UNIVERSITAT DE BARCELONA

FUNDAMENTAL PRINCIPLES OF DATA SCIENCE MASTER'S THESIS

Towards Fair Machine learning in Healthcare: Ensuring Non-Discrimination for Disease Prediction

Author:
Claudia Herron Mulet

Supervisor:
Dr. Polyxeni Gkontra
Dr. Karim Lekadir

A thesis submitted in partial fulfillment of the requirements for the degree of MSc in Fundamental Principles of Data Science

in the

Facultat de Matemàtiques i Informàtica

in collaboration with

BCN-AIM
ARTIFICIAL INTELLIGENCE IN MEDICINE

June 30, 2023

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica

MSc Fundamental Principles of Data Science

Towards Fair Machine learning in Healthcare: Ensuring Non-Discrimination for Disease Prediction

by Claudia Herron Mulet

Over the past few years, there has been a rise in the utilization of information and communication technologies (ICTs) and electronic health records (EHRs) within the healthcare system. This increase has led to a substantial gathering of medical data, opening up promising prospects for personalized medicine. Notably, one promising application is the creation of disease risk assessment tools, designed to precisely estimate an individual's predisposition to developing certain illnessess. These innovative tools empower healthcare professionals to conduct more targeted trials, closely monitor high-risk subjects, and implement timely interventions. However, as these systems start to be tested in real world scenarios, recent studies reveal that they might worsen off the situation of historically underprivileged groups in our society. These discriminatory biases might be caused by many reasons: unequal access to healthcare, false beliefs about biological differences, non-diverse datasets, machine learning (ML) models optimizing for the majority and disregarding underrepresented communities, etc. As a result, it becomes crucial to design and implement metrics and techniques to quantify and mitigate discriminatory biases.

In this work, we propose a comprehensive methodology that encompasses data wrangling, model evaluation, and the monitoring of both model performance and potential disparities. Building upon existing research on fairness in machine learning, we aim to adapt the fairness framework specifically for disease prediction, considering that some of the protected features also contribute to increased disease risk. Furthermore, we apply both in-processing and post-processing mitigation techniques to a classifier trained on a large-scale dataset. By experimenting with two diseases of increasing prevalence, Primary Hypertension and Parkinson's Disease, we seek to assess the effectiveness of these techniques in reducing discriminatory biases and ensuring equitable outcomes.

Acknowledgements

I would like to express my sincere gratitude and appreciation to all those who have supported and contributed to the completion of this master's thesis.

First and foremost, I would like to express my gratitude and appreciation to my supervisor, Polyxeni Gkontra, for her invaluable guidance, support, and expertise throughout the entire duration of this master's thesis. Her insightful feedback, and her exceptional mentorship have been instrumental in shaping the direction, methodology, and overall quality of this research. I would also like to extent my gratitude to my supervisor Karim Lekadir. His impecable leadership in directing the BCN-AIM research group, securing EU funding, and providing access to crucial datasets were instrumental in the project's overall success.

Moreover, I express my gratitude to all the members of the BCN-AIM research group for their outstanding contributions, which undoubtedly have a profound impact on the future of medicine in this rapid evolving digital world. I am especially thankful to Marina Camacho for generously offering her time and guidance during the initial phases of the project. I would also like to extend my appreciation to my fellow students in the group, namely Vien Dang Ngoc, Esmeralda Ruiz Pujadas, Joan Perramon, and Peter Brosten, with whom I had valuable conversations and exchanges of ideas.

During the course of these two years, I have met a lot of inspiring people. I would like to express my gratitude to the faculty members of Faculty of Mathematics and Computer Science at the University of Barcelona for their expertise and valuable feedback. Their passion for teaching and dedication to research have greatly enriched my learning experience. Furthermore, I want to transmit my personal gratitude to my classmates. Without cherishing the happy moments and sharing the difficult times in the company of such incredible people, this thesis would not have been possible.

Lastly, I want to express my gratitude to my friends. In particular I would like to thank my flatmates who have been there for me every step of the way, offering a shoulder to lean on. To my friends, both near and far, thank you for being my source of laughter and joy during the challenging times. Marta Soliño, I cannot find enough words to express my gratitude for your support throughout this demanding academic journey.

Finally, my deepest gratitude goes to my mother, my brother and my late father. Thank you for always believing in me and for being my constant source of love and motivation.

Thank you all.

