance, even in cases that we had still miss-predictions after applying our methodology (approach 1), there was an improvement in at least 60% of the cases in confidence that is related with that classification outcome for all datasets (e.g. we have a situation like the following in at least 60% of the miss-classified images in all datasets: we have an image with a real label of doctor, but the prediction of the model is nurse with confidence of 80% before applying our method and 65% after applying our method). Finally, for easier reference, we quote all the aforementioned comments for all datasets in Tables 5.29 and 5.30.

Statistical Parity (Before)	**Statistical Parity (After)**
Doctor: **14.8%**	Doctor: **8.5% (6.3% improvement)**
Chef: **27.4%**	Chef: **15.8%(11.6% improvement)**
Engineer: **6.7%**	Engineer: **2.2%(4.5% improvement)**

Table 5.29: Statistical parity in doctor, chef and engineer class, before and after applying our methodology (approach 1)

Class + Gender	Improvement in Accuracy (Validation and Test Set)
Doctor + Male	**+5.3%**
Doctor + Female	**+11.6%**
Nurse + Male	**+2%**
Nurse + Female	**+2.2%**
Chef + Male	**+5.3%**
Chef + Female	**+16.9%**
Waiter + Male	**+4.1%**
Waiter + Female	**+4.3%**
Engineer + Male	**+0%**
Engineer + Female	**+4.5%**
Farmer + Male	**+4.4%**
Farmer + Female	**+2.2%**

Table 5.30: Improvement in accuracy (validation and test set) for all datasets after applying our methodology (approach 1)

5.3.3. CONCLUSIONS

In this section, we evaluated the second and the third step of our methodology, the bias semantic interpretation and mitigation step (approach 1). We presented the experiments that we performed for our use-case of profession prediction from images (doctor/nurse, chef/waiter and engineer/farmer) in order to semantically describe the reason of a particular prediction of our classification model in miss-classified image data and to compensate for gender bias that is related with these predictions. We were able to draw three important conclusions.

Firstly, we were able to understand the reason of the discrimination in the predictions of our classification model with respect to the gender through semantically described it. More specifically, we shown that our classification model, in images that its prediction was wrong, was mostly looking in the presence of the face or body (person class) for the classes of doctor, chef and engineer, where the gender bias was huge. In the classes of nurse, waiter and farmer, its attention was in a lower grade in the presence of the face or body (person class) and therefore this was the reason actually that we did not have so much gender bias in these classes.

Secondly, we were able to compensate for this gender bias in the classes of doctor, chef and engineer. Particularly, through applying our bias mitigation step, we managed in decreasing the statistical parity from 14.8% to 8.5% in the doctor class, from 27.4% to 15.8% in the chef class and from 6.7% to 2.2% in the engineer class, via obfuscating the objects that introduced bias to the classification outcome (e.g person).

Finally, we achieved in increasing the general performance (accuracy) of our classification model among all the classes doctor, nurse, chef, waiter, engineer and farmer and for both genders. Meticulously, we were able to increase the accuracy of each class and for both genders with a percentage of at least 2% and at most 16.9% (engineer class for male gender is excluded, where we did notice any difference).

In the next section of this Chapter, and more specifically in the second (bias semantic interpretation step) and third (bias mitigation step) steps of our methodology (approach 2), we follow a similar procedure with this one presented in this section, but using crowdsourcing instead of object detection. Our goal is two-fold. Firstly, to observe which classes of objects that matter towards the classification outcome the crowd can suggest that an object detection algorithm cannot identify. Secondly, to understand whether the objects that people perceive that introduce bias to the classification result actually affect this outcome through comparing the final results of approach 2 with these of approach 1.

5.4. EVALUATION OF THE SEMANTIC INTERPRETATION AND MITIGATION OF BIAS STEP (APPROACH 2, CORRELATION ATTENTION MECHANISM-CROWDSOURCING)

As we mentioned in 4.1, the methodology that we propose consists of three main steps. In this section, we provide the evaluation of the second (bias semantic interpretation step) and third (bias mitigation step) steps of our methodology (approach 2, correlation attention mechanism-crowdsourcing) building on the results which we found in 5.2. More specifically and similar to the experimental evaluation that we provided in 5.3, in the second step (bias semantic interpretation step) we want to verify our hypothesis that is stated in 1.2, that the discrimination in the predictions of our classification model with respect to the gender comes from the presence of semantically meaningful visual clues that appear in the image data, which give away

the gender in a way that introduce bias on them. Finally, in the third step (bias mitigation step), we want to compensate for this gender bias by obfuscating these visual clues. The evaluation of these steps is based on the pipelines of the Figures 4.10, 4.11, 4.17 (step two) and 4.19 (step three). We start this section by providing the implementation details of the evaluation of these two steps. After that, the results are given and a comparison of this approach (2) with approach 1 is provided. Finally, we end up with a discussion of these results and with important conclusions that are drawn.

5.4.1. IMPLEMENTATION DETAILS

In this section, we provide the implementation details of the evaluation of the bias semantic interpretation and bias mitigation steps of our methodology (approach 2, correlation attention mechanism-crowdsourcing). Our implementation and similar to that in 5.3.1 follows the order of the pipelines 4.10, 4.11, 4.17 (step two) and 4.19 (step three).

The only difference here in comparison to the implementation details in 5.3.1 concerns the implementation details of the pipeline regarding the Figure 4.11. Thus, the difference in this approach (2) is that we use a crowdsourcing task instead of using an object detection algorithm. The goal of this crowdsourcing task is to ask people to draw bounding boxes around each visual clue that gives away the gender of the person of interest in the image.

Thereafter, our motivation of implementing this crowdsourcing part is actually two-fold: Firstly, to have an understanding of how the intuition of people about a potential cause of gender bias actually compares with the actual reason that affects the prediction of a Machine Learning classification model. Secondly, to gain an insight of how much the intuition of people for elements of gender bias actually matches the semantic description of the images that comes from an object detection algorithm.

In order to implement this crowdsourcing task, we designed it in the Figure Eight platform through recruiting some professional crowdworkers from USA in a way that we described in 4.4.4. More specifically, we chose each image to be annotated by 3 crowdworkers (in order to have a variety in annotations, as the property of our crowdsourcing task can be quite subjective). Also, each task contained 5 images and the total number of tasks was 40.

The rest of the implementation details are identical to those described in 5.3.1. Consequently, the goal of this part (and similar to that in 5.3) is three-fold: Firstly, to verify our hypothesis that is stated in 1.2, that the discrimination in the predictions of our classification model with respect to the gender comes from the presence of semantically meaningful visual clues that appear in the image data and which give away the gender in a way that introduce bias on them.

Secondly, to compensate for this gender bias by obfuscating these visual clues that give away the gender. Thirdly, to compare this approach (2) with approach (1) that was describing in 5.3 in terms of performance (improvement in accuracy and statistical parity) and semantic description (which objects matter towards a specific prediction of a Machine Learning classification model). In the next section, we present the results of our methodology (second approach using crowdsourcing) on our datasets (doctor/nurse, chef/waiter and engineer/farmer), followed by a discussion.

5.4.2. RESULTS AND DISCUSSION

In this section, we provide the results followed by a discussion that are related with the bias semantic interpretation and mitigation steps. As we stated in 5.4.1, and according to our pipeline in Figures 4.17 and 4.19 for these steps, we want to end up with a list of objects that matter towards the classification outcome and to compare the initial results that we found in 5.2.2 with the results after evaluating our second approach that we described in 5.4.1.

Hence, our goal in this section is two-fold: Firstly, to semantically describe the reason that a particular prediction of a Machine Learning classification model is made. Secondly, to compensate for gender bias that is related with the content of the image data. We are now ready to present these results on our datasets (doctor/nurse, chef/waiter and engineer/farmer). Firstly, we start the with the evaluation of the bias semantic interpretation step in the doctor/nurse dataset. It can be delineated in Figure 5.12 the number of occurrences of objects that matter towards classification for the missclassified images of the doctor class (approach 2).

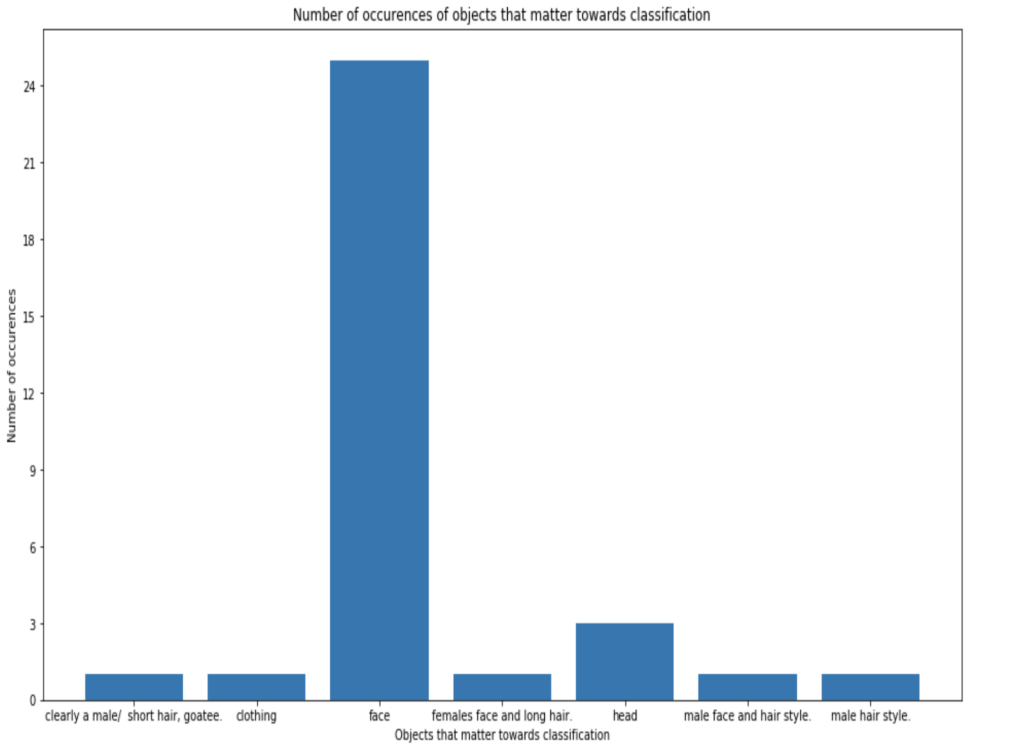


Figure 5.12: Number of occurrences of objects that matter towards classification for the doctor class (approach 2)

Based on Figure 5.12, we can observe that the primary reason of a classification of a person as a doctor is the presence of the face (and in accordance with approach 1). Thus, the model learns to act in a gender discriminative way. Finally, it looks also in visual clues like the head, the hair style and the clothing.