Contents

Chapter 1

Introduction

With the advent of Big Data and digital health, there has been considerable enthusiasm surrounding the potential opportunities that emerging technologies can bring to the future of healthcare. By leveraging advanced algorithms and vast amounts of data, it becomes possible to build systems that can provide valuable insights into early detection, prognosis, and personalized treatment of multiple diseases [1], [2]. Such predictive models offer healthcare professionals the opportunity to identify individuals at high risk of developing certain conditions, enabling proactive interventions and preventive measures.

However, as these technologies have been refined and tested, concerns have emerged regarding their ethical, societal, and legal implications. An instance that garnered significant media attention was a study published in Science [3], which revealed that a widely utilized risk assessment algorithm for patient referral in the US exhibited discriminatory behavior against Black patients. The authors estimated that addressing this disparity could increase the percentage of Black patients receiving additional assistance from 17.7% to 46.5%. As a result, in recent years, researchers, organizations, and influential figures have emphasized the necessity of developing unbiased and fair Artificial Intelligence (AI) algorithms that transcend biases related to sex, age, ethnicity, and other population groups.

As a testament to this effort, the European Commission established a High-Level Expert Group in 2019 to formulate ethical guidelines for trustworthy AI [4]. These guidelines outline seven essential principles, including diversity, non-discrimination, and fairness, that all AI systems must adhere to in order to be deemed trustworthy. In this line, extensive research has been carried out in the field of machine learning to develop statistical metrics and methods for discrimination detection and mitigation [5] [6]. Still, there are multiple challenges for their general adoption. For example, there is lack of an universal definition of fairness, as different applications and scenarios might require different fairness constraints. Moreover, in some cases, when efforts are made to reduce bias and ensure fairness, it may introduce some inaccuracies in the predictions or decisions, which can lower the overall accuracy of the model [7].

In this study, we aim to critically review and enhance existing approaches for evaluating fairness and mitigating bias by implementing them in the context of disease prediction using machine learning models. Our objective is to develop a system that exhibits fairness, transparency, and explainability, with the potential for application across various diseases. In particular, we carry out experiments for Primary Hypertension (PH) and Parkinson's disease (PD), two conditions that show increasing trends in the last years ([8], [9]).

1.1 Objectives

The objectives of this master's thesis are focused on addressing several key tasks related to disease prediction. Firstly, we aim to implement machine learning models that can effectively solve the disease prediction task, achieving performance that improves the current state-of-the-art approaches for various diseases. We aim to develop a reusable library that tackles the different stages of the data science pipeline and that can set the basis for experimentation on multiple diseases. By developing these models, we aim to contribute to the advancement of disease prediction techniques.

In addition to developing accurate prediction models, we also seek to address the issue of discriminatory bias in the input data and the resulting prediction outcomes. Given that the medical domain often involves sensitive attributes that may also act as risk factors for disease development, it is crucial to assess and quantify the level of bias present. To achieve this, we will implement a variety of criteria and metrics specifically tailored to the medical context. These tools will enable us to effectively evaluate the presence and extent of unwanted biases in both the data used for training the models and the prediction results they produce.

If discriminatory biases are identified during the evaluation process, our next objective is to mitigate these biases using appropriate strategies. We will explore different bias mitigation techniques that aim to reduce disparities both during model training (known as in-processing techniques) and after the model has been trained (known as post-processing techniques). By applying these strategies, we aim to improve the fairness and equity of the disease prediction models, ensuring that they do not perpetuate or amplify existing biases.

In summary, this master's thesis aims to implement advanced machine learning models for disease prediction, develop metrics to evaluate unwanted bias in the medical domain, and apply bias mitigation techniques to enhance fairness and equity in the prediction process. By addressing these objectives, we aim to contribute to the improvement of disease prediction methods and promote equitable healthcare practices.

1.2 Report Structure

The document is organized into several chapters, each addressing different aspects described as follows.

The Background chapter serves as an introduction, providing an overview of metrics and mitigation techniques employed to achieve fairness in machine learning. It also delves into prior works on machine learning for disease risk prediction, with a specific focus on Parkinson's disease and Primary Hypertension. Furthermore, it summarizes previous studies conducted in the field of fair disease prediction, establishing the foundation for the subsequent chapters.

In the Methodology section, we present our proposed stages for achieving fair disease prediction. This section outlines the step-by-step process required to ensure fairness in disease prediction models. We provide detailed explanations of the models, methods, and techniques selected for this specific use case. Additionally, we describe the data used in our research, highlighting its characteristics and relevance to the study.