It can be seen in Figure 5.13 the number of occurrences of objects that matter towards classification for the missclassified images of the nurse class (approach 2).

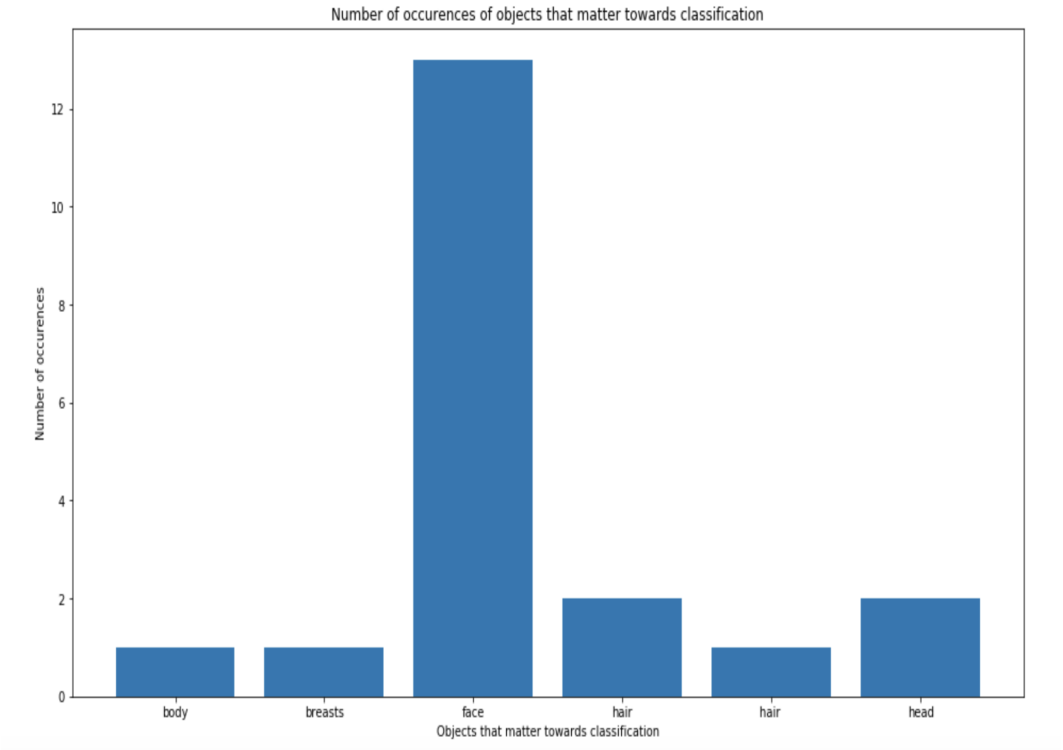


Figure 5.13: Number of occurrences of objects that matter towards classification for the nurse class (approach 2)

Based on Figure 5.13, we can observe that again the primary reason of a classification of a person as a nurse is the presence of the face. However (and in accordance with approach 1) we see now that our model takes into account the presence of a person only in 13 images and not in 26 as in the doctor class. Therefore, the model does not act in a gender-bias way as in the doctor class, something that it was also verified in Table 5.8, where we saw that there was no gender bias in the predictions of the model in the nurse class. Finally, it looks also in visual clues like the body, breast head and hair.

Now, we present the evaluation of the bias semantic interpretation step in the chef/waiter dataset. It can be depicted in Figure 5.14 the number of occurrences of objects that matter towards classification for the missclassified images of the chef class (approach 2).

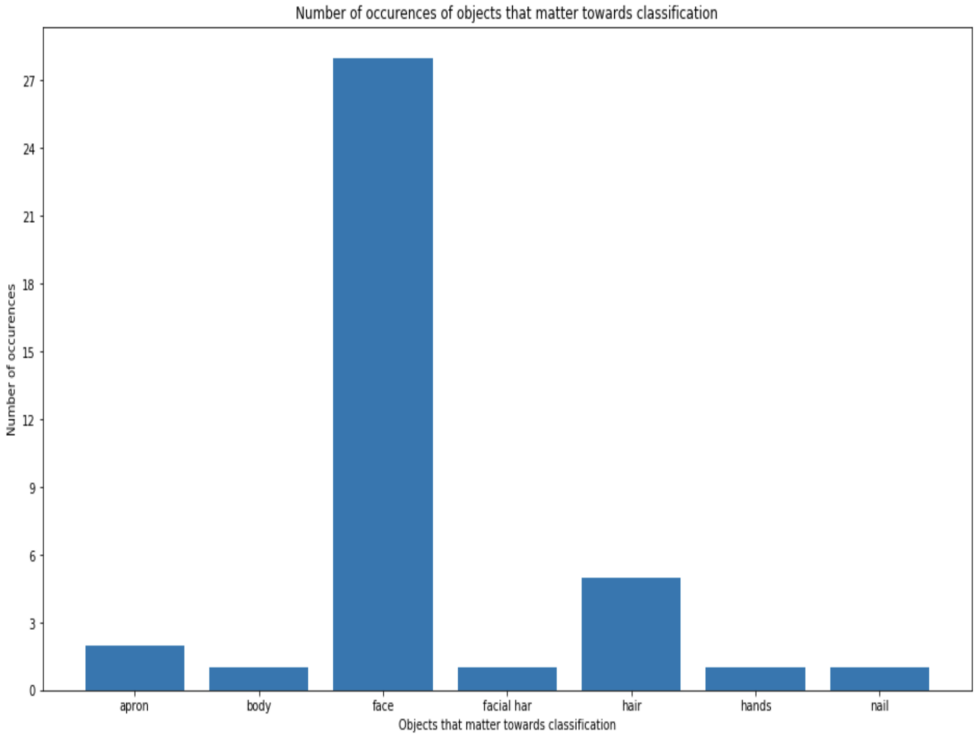


Figure 5.14: Number of occurrences of objects that matter towards classification for the chef class (approach 2)

Based on Figure 5.14, we can observe (and in accordance with approach 1) that the primary reason of a classification of a person as a chef is the presence of the face. Thus, the model learns to act in a gender discriminative way like in the doctor class. Finally, it looks also in visual clues like the apron, body, facial hair, hair, nails and hands.

It can be seen in Figure 5.15 the number of occurrences of objects that matter towards classification for the missclassified images of the waiter class (approach 2).

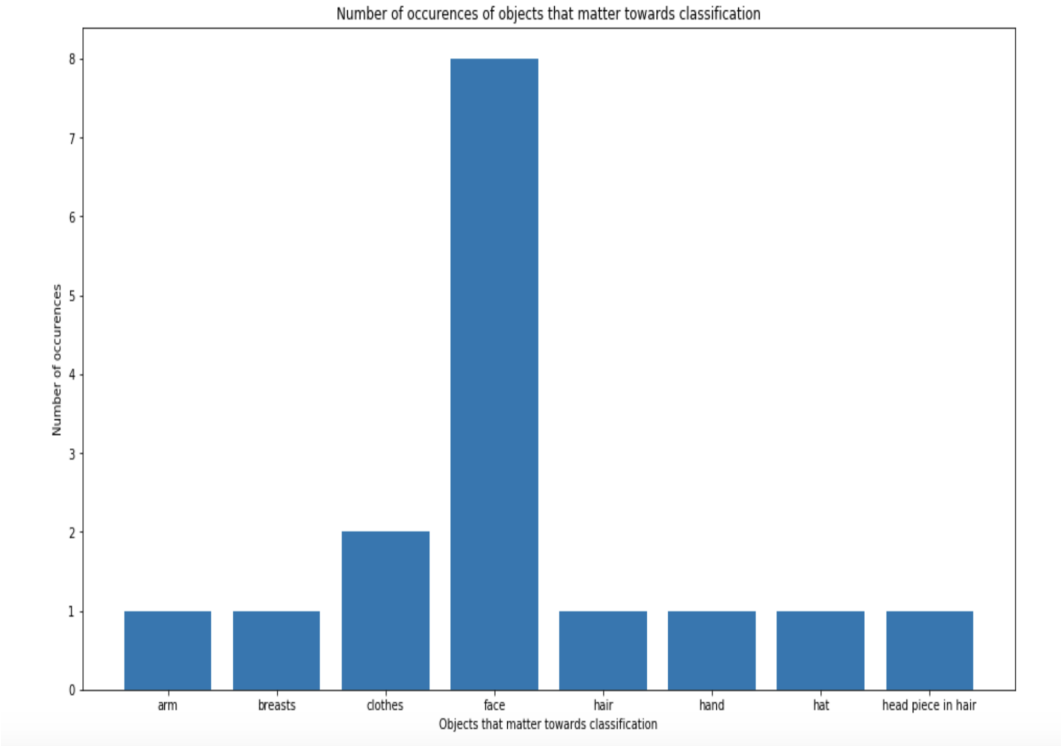


Figure 5.15: Number of occurrences of objects that matter towards classification for the waiter class (approach 2)

Based on Figure 5.15, we can observe that again the primary reason of a classification of a person as a waiter (and in accordance with approach 1) is the presence of the face. However we see now that our model takes into account the presence of a face in a smaller number of images (8 in comparison to 28 in chef class). Therefore, the model does not act so much in a gender-bias way as in the chef class, something that it was also verified in Table 5.12, where we saw that there was no gender bias in the predictions of the model in the waiter class. Finally, it looks also in visual clues like arm, breast, clothes, hair and hands.

Finally, we present the evaluation of the bias semantic interpretation step in the engineer/farmer dataset. It can be depicted in Figure 5.16 the number of occurrences of objects that matter towards classification for the missclassified images of the engineer class (approach 2).