Moving on to the Results section, we provide a comprehensive analysis of the outcomes obtained at each stage described in the Methodology section. Specifically, we present detailed results for two specific diseases: Primary Hypertension and Parkinson's disease. By analyzing the results, we aim to assess the effectiveness and fairness of our proposed approach in predicting these diseases.

Finally, in the last chapter, we summarize the achieved objectives of our research. We provide a concise overview of the key findings, highlighting the main contributions. Additionally, we outline potential future directions for further research, identifying areas of improvement and expansion in order to advance the field and continue promoting fairness and equity in disease prediction models.

Chapter 2

Background

What is a fairness? According to the Collins dictionary, fairness is *the quality of being reasonable, right, and just.* Numerous authors have addressed these ethical concepts in the last centuries, from Plato and Aristotle to Philippa Foot, going through Kant and John Rawls. Fairness definitions often incorporate the notion of equality, such as "equality of treatment" (where goods are distributed evenly irrespective of individual differences) or "equality of opportunity" (where goods are distributed to ensure everyone has the same chances). However, various authors present diverse criteria for defining fairness. Consequently, this concept proves to be intricate, as it encompasses multiple interpretations not only from a philosophical standpoint but also from legal and technical perspectives.

In recent years, machine learning has emerged as a powerful tool for analyzing complex data sets and making predictions across various domains. In the field of healthcare, machine learning algorithms have shown immense potential for disease risk prediction, offering new avenues for early detection, intervention, and personalized treatment [2], [10]. However, as these algorithms become increasingly integrated into clinical practice, concerns related to fairness, and non-discrimination have come to the forefront [3][11].

In this chapter, we first present the main definitions and techniques in the field of fair ML. Then, we review the state of the art for disease prediction, focusing on two specific diseases: Primary Hypertension and Parkinson's disease. Finally, we include a summary of prior works in the area of fair disease prediction.

2.1 Fairness in machine learning

In the field of Data Science, **bias** refers to the systematic error or deviation of a model's predictions from the true or expected values. Fairness in machine learning is a newly emerged discipline that explores methods to prevent biases in data and models from treating individuals unfairly based on sensitive characteristics [6]. These sensitive characteristics are commonly referred as **protected attributes** or **socially salient characteristics**, and they describe groups of people that systematically experience social disadvantages (unprivilege). Common examples of protected attributes are sex, gender, sexual orientation, ethnicity, country of origin, age, disability and many more. In some countries, it is explicitly unlawful to discriminate based on these characteristics.

Unfair biases might appear because of many reasons: biased device measurements, historical human-biased decisions encoded in the data sets, missing data that do not represent the target population, minimization of objective functions that leave out minorities, etc [12]. For this reason, it becomes crucial to incorporate bias evaluation methodologies in the different steps of the machine learning pipeline.

There are two main frameworks for tackling unfair bias in machine learning: **statistical** and **causal** methodologies [6]. Statistical methodologies focus on identifying and measuring bias within the data and model outputs. They typically involve analyzing the distribution of data and the predictions made by the model to detect any disparities across different demographic groups. Statistical methods often include fairness metrics, such as Disparate Impact or Equalized Odds, to quantify and assess the degree of bias present [6].

On the other hand, causal methodologies aim to understand the underlying causal relationships that contribute to bias in machine learning systems. They go beyond statistical associations and seek to identify the direct causes of bias. Causal methodologies involve techniques such as causal inference and counterfactual analysis to explore how changes in input variables affect the outcome and uncover potential sources of bias. By understanding the causal mechanisms, these methods enable interventions to mitigate bias directly at its root causes [6].

Both statistical and causal methodologies have their strengths and limitations. In this project, we focus on the statistical framework as there are more available tools and resources from which to start. There are several python libraries, such as IBM's AIF360 [13] or Microsoft's Fairlearn [14], with fairness evaluation metrics and bias mitigation techniques that can be included into the disease prediction pipeline. In the following subsections, we will describe these metrics and mitigation techniques.

2.1.1 Fairness metrics

Fairness metrics are statistical criteria that measure the presence or lack of discrimination. These criteria are written as a function of random variables such as the protected attribute(s), the target variable in the classification task, the classifier score or the classifier label. Different metrics enclose different definitions of what is considered *fair*, and in some occasions, these criteria cannot hold simultaneously [5].

Usually, fairness metrics are divided into three big groups:

- **Independence**: we say that a classifier satisfies independence when the random variables representing the protected attribute, A, and the classifier score, S, are marginally independent. The core idea behind independence fairness is that the decision or outcome should be independent of the sensitive attribute. In other words, knowing someone's sensitive attribute should not provide any additional information about the decision or outcome beyond what is already known about them through non-sensitive attributes. If $\hat{Y}$ is the random variable for the classifier labelling and A has domain a_1, a_2, independence implies:

$$P(\hat{Y}|A = a_1) = P(\hat{Y}|A = a_2)$$

- **Separation**: for separation to hold, the random variables for the score, S, and the protected attribute, A, must be conditionally independent given the true label Y. It aims to address situations where there are significant imbalances or disparities in outcomes between different groups based on sensitive attributes. In the binary classification scenario, separation implies equal true and false positive rates:

$$P(\hat{Y} = 1|Y = 1, A = a_1) = P(\hat{Y} = 1|Y = 1, A = a_2)$$
$$P(\hat{Y} = 1|Y = 0, A = a_1) = P(\hat{Y} = 1|Y = 0, A = a_2)$$

- **Sufficiency**: sufficiency is satisfied when the true label Y is conditionally independent to the protected attribute, A, provided the score S. In practice, sufficiency enforces parity in positive/negative predictive values. Mathematically,

$$P(Y = 1|S = s, A = a_1) = P(Y = 1|S = s, A = a_2)$$

  Sufficiency is closely related to the concept of calibration. We say that a model is calibrated when we can interpret scores as probabilities that match true data distributions. It can be shown that a classifier that satisfies calibration for all protected groups then it is sufficient [5].

These criteria are often measured by taking the difference or the ration of the probabilities for each groups, and a certain threshold is set to identify a system as unfair. In this project, we explicitly use the following metrics:

- **Disparate Impact Ratio (DIR)**: a metric related to the measure of independence. Mathematically:

$$DIR = \frac{P(\hat{Y}|A = a_1)}{P(\hat{Y}|A = a_2)}$$

- **Average Odds Error (AOE)**: related to the measure of separation.

$$AOD = \frac{|FPR_{A=a_1} - FPR_{A=a_2}| + |TPR_{A=a_1} - TPR_{A=a_2}|}{2}$$

- **Equal Opportunity Difference (EOD)**: a relaxation of the separation criteria that only asks for equality of true positive rates (equal recalls for groups):

$$EOD = TPR_{A=a_1} - TPR_{A=a_2}$$

- **False Positive Rate Difference (FPRD)**: also a relaxation of the separation criteria.

$$FPRD = FPR_{A=a_1} - FPR_{A=a_2}$$

An interpretation of these metrics, and a justification for their selection is presented in Chapter 3.

2.1.2 Discrimination mitigation

Once the fairness evaluation is performed and significant biases are identified, the question arises as to whether it is possible to mitigate these biases without compromising performance. Various approaches have been proposed and explored, aiming to address the issue of bias and ensure fair and equitable outcomes for all individuals involved.

One such approach, known as **fairness through unawareness**, suggests that ignoring sensitive attributes or withholding certain information during the decision-making process can lead to fairer outcomes. Fairness through unawareness is an intuitive concept, rooted in the belief that if an algorithm does not have access to sensitive attributes, it will be less likely to discriminate based on those attributes. This approach, in theory, seems promising, as it appears to promote fairness by treating all individuals as equals, regardless of their personal characteristics. However, this method is unable to address the root causes of discrimination and biases present in the underlying data used to train AI systems. Even without explicit access to sensitive attributes, AI models can still indirectly learn and infer such attributes from other correlated features present in the data. Consequently, discrimination can persist and even worsen despite attempts to ignore or hide sensitive attributes during the decision-making process [5].

Therefore, to achieve meaningful discrimination mitigation, it is crucial to employ more sophisticated methods that go beyond fairness through unawareness. Mitigation strategies are often categorized into three big groups depending on what moment of the machine learning pipeline are applied.

Pre-processing techniques modify the training data before fitting machine learning models. The idea in this case is to remove prior disparities encoded in the data. For example, one strategy consists of creating synthetic data or oversampling minority groups when real data is not available [15]. Another strategy that aims to reduce disparate impact is known as **Feature modification** [16]. This method applies transformation techniques to modify sensitive attributes. The goal is to preserve the information encoded in these features while making them less discriminate. Other methods impose demographic parity with **re-weighting** techniques [17] or by learning **fair representations** of data [18].