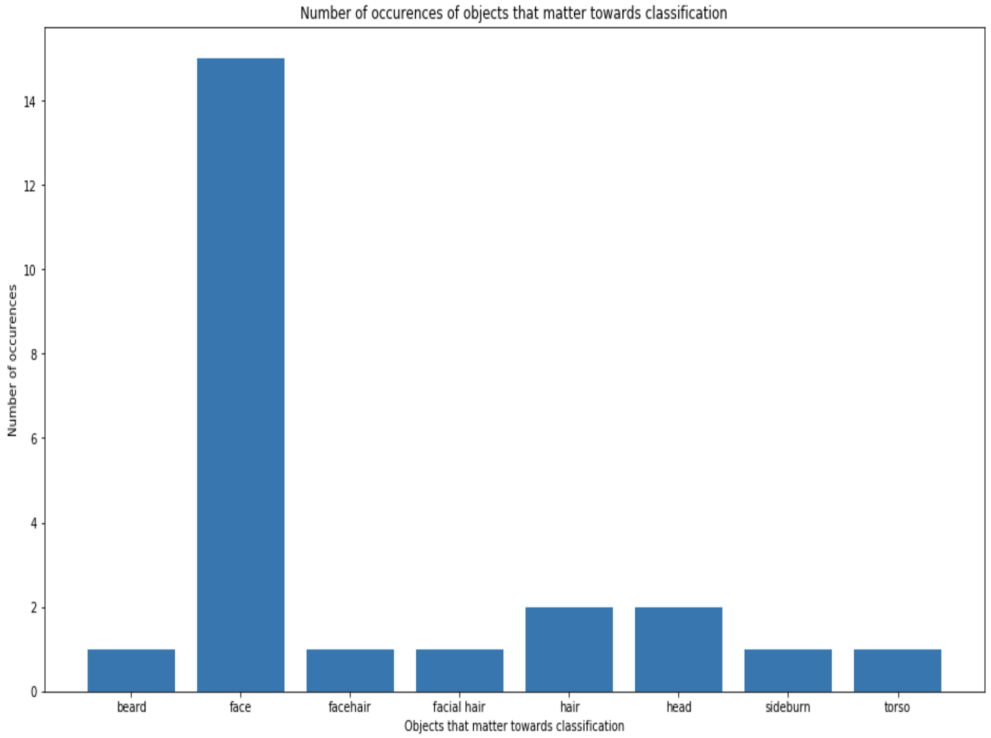


Figure 5.16: Number of occurrences of objects that matter towards classification for the engineer class (approach 2)

Based on Figure 5.16 , we can observe that the primary reason of a classification of a person as a engineer (and in accordance with approach 1) is the presence of the face. Thus, the model learns to act in a gender discriminative way like in the doctor and chef class. Finally, it looks also in visual clues like beard, facial hair, hair, head and torso.

It can be seen in Figure 5.17 the number of occurrences of objects that matter towards classification for the missclassified images of the farmer class (approach 2).

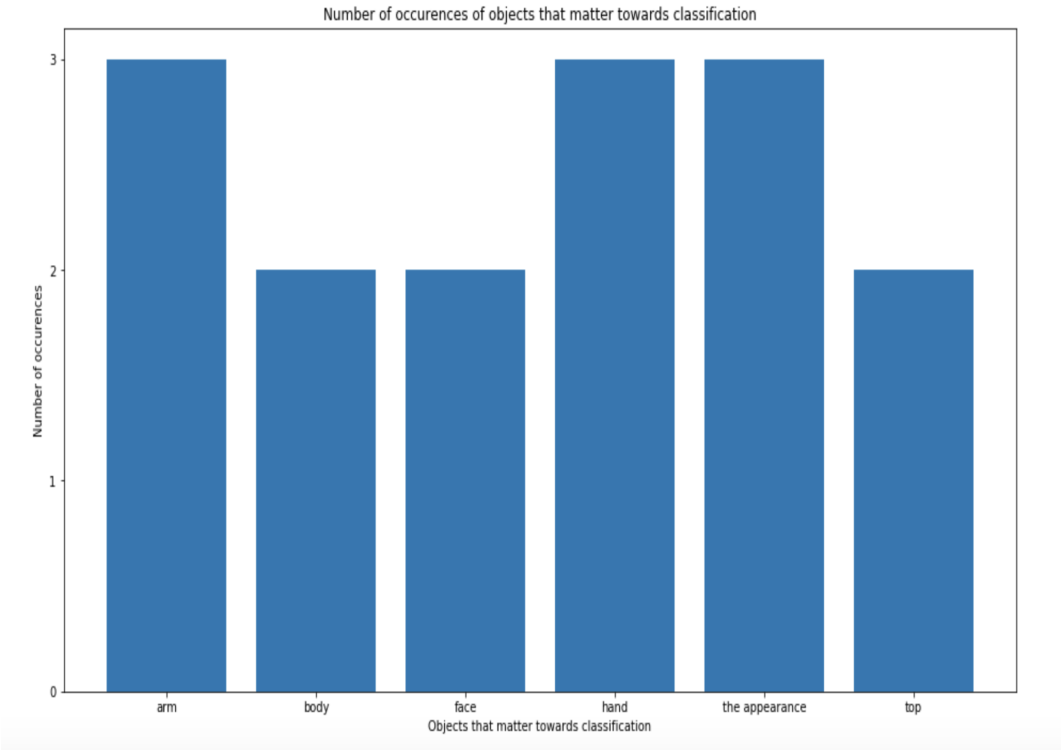


Figure 5.17: Number of occurrences of objects that matter towards classification for the farmer class (approach 2)

Based on Figure 5.17, we can observe that here there is a variety in the primary reason of the classification of a person as a farmer. The model look in visual clues like the the arm, face, body and hand. However we see now that our model takes into account the presence of these give ways elements of gender in a smaller number of images in comparison to those in engineer class. Therefore, the model does not act so much in a gender-bias way as in the engineer class, something that it was also verified in Table 5.16, where we saw that there was no gender bias in the predictions of the model in the farmer class.

Now, we continue with the evaluation of the third and last step of our methodology (approach 2), the bias mitigation step in the doctor/nurse dataset. It can be delineated in Table 5.31 the prediction performance of our classification model per class (doctor/nurse) on validation and test set after applying our methodology (approach 2).

Validation Set	**Test Set**
Doctor Accuracy: 90%	Doctor Accuracy: 89%
Nurse Accuracy: 92.5%	Nurse Accuracy: 85%

Table 5.31: Prediction performance per class on validation and test set (doctor-nurse dataset) (approach 2)

According to Table 5.31 we can stress two important points. Firstly, we see that our model now achieves a better performance. Overall, it has an accuracy now in the doctor class of at least 9% more than before applying our methodology on both validation and test set. Also, it has an accuracy in the nurse class of at least 3.7% more than before applying our methodology on both validation and test set.

Secondly, we observe that there is a very small difference now in the prediction performance in validation set for the class of doctor and nurse. Particularly, this difference is only 2.5%, in comparison to 8.8% that was before.

It can be depicted in Table 5.32 the prediction performance of our classification model after applying our methodology (approach 2) per gender (male/female) on validation set for the doctor and nurse class.

Doctor Class	**Nurse Class**
Male Accuracy: 90%	Male Accuracy: 90%
Female Accuracy: 90%	Female Accuracy: 95%

Table 5.32: Prediction performance per gender (male/female) on validation set for the doctor and nurse class (approach 2)

According to Table 5.32 we can stress three important points. Firstly and most importantly, we see that there is a significant decrease in gender bias in the predictions in doctor class (0% difference in accuracy, in comparison to 13.3% that it was before). Based on that, we can infer that there is smaller discrimination in the predictions of our classification model with respect to the gender for the doctor class in the validation set after applying our method.

Secondly, we observe that still there is no big difference (5% difference in accuracy) in the predictions of our classification model with respect to the gender for the nurse class in the validation set. Finally, we can also mention that there is a significant increase in accuracy of at least 3.4% for both classes and for both male and female.

It can be seen in Table 5.33 the prediction performance of our classification model after applying our methodology (approach 2) per gender (male/female) on test set for the doctor and nurse class.

Doctor Class	**Nurse Class**
Male Accuracy: 92%	Male Accuracy: 88%
Female Accuracy: 86%	Female Accuracy: 82%

Table 5.33: Prediction performance per gender (male/female) on test set for the doctor and nurse class (approach 2)

According to Table 5.33 we can stress three important points. Firstly and most importantly, we see that there is a significant decrease in gender bias in the predictions in doctor class (6% difference in accuracy, in comparison to 16% that it was before). Based on that, we can infer that there is smaller discrimination in the predictions of our classification model with respect to the gender for the doctor class in the test set after applying our method.

Secondly, we observe that still there is no big difference (6% difference in accuracy) in the predictions of our classification model with respect to the gender for the nurse class in the test set. Finally, we can also mention that there is a significant increase in accuracy of at least 4% for both classes and for both male and female.

Finally, it can be shown in Table 5.34 the prediction performance of our classification model after applying our methodology (approach 2) per gender (male/female) on validation and test set for the doctor and nurse class.

Doctor Class	**Nurse Class**
Male Accuracy: **91.1%**	**Male** Accuracy: **88.8%**
Female Accuracy: **87.7%**	**Female** Accuracy: **88.2%**

Table 5.34: Prediction performance per gender (male/female) on validation and test set for the doctor and nurse class (approach 2)

According to Table 5.34 we can stress three important points (similar to Tables 5.32-5.33. Firstly and most importantly, we see that there is a significant decrease in gender bias in the predictions in doctor class (3.4% difference in accuracy, in comparison to 14.8% that it was before). Based on that, we can infer that there is smaller discrimination in the predictions of our classification model with respect to the gender for the doctor class in both validation and test set after applying our method.

Secondly, we observe that still there is almost no difference (0.6% difference in accuracy) in the predictions of our classification model with respect to the gender for the nurse class in both validation and test set. Finally, we can also mention that there is a significant increase in accuracy of at least 3.5% for the nurse class for both male and female and at least 3.7% for the doctor class for both genders.

Thereafter, the conclusion that we draw after applying our methodology (approach 2) for the doctor-nurse dataset is that there is a significant mitigation of gender bias in the predictions in doctor class and a significant increase in accuracy in both classes for both male and female.

Now we present a similar analysis of the results after applying our methodology (approach 2) for the

chef/waiter dataset. It can be delineated in Table 5.35 the prediction performance of our classification model after applying our methodology (approach 2) per class (chef/waiter) on validation and test set.

Validation Set	**Test Set**
Chef Accuracy: 86.6%	Chef Accuracy: 83%
Waiter Accuracy: 87.7%	Waiter Accuracy: 89%

Table 5.35: Prediction performance per class on validation and test set (chef-waiter dataset) (approach 2)

According to Table 5.35 we can stress two important points. Firstly, we see that our model now achieves a better performance. Overall, it has an accuracy now in the chef class of at least 12% more than before applying our methodology on both validation and test set. Also, it has an accuracy in the waiter class of at least 3% more than before applying our methodology on both validation and test set.

Secondly, we observe that there is a very small difference now in the prediction performance in validation and test set for the class of chef and waiter. Particularly, this difference is only 1.1%, in comparison to 10% that was before for the validation set and 6%, in comparison to 15% for the test set that was before.

It can be depicted in Table 5.36 the prediction performance of our classification model after applying our methodology (approach 2) per gender (male/female) on validation set for the chef and waiter class.

Chef Class	**Waiter Class**
Male Accuracy: 95.5%	Male Accuracy: 93.3%
Female Accuracy: 77.7%	Female Accuracy: 82.2%

Table 5.36: Prediction performance per gender (male/female) on validation set for the chef and waiter class (approach 2)

According to Figure Table 5.36 we can stress three important points. Firstly and most importantly, we see that there is a significant decrease in gender bias in the predictions in chef class (17.8% difference in accuracy, in comparison to 28.8% that it was before). Based on that, we can infer that there is smaller discrimination in the predictions of our classification model with respect to the gender for the chef class in the validation set after applying our method.

Secondly, we observe that still there is no big difference (11.2% difference in accuracy) in the predictions of our classification model with respect to the gender for the waiter class in the validation set. Finally, we can also mention that there is a significant increase in accuracy of at least 2.2% for both classes and for both male and female.

It can be seen in Table 5.37 the prediction performance of our classification model after applying our methodology (approach 2) per gender (male/female) on test set for the chef and waiter class.

Chef Class	**Waiter Class**
Male Accuracy: 90%	Male Accuracy: 84%
Female Accuracy: 76%	Female Accuracy: 94%

Table 5.37: Prediction performance per gender (male/female) on test set for the chef and waiter class (approach 2)

According to Table 5.37 we can stress three important points. Firstly and most importantly, we see that there is a significant decrease in gender bias in the predictions in chef class (14% difference in accuracy, in comparison to 26% that it was before). Based on that, we can infer that there is smaller discrimination in the predictions of our classification model with respect to the gender for the chef class in the test set after applying our method.

Secondly, we observe that still there is no big difference (10% difference in accuracy) in the predictions of our classification model with respect to the gender for the waiter class in the test set. Finally, we can also mention that there is a significant increase in accuracy of at least 2% for both classes and for both male and female.

Finally, it can be shown in Table 5.38 the prediction performance of our classification model after applying our methodology (approach 2) per gender (male/female) on validation and test set for the chef and waiter class.

Chef Class	**Waiter Class**
Male Accuracy: **92.6%**	**Male** Accuracy: **88.7%**
Female Accuracy: **76.8%**	**Female** Accuracy: **88.1%**

Table 5.38: Prediction performance per gender (male/female) on validation and test set for the chef and waiter class (approach 2)

According to Table 5.38 we can stress three important points (similar to Tables 5.36-5.37. Firstly and most importantly, we see that there is a significant decrease in gender bias in the predictions in chef class (15.8% difference in accuracy, in comparison to 27.4% that it was before). Based on that, we can infer that there is smaller discrimination in the predictions of our classification model with respect to the gender for the chef class in both validation and test set after applying our method.

Secondly, we observe that still there is almost no difference (0.6% difference in accuracy) in the predictions of our classification model with respect to the gender for the waiter class in both validation and test set. Finally, we can also mention that there is a significant increase in accuracy of at least 3.1% for the waiter class for both male and female and at least 6.3% for the chef class for both genders.

Thereafter, the conclusion that we draw after applying our methodology (approach 2) for the chef-waiter dataset (and similar to the doctor/nurse dataset) is that there is a significant mitigation of gender bias in the predictions in chef class and a significant increase in accuracy in both classes for both male and female.

Finally we present a similar analysis of the results after applying our methodology (approach 2) for the engineer/farmer dataset. It can be delineated in Table 5.39 the prediction performance of our classification model after applying our methodology (approach 2) per class (engineer/farmer) on validation and test set.

Validation Set	**Test Set**
Engineer Accuracy: 93.8%	Engineer Accuracy: 97%
Farmer Accuracy: 93.8%	Farmer Accuracy: 100%

Table 5.39: Prediction performance per class on validation and test set (engineer-farmer dataset) (approach 2)

According to Table 5.39 we can stress two important points. Firstly, we see that our model now achieves a better performance. Overall, it has an accuracy now in the engineer class of at least 2.5% more than before applying our methodology on both validation and test set. Also, it has an accuracy in the farmer class of at least 1% more than before applying our methodology on both validation and test set.

Secondly, we observe that there is a smaller difference now in the prediction performance in validation and test set for the class of engineer and farmer with an improvement of 3%.

It can be depicted in Table 5.40 the prediction performance of our classification model after applying our methodology (approach 2) per gender (male/female) on validation set for the engineer and farmer class.

Engineer Class	**Farmer Class**
Male Accuracy: 95%	Male Accuracy: 90%
Female Accuracy: 92.5%	Female Accuracy: 97.5%

Table 5.40: Prediction performance per gender (male/female) on validation set for the engineer and farmer class (approach 2)

According to Table 5.40 we can stress three important points. Firstly and most importantly, we see that there is a significant decrease in gender bias in the predictions in engineer class (2.5% difference in accuracy, in comparison to 7.5% that it was before). Based on that, we can infer that there is smaller discrimination in the predictions of our classification model with respect to the gender for the engineer class in the validation set after applying our method.

Secondly, we observe that there is a small difference (7.5% difference in accuracy) in the predictions of our classification model with respect to the gender for the farmer class in the validation set. Finally, we can also mention that there is a significant increase in accuracy of at least 5% for both classes and for female gender.

It can be seen in Table 5.41 the prediction performance of our classification model after applying our methodology (approach 2) per gender (male/female) on test set for the engineer and farmer class.

Engineer Class	**Farmer Class**
Male Accuracy: 98%	Male Accuracy: 100%
Female Accuracy: 96%	Female Accuracy: 100%

Table 5.41: Prediction performance per gender (male/female) on test set for the engineer and farmer class (approach 2)

According to Table 5.41 we can stress three important points. Firstly and most importantly, we see that there is a significant decrease in gender bias in the predictions in engineer class (2% difference in accuracy, in comparison to 6% that it was before). Based on that, we can infer that there is smaller discrimination in the predictions of our classification model with respect to the gender for the engineer class in the test set after applying our method.

Secondly, we observe that there is no difference (0% difference in accuracy) in the predictions of our classification model with respect to the gender for the farmer class in the test set. Finally, we can also mention that there is an increase in accuracy of 2% in engineer class for male gender, 6% in engineer class for female gender and 1% in farmer class for male gender.

Finally, it can be shown in Table 5.42 the prediction performance of our classification model after applying our methodology (approach 2) per gender (male/female) on validation and test set for the engineer and farmer class.

Engineer Class	**Farmer Class**
Male Accuracy: **96.6%**	**Male** Accuracy: **95%**
Female Accuracy: **94.4%**	**Female** Accuracy: **98.8%**

Table 5.42: Prediction performance per gender (male/female) on validation and test set for the engineer and farmer class (approach 2)

According to Table 5.42 we can stress three important points (similar to Tables 5.40-5.41). Firstly and most importantly, we see that there is a significant decrease in gender bias in the predictions in engineer class (2.2% difference in accuracy, in comparison to 6.7% that it was before). Based on that, we can infer that there is smaller discrimination in the predictions of our classification model with respect to the gender for the engineer class in both validation and test set after applying our method.

Secondly, we observe that still there is small difference (3.8% difference in accuracy) in the predictions of our classification model with respect to the gender for the farmer class in both validation and test set. Finally, we can also mention that there is a significant increase in accuracy of 1.1% in the engineer class for male gender, 5.6% in the engineer class for female gender, 0.6% in the farmer class for male gender and 2.2% in the farmer class for female gender.

Thereafter, the conclusion that we draw after applying our methodology (approach 2) for the engineer-farmer dataset (and similar to the doctor/nurse and chef/waiter datasets) is that there is a significant mitigation of gender bias in the predictions in engineer class and a significant increase in accuracy in both classes for both male and female.

Finally, we would like to stress that for all datasets, after applying our methodology (approach 2), the probabilities that are related with a specific classification outcome were improved in at least 65% of the cases. For instance, even in cases that we had still miss-predictions after applying our methodology (approach 2), there was an improvement in at least 65% of the cases in confidence that is related with that classification outcome for all datasets (e.g. we have a situation like the following in at least 65% of the miss-classified images in all datasets: we have an image with a real label of doctor, but the prediction of the model is nurse with confidence of 80% before applying our method and 65% after applying our method). Finally, for easier reference, we quote all the aforementioned comments for all datasets in Tables 5.43 and 5.44.

Statistical Parity (Before)	**Statistical Parity (After)**
Doctor: **14.8%**	Doctor: **3.4% (11.4% improvement)**
Chef: **27.4%**	Chef: **15.8%(11.6% improvement)**
Engineer: **6.7%**	Engineer: **2.2%(4.5% improvement)**

Table 5.43: Statistical parity in doctor, chef and engineer class, before and after applying our methodology (approach 2)

Class + Gender	Improvement in Accuracy (Validation and Test Set)
Doctor + Male	**+3.7%**
Doctor + Female	**+15.1%**
Nurse + Male	**+3.5%**
Nurse + Female	**+3.7%**
Chef + Male	**+6.3%**
Chef + Female	**+17.9%**
Waiter + Male	**+3.3%**
Waiter + Female	**+3.1%**
Engineer + Male	**+1.1%**
Engineer + Female	**+5.6%**
Farmer + Male	**+0.6%**
Farmer + Female	**+2.2%**

Table 5.44: Improvement in accuracy (validation and test set) for all datasets after applying our methodology (approach 2)

Therefore, based on the results of Tables 5.43 and 5.44 we can notice that the use of this approach (approach 2) was able to increase the general performance (accuracy) of our classification model among all the classes doctor, nurse, chef, waiter, engineer and farmer and for both genders with a percentage of at least 0.6% and at most 17.9%. Thus, we can infer based on this that the intuition of people about a potential cause of gender bias actually matches in a very significant degree with the actual reason that affects the prediction of a Machine Learning classification model, as obfuscating these clues leads to a better performance.

5.4.3. COMPARISON OF THE RESULTS BETWEEN APPROACH 1 AND 2

In this section we provide a comparison of the results of these two approaches (Approach 1)-(Approach 2) with respect to their performance and differences. Firstly, we want to compare these approaches with respect to the bias semantic interpretation step. Secondly, we compare them with respect to the bias mitigation step at two layers, namely: bias mitigation and increase in overall performance.

Firstly, we provide the comparison with respect to the bias semantic interpretation step. It can be inferred through Figures 5.5, 5.7, 5.8, 5.10 (approach 1) and Figures 5.12, 5.14, 5.16 (approach 2) that in images that the prediction of the classification model was wrong, the model was mostly looking in the presence of the face or other direct give away elements of the gender of people like hair, facial hairs, nail and body for the classes of doctor, chef and engineer, where the gender bias was huge. Therefore, we had an agreement between the two approaches.

Also, in Figures 5.6, 5.9, 5.11 (approach 1) and Figures 5.13, 5.15, 5.17 (approach 2) and more specifically in the classes of nurse, waiter and farmer, the attention of the model was in a lower degree in the presence of direct give away elements of the gender of people like the face or body and therefore this was the reason actually that we did not have so much gender bias in these classes. Hence, we also had an agreement between the two approaches for these classes.

One thing that we would like also to stress here is that in approach 1 we were able to end up also with objects that are related with the specific profession (e.g. uniform). In contrast in approach 2, we had the opportunity to have a better understanding of the reason of a specific prediction through having a more detailed description (e.g. face, hair, nails and not just person (like in approach 1)). Thus, we could say that the two different approaches can be used in a complementary way.

Now, we continue with the comparison of these two approaches with respect to the bias mitigation step (bias mitigation layer). It can be seen in Table 5.45 the statistical parity in doctor, chef and engineer class, before and after applying our methodology (approach 1 and 2). We can infer through this that there is a balance between approach 1 and approach 2 with respect to the bias mitigation. However, we could notice that approach 2 produces better results in doctor class (5.1% improvement in statistical parity in comparison to approach 1).

Statistical Parity (Before)	Statistical Parity (After Approach 1)	Statistical Parity (After Approach 2)
Doctor: **14.8%**	Doctor: **8.5%**	Doctor: **3.4%**
Chef: **27.4%**	Chef: **15.8%**	Chef: **15.8%**
Engineer: **6.7%**	Engineer: **2.2%**	Engineer: **2.2%**

Table 5.45: Statistical parity in doctor, chef and engineer class, before and after applying our methodology (approach 1 and 2)

Finally, we compare the two approaches with respect to the general performance (accuracy). It can be depicted in Table 5.46 the improvement in accuracy (validation and test set) for all datasets after applying our methodology (approach 1 and 2). We can infer through this at in seven out of twelve of the cases approach 2 produces better results than approach 1. In only four cases approach 1 produces better results than approach 2 and in one case the result is identical. Therefore, we can conclude that in general also in improvement in accuracy, approach 2 produces better results than approach 1.

Class + Gender	Improvement in Accuracy (Validation and Test Set Approach 1)	Improvement in Accuracy (Validation and Test Set Approach 2)
Doctor + Male	**+5.3%**	**+3.7%**
Doctor + Female	**+11.6%**	**+15.1%**
Nurse + Male	**+2%**	**+3.5%**
Nurse + Female	**+2.2%**	**+3.7%**
Chef + Male	**+5.3%**	**+6.3%**
Chef + Female	**+16.9%**	**+17.9%**
Waiter + Male	**+4.1%**	**+3.3%**
Waiter + Female	**+4.3%**	**+3.1%**
Engineer + Male	**+0%**	**+1.1%**
Engineer + Female	**+4.5%**	**+5.6%**
Farmer + Male	**+4.4%**	**+0.6%**
Farmer + Female	**+2.2%**	**+2.2%**

Table 5.46: Improvement in accuracy (validation and test set) for all datasets after applying our methodology (approach 1 and 2)

5.4.4. QUALITATIVE ANALYSIS AND GENERALIZABILITY

In this section we are going to provide a qualitative analysis of the results of our experiments and explaining the reason and the cases that our methodology works well or not. Also, we provide some comments regarding the generalization capabilities of our method.

We saw in the evaluation of the bias detection step that there was a significant gender bias in the predictions of our classification model with respect to the gender in doctor class, in chef class and in engineer class. Moreover we depicted that there was a negligible difference in accuracy in the predictions of our classification model with respect to the gender for the nurse, waiter and farmer class.

In order to be able to reason upon such a difference in performance, we evaluated the bias semantic interpretation step. Based on this, we found that our classification model, in images that its prediction was wrong, was mostly looking in the presence of the face or body (person class) for the classes of doctor, chef and engineer, where the gender bias was huge and in the classes of nurse, waiter and farmer, its attention was in a lower grade in the presence of the face or body (person class). Therefore this was the reason actually that we did not have so much gender bias in these classes.

Finally, our goal was to compensate for this gender bias in the classes of doctor, chef and engineer and to increase the performance of our model in all classes. Particularly, through applying our bias mitigation step, we managed in decreasing the statistical parity in the doctor, chef and engineer class, via obfuscating the objects that introduced bias to the classification outcome (e.g person). Moreover, it is worth mentioning that we achieved in increasing the general performance (accuracy) of our classification model among all the classes doctor, nurse, chef, waiter, engineer and farmer and for both genders.

We should stress here (and based on the results of Table 5.46) that our methodology enables the mitigation of bias and the increase in accuracy in all cases. However, it is worth mentioning that this improvement varies significantly among different cases and sometimes our methodology provides only minor improvement. The reason behind this, depends mainly on the evaluation of the second step (bias semantic interpretation step). As we found on this step, the classification model, was mostly looking in the presence of the face or body (person class) in order to classify a data point. However, in some data points the face or body of the professional of interest is not visible enough or it is only a small portion of the whole image. Hence, obfuscating the face or the body and re-classifying this data point, has only a minor effect in the final outcome and the improvement would be slight. Therefore, in cases like this, where the main reason of a specific classification (face or the body), is not a significant portion of the data point, our methodology does not work so well.

As far as the generalization capabilities of our methodology are concerned, we could mention the following comments. Based on the way that our methodology works (3-steps, bias detection, semantic interpretation and mitigation) it is pretty straight forward to apply it on different use cases, data, protected attributes and to use different Machine Learning models. Moreover, based on the fact that our methodology works very well in a variety of different datasets coming from different distributions, we can infer that it is able to generalize with a good performance on new and unseen settings.

5.4.5. CONCLUSIONS

In this section, we evaluated the second and the third step of our methodology, the bias semantic interpretation and mitigation step (approach 2). We presented the experiments that we performed for our use-case of profession prediction from images (doctor/nurse, chef/waiter and engineer/farmer) in order to semantically describe the reason of a particular prediction of our classification model in miss-classified image data and to compensate for gender bias that is related with these predictions. Also we provided a comparison of the approach 1 and 2. We were able to draw some important conclusions.

Firstly, we were able to understand the reason of the discrimination in the predictions of our classification model with respect to the gender through semantically described it. More specifically, we shown that our classification model, in images that its prediction was wrong, was mostly looking in the presence of the face or other direct give away elements of the gender of people like hair, facial hairs, nail and body for the classes of doctor, chef and engineer, where the gender bias was huge. In the classes of nurse, waiter and farmer, its attention was in a lower degree in the presence of the face or in these direct give away elements of the gender and therefore this was the reason actually that we did not have so much gender bias in these classes.

Secondly, we were able to compensate for this gender bias in the classes of doctor, chef and engineer. Particularly, through applying our bias mitigation step, we managed in decreasing the statistical parity from 14.8% to 3.4% in the doctor class, from 27.4% to 15.8% in the chef class and from 6.7% to 2.2% in the engineer class, via obfuscating the direct give away elements of the gender that introduced bias to the classification outcome (e.g face, hairstyle etc).

Thirdly, we achieved in increasing the general performance (accuracy) of our classification model among all the classes doctor, nurse, chef, waiter, engineer and farmer and for both genders. Meticulously, we were able to increase the accuracy of each class and for both genders with a percentage of at least 0.6% and at most 17.9%.

Moreover, we inferred that the intuition of people about a potential cause of gender bias actually matches in a very significant degree with the actual reason that affects the prediction of a Machine Learning classification model, as obfuscating these clues leads to a better performance.

Furthermore, we saw that in terms of bias semantic interpretation the two approaches had a similar behavior. Particularly, they can be used in a complementary way as approach 1 ends up also with objects that are related with the specific profession (e.g. uniform) in contrast to approach 2, and approach 2 ends up also with objects that are more descriptive (e.g. face, nails etc.) in contrast to approach 1.

In addition, as far as the bias mitigation is concerned, we saw that approach 1 and 2 had a similar behavior. However, approach 2 had a better performance in the doctor class. Finally, approach 2 also had a better performance in comparison to approach 1 with respect to the improvement in accuracy in most of the cases (seven out of the twelve cases). In only four cases approach 1 produced better results than approach 2 and in one case the result was identical. Therefore, we concluded that in general also in improvement in accuracy, approach 2 produced better results than approach 1.

5.5. SUMMARY

In this chapter, we described the evaluation of the methodology that we propose for bias detection, semantic interpretation and mitigation in chapter 4. Particularly, we wanted to have an understanding of how our methodology performs towards answering our second and third research sub-questions (**RSQ2**)+(**RSQ3**) and more specifically of semantically describe at scale the reason that a particular prediction of a Machine Learning classification model is made and compensating for gender bias that is related with the content of the image.

We presented the experiments that we performed for the use-case of profession prediction from images (doctor/nurse, chef/waiter and engineer/farmer). Especially, we evaluated the bias detection step and the semantic interpretation and mitigation of bias step (two approaches, attention mechanism-object detection (approach 1) and attention mechanism-crowdsourcing (approach 2)) in terms of the evaluation metrics that

we adopt.

Notably, we measured accuracy (good indicator of performance) and statistical parity (good indicator of bias) before and after applying our methodology towards compensating for gender bias that is related with the content of the image data. Finally, we compared these two approaches, with respect to the semantic description of the reason that a particular prediction of a Machine Learning classification model is made, that they end up.

More specifically, we started by providing the evaluation of the bias detection step, where we wanted to observe whether there is discrimination in the predictions of the Machine Learning classification model with respect to the gender. Based on this we found that there was a significant gender bias in the predictions of our classification model with respect to the gender in doctor class (doctor/nurse dataset) with a 14.8% difference in accuracy, in chef class (chef/waiter dataset) with a 27.4% difference in accuracy and in engineer class (engineer/farmer dataset) with a 6.7% difference in accuracy even though there was an equal distribution among the doctors, chefs and engineers with respect to the gender in the training set.

Moreover, we observed that there was a negligible difference (0.8% difference in accuracy), (0.4% difference in accuracy) and (2.2% difference in accuracy) in the predictions of our classification model with respect to the gender for the nurse, waiter and farmer class in the validation and test sets respectively.

As a next step, we evaluated the second and the third step of our methodology, the bias semantic interpretation step and the bias mitigation step in order to semantically describe the reason that a particular prediction of a Machine Learning classification model is made and to observe whether the application of our methodology is able to compensate for gender bias that is related with the content of the image.

Particularly, we evaluated firstly our first approach (Approach 1) which uses the correlation or overlapping between the attention mechanism and the object detection. Based on this, we shown that our classification model, in images that its prediction was wrong, was mostly looking in the presence of the face or body (person class) for the classes of doctor, chef and engineer, where the gender bias was huge and in the classes of nurse, waiter and farmer, its attention was in a lower grade in the presence of the face or body (person class) and therefore this was the reason actually that we did not have so much gender bias in these classes.

Also, we were able to compensate for this gender bias in the classes of doctor, chef and engineer. Particularly, through applying our bias mitigation step, we managed in decreasing the statistical parity from 14.8% to 8.5% in the doctor class, from 27.4% to 15.8% in the chef class and from 6.7% to 2.2% in the engineer class, via obfuscating the objects that introduced bias to the classification outcome (e.g person).

It is worth mentioning that we achieved in increasing the general performance (accuracy) of our classification model among all the classes doctor, nurse, chef, waiter, engineer and farmer and for both genders. We were able to increase the accuracy of each class and for both genders with a percentage of at least 2% and at most 16.9% (engineer class for male gender is excluded, where we did notice any difference).

As a next step, we evaluated the second approach (approach 2) which uses the correlation or overlapping between the attention mechanism and the crowdsourcing task. We found that our classification model, in images that its prediction was wrong, was mostly looking in the presence of the face or other direct give away elements of the gender of people like hair, facial hairs, nail and body for the classes of doctor, chef and engineer, where the gender bias was huge and in the classes of nurse, waiter and farmer, its attention was in a lower grade in the presence of the face or in these direct give away elements of the gender and therefore this was the reason actually that we did not have so much gender bias in these classes.

Furthermore, we were able to compensate for this gender bias in the classes of doctor, chef and engineer. Particularly, through applying our bias mitigation step, we managed in decreasing the statistical parity from 14.8% to 3.4% in the doctor class, from 27.4% to 15.8% in the chef class and from 6.7% to 2.2% in the engineer class, via obfuscating the direct give away elements of the gender that introduced bias to the classification outcome (e.g face, hairstyle etc).

In addition, we achieved in increasing the general performance (accuracy) of our classification model among all the classes doctor, nurse, chef, waiter, engineer and farmer and for both genders. Meticulously, we were able to increase the accuracy of each class and for both genders with a percentage of at least 0.6% and at most 17.9%.

Last but not least, we inferred that the intuition of people about a potential cause of gender bias actually matches in a very significant degree with the actual reason that affects the prediction of a Machine Learning classification model, as obfuscating these clues leads to a better performance.

Finally, we provided a comparison of the results of these two approaches (Approach 1)-(Approach 2) with respect to their performance and differences. We saw that in terms of bias semantic interpretation the two approaches had a similar behavior. Particularly, they can be used in a complementary way as approach 1

ends up also with objects that are related with the specific profession (e.g. uniform) in contrast to approach 2, and approach 2 ends up also with objects that are more descriptive (e.g. face, nails etc.) in contrast to approach 1.

As far as the bias mitigation is concerned, we saw that approach 1 and 2 had a similar behavior. However, approach 2 had a better performance in the doctor class. Finally, approach 2 also had a better performance in comparison to approach 1 with respect to the improvement in accuracy in most of the cases (seven out of the twelve cases). In only four cases approach 1 produced better results than approach 2 and in one case the result was identical. Therefore, we concluded that in general also in improvement in accuracy, approach 2 produced better results than approach 1.

6

CONCLUSION

In this chapter, we start by discussing the work done in the thesis and draw conclusions to answer our research questions. Finally we propose future work considering also our conclusions.

6.1. DISCUSSION OF CURRENT WORK

In this section, we highlight the focus of our work, the potential strong points and the limitations of our methodology.

6.1.1. FOCUS OF THE WORK

As we shown, imposing an equal distribution of the data with respect to the protected attribute of people, does not always solve the problem of bias. Hence, it is crucial to be able to investigate the data, the algorithm and the output of Machine learning models which are made for predicting decisions that are highly related to different groups of people in order to identify potential issues that are related to bias towards them.

Therefore, in this work, we proposed a methodology to identify, explain and mitigate gender bias in Machine Learning data. We focused on this specific angle of the problem, because we strongly believe that it is extremely important to tackle these aspects of the problem properly, as in a opposite case there is a high risk of harmful consequences and negative impacts towards the lives of the groups of people who receive biased, unfair and undesirable decisions.

6.1.2. METHODOLOGY OF OUR APPROACH

Our thesis work is organized along a methodology that we propose and evaluate. Particularly, our methodology has three main steps. We believe that combining these three steps is a strong and innovative point of our project.

More specifically, the three main steps of our methodology are: A bias detection step, a bias semantic interpretation step and a bias mitigation step. Our methodology takes as an input some image data for which we define a protected attribute (we adopt the gender in our work). After that, we start with the first step of our methodology (bias detection step), where we pass these images to a Machine Learning classification model and we end up with some miss-classified data. The goal of this step is to observe whether there is discrimination in the predictions of this Machine Learning classification model with respect to the gender.

We continue with the second step of our methodology (bias semantic interpretation step), where we pass these miss-classified data to a debug mechanism that has three parts (attention mechanism, object detection and crowdsourcing). The goal of this step is to semantically describe the reason of these miss-classifications in the predictions of the Machine Learning classification model in a human interpretable way.

Finally, in the third step of our methodology (bias mitigation), we pass the output of the previous step to a correction mechanism (obfuscation task) and we end up with some new image data that we feed them again to the the same Machine Learning classification model. The goal of this step is to compensate for gender bias that is related with the content of the image data.

To the best of our knowledge, most research tackle solely one or two of the three steps, and more specifically they try to detect and mitigate gender bias and do not pay any attention to explain their reasoning in a human interpretable way. Also, the main method that they use is to balance the distribution of the train-

ing data with respect to the protected attribute. However, as we shown, this is not always the solution to the problem. On the contrary, we decided to study concurrently these three steps to overcome these shortcoming and limitations and we performed an extensive evaluation of our method in order to verify its efficiency and effectiveness.

We presented the experiments that we performed for the use-case of profession prediction from images (doctor/nurse, chef/waiter and engineer/farmer). Especially, we evaluated the bias detection step and the semantic interpretation and mitigation of bias step (two approaches, attention mechanism-object detection (approach 1) and attention mechanism-crowdsourcing (approach 2)) in terms of the evaluation metrics that we adopt.

Notably, we measured accuracy (good indicator of performance) and statistical parity (good indicator of bias) before and after applying our methodology towards compensating for gender bias that is related with the content of the image data. Finally, we compared these two approaches, with respect to the semantic description of the reason that a particular prediction of a Machine Learning classification model is made, that they end up.

To this end, we observed that there was a discrimination in the predictions of the Machine Learning classification model with respect to the gender (bias detection) even though there was an equal distribution of the labels with respect to the gender in the training data and mostly for the classes of doctor, chef and engineer. We were able to semantically describe at scale the reason that a particular prediction of a Machine Learning system is made in a human interpretable way (bias semantic interpretation) and we saw that the main reason on these classes (doctor, chef and engineer) was the presence of the face (direct give away element of the gender). Finally, based on our two approaches, we compensated for gender bias that is related with the content of the image data (bias mitigation) and particularly we observed a significant mitigation of gender bias in the predictions in doctor, chef and engineer class and a significant increase in accuracy in all classes for both male and female gender.

6.1.3. LIMITATIONS OF OUR APPROACH

The most important limitation of our approach is the fact that we did not verify its effectiveness in a variety of professions and/or visual tasks that may related with bias. Also, we did not explore its capabilities in other media like text, video etc. The main reason that we did not perform such an exploration is a fact that is also related with the main limitation of Machine Learning algorithms: they require massive stores of training data. Therefore, in order to do such a verification, more data are needed and it is not always easy to have access on them.

Moreover, something closely related to that, is that labeling training data is a tedious process and based on that it would be difficult for us to collect more data for the professions that we used or to experiment with more professions or a different visual task. In addition, we only made experiment with gender as the chosen protected attribute. In would be nice to validate our methodology on other kind of protected attributes like race or age in order to observe its generalization performance. Finally, even though we tried to identify, explain and mitigate gender bias that appears in the data, based on the fact that we use a pre-collected dataset, there is always the case for these data to reflect human biases in the way that they collected.

6.2. CONCLUSIONS

Machine Learning models are increasingly used to assist or replace humans in a variety of decision-making domains. However, there is the case these decisions in the aforementioned life-affecting scenarios to have important negative impacts on the lives of people who are involved as there be might the case that there is bias on these decisions. Also, there is a lack of methods that actually interpret and explain the predictions of these ML systems in a human interpretable way which are then used to help decision making. That is why we had set up a methodology to study how to **analyze, reason upon and fix** Machine Learning **training data** with respect to the gender and to content of the images that are **balanced** in terms of **gender bias** to the **output** of these models (**MRQ**). Particularly, we broke this main research question (**MRQ**) into three research sub-questions, namely:

RSQ1: Which are the **current methods** and their **limitations** related to **bias in Machine learning data** with respect to **protected attributes** of people?

RSQ2: How can we **describe** in a **semantically rich fashion at scale** the **features** in the **data** that are likely to

be **related** to a particular **biased prediction** of a **Machine Learning system**?

RSQ3: How can we **compensate** for **gender bias** that is related with the **content** of the **image data** in **Machine Learning** systems?

RSQ3a: How much is the **content** of the image **correlated** to the **prediction errors** with respect to the **gender**?

RSQ3b: How can we use **crowdsourcing** to help **uncover** potential **unknown elements** of **gender bias** that may reside in Machine Learning **data**?

In order to answer there research question we did the following procedure: As a first step **(Chapter 2)** we conducted a **literature review** of the different fields concerned with our main research and sub-research questions. As a next step **(Chapter 3)**, we presented the **use case** (profession prediction from images), **dataset** and the **classification task** to study detection, semantic interpretation and mitigation of gender bias. Then we tackled the second and third research sub-questions (**RSQ2**)+(**RSQ3**), where we described the **methodology** that we made in order to answer these research sub-questions **(Chapter 4)**.Finally, we gave the **evaluation** and the results of the experiments that we did in order to verify the effectiveness of our scheme **(Chapter 5)**. The main contributions of our work are as follows:

CO1: The first contribution of the thesis was an **in-depth systematic literature review** of the state-of-the-art methods and their limitations related to bias with respect to protected attributes of people in Machine Learning. We investigated existing literature on the topics of definitions and evaluation measures of bias, bias mitigation algorithms and bias identification using the crowd which enabled us to highlight current limitations as well as to come up with possible directions to develop our methodology. It enabled to answer the first research sub-question **(RSQ1)**.

CO2: The second contribution was the answer to the second research sub-question (**RSQ2)**. More specifically, it was the **one part** of the **methodology** that we proposed of describing in a **semantically rich fashion at scale** the **features** in the **data** that are likely to be **related** to a particular **biased prediction** of a **Machine Learning system** in a human interpretable way.

CO3: The third contribution was the **other part** of the **methodology** that we proposed and was the answer to our third research sub-question (**RSQ3)**. Particularly, of **compensating** for **gender bias** that is related with the **content** of the **image data** in **Machine Learning** systems.

To summarize, we proposed a methodology that firstly allows us to identify gender bias in Machine Learning data through observing whether there is discrimination in predictions of a Machine learning classification model with respect to the gender. Secondly, it enables us to semantically describe at scale the reason that a particular prediction of a Machine Learning system is made in a human interpretable way. Finally, it can help towards compensating for gender bias that is related with the content of the image data in Machine Learning systems. Through the results of our experiments, we were able to validate the effectiveness of our methodology with respect to these three goals. Therefore, the main hypothesis (**H**) of the thesis is verified. The **presence** of **visual clues** in image data that **give away** the protected attribute (e.g. **gender**), **affect** the **classification** outcome and **introduce bias** on that.

6.3. PROPOSITION OF FUTURE WORK

In this section, we propose future work in order to make our methodology more effective and more generalizable.

6.3.1. APPLICATION TO DIFFERENT USE-CASES

The complete thesis work is based on Machine Learning systems for profession prediction from images. However we aim at studying the prediction of Machine Learning classification models for visual tasks that bias may appear. Consequently future work should also address the generalization of our method to other visual tasks which may involve bias with respect to protected attributes of people. The main steps of our methodology (bias detection, semantic interpretation and mitigation) can be directly applied to any Machine Learning

classification task that involves image data without any modification.

6.3.2. APPLICATION TO DIFFERENT MACHINE LEARNING TASKS

The complete thesis work is based on a Machine Learning image classification task. Nonetheless, our methodology could be easily adapted to different Machine Learning tasks. For instance it can be used for bias detection, semantic interpretation and mitigation of object detection tasks. Another possible task, would be in Captioning Models, where the goal is to provide a description given an image. On the other hand, through making some slight modifications, our methodology could be also extended to other media (e.g. text and video). For example, for text classification or Question Answering (QA).

6.3.3. MODIFICATION OF THE BUILDING BLOCKS OF THE METHODOLOGY

One key element in our methodology of attaching semantic interpretation on top of the attention mechanism was the use of the object detection part. However, in case that we have more computational power and more time to deploy such a task, other approaches like semantic segmentation and instance segmentation could be also employed. Furthermore, one key element that we used in order to end up with the objects that matter towards the classification was the use of the Intersection over Union score. We said that, in case that this overlap between the bounding box coming from the attention mechanism with the bounding box coming from the object detection or the crowdsourcing is larger that 0.5 then we add the object that corresponds to that bounding box, to the list of objects that matter.

One extra thing that would be interesting to be investigated as a future work would be to extend this metric in a way that involves a gray-scale calculation in terms of percentages and not in a binary way. Our reasoning of choosing this binarization was the fact that we wanted to tie visual clues to objects this yes/no overlapping makes things easier towards explainability. Moreover, the proposition of one new metric of measuring the overlap between objects (and/or maybe designing a crowdsourcing task for this?) is something that is worth of extra exploration. In addition, we focused only on the case that the protected attribute is the gender. We could also have explored other protected attributes like race and age. Finally, in order to mitigate the gender bias we used an obfuscation part. Instead of that, we could try some other modifications like removing or adding objects to the images.

6.3.4. CREATION OF A NEW IMAGE DATASET

Finally, one thing that is worth mentioning with which we can deal as future work is the creation of a new image dataset. More specifically, based on the results of our methodology and particularly of step two (semantic interpretation of the bias), we ended up with some objects. As we said, we obfuscated these objects and we passed the images that derived from this modification to the initial Machine Learning classification model. However, one thing that it would be interesting to do, is to create an image dataset where the images do not consist of these objects that matter towards the classification and introduce bias on this. Therefore, to construct a dataset with images without objects (or being obfuscated) that give away the gender directly (e.g. face, tie, nails etc) and based on that data to train a new Machine Learning classification model and observe its performance with respect to gender bias.

BIBLIOGRAPHY

[1] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney, *Optimized pre-processing for discrimination prevention,* in *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017) pp. 3992–4001.

[2] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, *Algorithmic decision making and the cost of fairness,* CoRR **abs/1701.08230** (2017), arXiv:1701.08230 .

[3] A. Chouldechova, *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments,* CoRR **abs/1610.07524** (2016), arXiv:1610.07524 .

[4] R. Nabi and I. Shpitser, *Fair inference on outcomes,* Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence **2018**, 1931 (2018).

[5] G. Pleiss, M. Raghavan, F. Wu, J. M. Kleinberg, and K. Q. Weinberger, *On fairness and calibration,* CoRR **abs/1709.02012** (2017).

[6] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, *Fairness beyond disparate treatment disparate impact: Learning classification without disparate mistreatment,* in *WWW* (2017).

[7] K. Lum and J. Johndrow, *A statistical framework for fair predictive algorithms,* (2016).

[8] N. Grgic-Hlaca, E. M. Redmiles, K. P. Gummadi, and A. Weller, *Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction,* in *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018* (2018) pp. 903–912.

[9] N. Grgic-Hlaca, M. B. Zafar, K. P. Gummadi, and A. Weller, *Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning,* in *AAAI* (2018).

[10] A. Datta, M. C. Tschantz, and A. Datta, *Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination,* CoRR **abs/1408.6491** (2014), arXiv:1408.6491 .

[11] L. Sweeney, *Discrimination in online ad delivery,* Queue **11**, 10:10 (2013).

[12] M. Hardt, E. Price, and N. Srebro, *Equality of opportunity in supervised learning,* CoRR **abs/1610.02413** (2016).

[13] R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, *Learning fair representations,* in *ICML* (2013).

[14] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, *The variational fair auto encoder,* (2015).

[15] T. Calders and S. Verwer, *Three naive bayes approaches for discrimination-free classification,* Data Mining and Knowledge Discovery **21**, 277 (2010).

[16] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, *Certifying and removing disparate impact,* in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15 (ACM, New York, NY, USA, 2015) pp. 259–268.

[17] D. Pedreschi, S. Ruggieri, and F. Turini, *Discrimination-aware data mining,* Tech. Rep. (2007).

[18] S. Verma and J. Rubin, *Fairness definitions explained,* in *Proceedings of the International Workshop on Software Fairness*, FairWare '18 (ACM, New York, NY, USA, 2018) pp. 1–7.

[19] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, *Fairness constraints: Mechanisms for fair classification,* in *AISTATS* (2017).

[20] R. Kohavi, *Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid,* in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96 (AAAI Press, 1996) pp. 202–207.

[21] T. Kamishima, S. Akaho, and J. Sakuma, *Fairness-aware learning through regularization approach,* in *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*, ICDMW '11 (IEEE Computer Society, Washington, DC, USA, 2011) pp. 643–650.

[22] I. Zliobaite, *On the relation between accuracy and fairness in binary classification,* CoRR **abs/1505.05723** (2015).

[23] B. H. Zhang, B. Lemoine, and M. Mitchell, *Mitigating unwanted biases with adversarial learning,* CoRR **abs/1801.07593** (2018).

[24] N. Kallus and A. Zhou, *Residual unfairness in fair machine learning from prejudiced data,* in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018* (2018) pp. 2444–2453.

[25] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva, *Counterfactual fairness,* in *NIPS* (2017).

[26] C. Simoiu, S. Corbett-Davies, and S. Goel, *The problem of infra-marginality in outcome tests for discrimination,* The Annals of Applied Statistics **11**, 1193 (2017).

[27] J. M. Kleinberg, S. Mullainathan, and M. Raghavan, *Inherent trade-offs in the fair determination of risk scores,* CoRR **abs/1609.05807** (2016).

[28] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel, *Fairness through awareness,* CoRR **abs/1104.3913** (2011).

[29] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, *Fairness in criminal justice risk assessments : The state of the art,* (2017).

[30] S. Galhotra, Y. Brun, and A. Meliou, *Fairness testing: Testing software for discrimination,* CoRR **abs/1709.03221** (2017).

[31] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, *Avoiding discrimination through causal reasoning,* in *Advances in Neural Information Processing Systems 30* (Curran Associates, Inc., 2017) pp. 656–666.

[32] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. A. Baeza-Yates, *Fa*ir: A fair top-k ranking algorithm,* CoRR **abs/1706.06368** (2017).

[33] D. Pedreschi, S. Ruggieri, and F. Turini, *Measuring discrimination in socially-sensitive decision records,* in *SDM* (SIAM, 2009) pp. 581–592.

[34] D. Pedreschi, S. Ruggieri, and F. Turini, *Discrimination-aware data mining,* in *KDD* (2008).

[35] T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, A. Weller, and M. B. Zafar, *A unified approach to quantifying algorithmic unfairness: Measuring individual and group unfairness via inequality indices,* Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD) (2018).

[36] F. Kamiran and T. Calders, *Classifying without discriminating,* in *2009 2nd International Conference on Computer, Control and Communication* (2009) pp. 1–6.

[37] F. Kamiran and T. Calders, *Data preprocessing techniques for classification without discrimination,* Knowledge and Information Systems **33**, 1 (2011).

[38] T. Calders, F. Kamiran, and M. Pechenizkiy, *Building classifiers with independency constraints,* 2009 IEEE International Conference on Data Mining Workshops , 13 (2009).

[39] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets,* in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14 (MIT Press, Cambridge, MA, USA, 2014) pp. 2672–2680.

[40] T. Bolukbasi, K. Chang, J. Y. Zou, V. Saligrama, and A. Kalai, *Man is to computer programmer as woman is to homemaker? debiasing word embeddings,* CoRR **abs/1607.06520** (2016).

[41] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi, *Data decisions and theoretical implications when adversarially learning fair representations,* CoRR **abs/1707.00075** (2017).

[42] L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi, *Classification with fairness constraints: A meta-algorithm with provable guarantees,* CoRR **abs/1806.06055** (2018).

[43] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, *Fairness-aware classifier with prejudice remover regularizer,* in *ECML/PKDD* (2012).

[44] F. Kamiran, A. Karim, and X. Zhang, *Decision theory for discrimination-aware classification,* in *2012 IEEE 12th International Conference on Data Mining* (2012) pp. 924–929.

[45] I. Zliobaite, F. Kamiran, and T. Calders, *Handling conditional discrimination,* in *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*, ICDM '11 (IEEE Computer Society, Washington, DC, USA, 2011) pp. 992–1001.

[46] F. Kamiran, T. Calders, and M. Pechenizkiy, *Discrimination aware decision tree learning,* 2010 IEEE International Conference on Data Mining , 869 (2010).

[47] J. W. Vaughan, *Making better use of the crowd: How crowdsourcing can advance machine learning research,* Journal of Machine Learning Research **18**, 1 (2018).

[48] A. J. Berinsky, G. A. Huber, G. Lenz, and R. Michael Alvarez, *Evaluating online labor markets for experimental research: Amazon.com's mechanical turk,* Political Analysis **20**, 351 (2012).

[49] E. Pavlick, M. Post, A. Irvine, D. Kachaev, and C. Callison-Burch, *The language demographics of amazon mechanical turk,* Transactions of the Association for Computational Linguistics **2**, 79 (2014).

[50] P. Miller and A. Sønderlund, *Using the internet to research hidden populations of illicit drug users: A review,* Addiction (Abingdon, England) **105**, 1557 (2010).

[51] R. Epstein and R. E. Robertson, *The search engine manipulation effect (seme) and its possible impact on the outcomes of elections.* Proceedings of the National Academy of Sciences of the United States of America **112 33**, E4512 (2015).

[52] J. Kulshrestha, M. Eslami, J. Messias, M. B. Zafar, S. Ghosh, K. P. Gummadi, and K. Karahalios, *Quantifying search bias: Investigating sources of bias for political searches in social media,* CoRR **abs/1704.01347** (2017).

[53] B. Pan, H. Hembrooke, T. Joachims, L. Lorigo, G. Gay, and L. Granka, *In google we trust: Users' decisions on rank, position, and relevance,* Journal of Computer-Mediated Communication **12**, 801 (2007), /oup/backfile/content$_public/journal/jcmc/12/3/10.1111_j$.1083 – 6101.2007.00351.$x/2/jjcmcom$0801.$pdf$.

[54] R. White, *Beliefs and biases in web search,* in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13 (ACM, New York, NY, USA, 2013) pp. 3–12.

[55] J. Otterbacher, A. Checco, G. Demartini, and P. Clough, *Investigating user perception of gender bias in image search: The role of sexism,* in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18 (ACM, New York, NY, USA, 2018) pp. 933–936.

[56] P. Glick and S. Fiske, *The ambivalent sexism inventory: Differentiating hostile and benevolent sexism,* (1996) pp. 116–160.

[57] S. Tang, X. Zhang, J. Cryan, M. J. Metzger, H. Zheng, and B. Y. Zhao, *Gender bias in the job market: A longitudinal analysis,* Proc. ACM Hum.-Comput. Interact. **1**, 99:1 (2017).

[58] Z. Hu and J. Strout, *Exploring stereotypes and biased data with the crowd,* CoRR **abs/1801.03261** (2018).

[59] F. Durupinar, K. Wang, A. Nenkova, and N. Badler, *An environment for transforming game character animations based on nationality and profession personality stereotypes,* (2016).

[60] J. Attenberg, P. G. Ipeirotis, and F. J. Provost, *Beat the machine: Challenging workers to find the unknown unknowns,* in *Human Computation* (2011).

[61] A. Torralba and A. A. Efros, *Unbiased look at dataset bias,* in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11 (IEEE Computer Society, Washington, DC, USA, 2011) pp. 1521–1528.

[62] B. Green and Y. Chen, *Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments,* (2019) pp. 90–99.

[63] Z. Song, M. Wang, X. sheng Hua, and S. Yan, *Predicting occupation via human clothing and contexts,* in *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11 (IEEE Computer Society, Washington, DC, USA, 2011) pp. 1084–1091.

[64] W.-T. Chu and C.-H. Chiu, *Predicting occupation from images by combining face and body context information,* ACM Trans. Multimedia Comput. Commun. Appl. **13**, 7:1 (2016).

[65] M. Shao, L. Li, and Y. Fu, *What do you do? occupation recognition in a photo via social context,* in *2013 IEEE International Conference on Computer Vision* (2013) pp. 3631–3638.

[66] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg, *Smoothgrad: removing noise by adding noise,* CoRR **abs/1706.03825** (2017), arXiv:1706.03825 .

[67] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller, *How to explain individual classification decisions,* J. Mach. Learn. Res. **11**, 1803 (2010).

[68] M. D. Zeiler and R. Fergus, *Visualizing and understanding convolutional networks,* CoRR **abs/1311.2901** (2013), arXiv:1311.2901 .

[69] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, *Striving for simplicity: The all convolutional net,* in *ICLR (workshop track)* (2015).

[70] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, *Learning deep features for discriminative localization,* CoRR **abs/1512.04150** (2015), arXiv:1512.04150 .

[71] M. Sundararajan, A. Taly, and Q. Yan, *Axiomatic attribution for deep networks,* CoRR **abs/1703.01365** (2017), arXiv:1703.01365 .

[72] L. M. Zintgraf, T. S. Cohen, and M. Welling, *A new method to visualize deep neural networks,* CoRR **abs/1603.02518** (2016), arXiv:1603.02518 .

[73] K. Simonyan, A. Vedaldi, and A. Zisserman, *Deep inside convolutional networks: Visualising image classification models and saliency maps,* CoRR **abs/1312.6034** (2013).

[74] C. M. Bishop, *Training with noise is equivalent to tikhonov regularization,* Neural Comput. **7**, 108 (1995).

[75] P. Viola and M. Jones, *Rapid object detection using a boosted cascade of simple features,* (2001).

[76] P. Viola and M. Jones, *Robust real-time object detection,* in *International Journal of Computer Vision* (2001).

[77] D. G. Lowe, *Object recognition from local scale-invariant features,* in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Vol. 2 (1999) pp. 1150–1157 vol.2.

[78] N. Dalal and B. Triggs, *Histograms of oriented gradients for human detection,* in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1 (2005) pp. 886–893 vol. 1.

[79] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, *Rich feature hierarchies for accurate object detection and semantic segmentation,* CoRR **abs/1311.2524** (2013), arXiv:1311.2524 .

[80] R. B. Girshick, *Fast R-CNN,* CoRR **abs/1504.08083** (2015), arXiv:1504.08083 .

[81] S. Ren, K. He, R. B. Girshick, and J. Sun, *Faster R-CNN: towards real-time object detection with region proposal networks,* CoRR **abs/1506.01497** (2015), arXiv:1506.01497 .

[82] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, *SSD: single shot multibox detector,* CoRR **abs/1512.02325** (2015), arXiv:1512.02325 .

[83] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, *You only look once: Unified, real-time object detection,* in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016) pp. 779–788.

[84] J. Redmon and A. Farhadi, *YOLO9000: better, faster, stronger,* CoRR **abs/1612.08242** (2016), arXiv:1612.08242 .

[85] J. Redmon and A. Farhadi, *Yolov3: An incremental improvement,* CoRR **abs/1804.02767** (2018), arXiv:1804.02767 .

[86] H. Samet and M. Tamminen, *Efficient component labeling of images of arbitrary dimension represented by linear bintrees,* IEEE Transactions on Pattern Analysis and Machine Intelligence **10**, 579 (1988).

[87] M. B. Dillencourt, H. Samet, and M. Tamminen, *A general approach to connected-component labeling for arbitrary image representations,* J. ACM **39**, 253 (1992).

[88] E. Hossain and M. A. Rahaman, *Detection classification of tumor cells from bone mr imagery using connected component analysis neural network,* 2018 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE) , 1 (2018).

[89] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, *Gradient-based learning applied to document recognition,* Proceedings of the IEEE **86**, 2278 (1998).

[90] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition,* CoRR **abs/1512.03385** (2015), arXiv:1512.03385 .