In-processing techniques are discrimination mitigation methods that directly modify the learning algorithms themselves to promote fairness and mitigate bias during the training process. Some methods add a **regularization** term to the objective function that enforces fairness constraints [19]. The **reductions approach** [20], decomposes fair classification as a sequence of cost-sensitive classification problems, subject to fairness constraints. This methodology captures two techniques: Exponentiated Gradient and Grid Search Reduction. Exponentiated Gradient Reduction uses iterative updates based on gradients to find the best parameters, while Grid Search Reduction performs an exhaustive search over a grid of hyperparameters to find the optimal configuration. Finally, another technique worth mentioning is **Adversarial Debiasing** [21], that adopts the adversarial framework for training. The weights of the debiasing network are adjusted to minimize its ability to predict the sensitive features accurately, while the classifier model is trained to maximize its accuracy. This adversarial optimization process seeks to create a balance between accurate prediction and reducing the influence of sensitive features.

Ultimately, **post-processing** methods adjust the model output to satisfy fairness criteria. One of the advantages of these techniques is that they can be applied to black-box models as neither the training data or the model specifications are needed. The simplest post-processing method consists of assigning different classification thresholds based on protected attributes in order to achieve parity in acceptance rates [5]. This is the intuitive idea behind the **Equalized Odds classifier** [22], a method that solves a linear program to equalize false and true positive rates. Another method is the **Rejection Option Classifier**, that favours the underprivileged group in the neighborhood of the decision threshold, where the uncertainty is high [23].

2.2 Machine learning for disease risk prediction

Machine learning for disease risk prediction involves the application of machine learning algorithms to assess an individual's likelihood of developing a particular disease. As opposed to the task of **diagnosis**, in disease prediction, the input data comes from healthy individuals. The goal is to estimate the probabilities of developing target diseases in the future.

There are common challenges in the field of ML disease prediction. One of these potential issues is **data imbalance**. Most of disease data sets exhibit class imbalance, where the number of instances belonging to different classes (e.g., healthy vs. diseased) is highly unequal. This imbalance can have repercussions on the model's performance, as the minority class receives insufficient attention during the training process.

Another challenge in disease prediction is ML **generalization** [24]. Machine learning models need to generalize well to unseen data to be useful in real-world clinical settings. However, healthcare data sets often have specific characteristics and may not fully represent the diversity of patients and healthcare systems. Models developed on one population or data set may not perform as well on different populations or in different healthcare settings, leading to challenges in achieving broad generalizability.

Finally, another requirement in ML for disease prediction is **explainability** [25]. Developing explainable machine learning models is critical in the medical domain to build trust and understand the factors influencing predictions. Many advanced machine learning algorithms, such as deep neural networks, are considered black boxes, making it difficult to interpret their decision-making process. Balancing the need for accurate predictions with the requirement for explainability is an ongoing challenge.

Up to our knowledge, there are no multidisciplinary works that tackle potentially any disease in the same methodology. This is probably because for developing these models, it is often required a high understanding and domain expertise of the disease being studied. Some works tackle sets of diseases, such as cardiovascular diseases, CVDs, [2][26], but most of the works focus on a single one. In the following subsections, we go through the prior works on Parkinson and Primary Hypertension ML prediction, two diseases that we will experiment with.

2.2.1 Parkinson's disease

Parkinson's disease (PD) is a neurodegenerative disorder that affects over 10 million people in the world [27]. Characterized by the loss of dopamine-producing cells in the brain, PD leads to a wide range of motor and non-motor symptoms, significantly impacting the quality of life for individuals diagnosed with the condition and their caretakers. Nowadays there is no cure for this disease, but there exist treatments to mitigate the effect of the symptoms. As the incidence of Parkinson's rises significantly with age, and people are living longer in high-income countries, the prevalence of Parkinson's is set to rise dramatically in the future.

Parkinson's disease is typically diagnosed after the onset of symptoms. For this reason, developing risk models is highly important to enable early-detection leading to access to clinical trials and personalized medication. However, this is a challenging task because Parkinson's disease can appear because a combination of complex causes, from genetic predisposition to exposure to environmental factors, such as pesticides [28]. As a result, it is an active area of research in medicine.

Most of the works from the ML community on Parkinson tackle the diagnosis case, where the individual is subject to having the disease in the moment of evaluation. Moreover, the majority of studies use speech data and voice recordings, while in this project we have we have readily available variables regarding the exposome. The exposome encompasses all the exposures an individual experiences throughout their lifetime and can be divided into external and internal factors. The external exposome includes various elements such as lifestyle, mental health, sociodemographic factors, early-life influences, environmental aspects, and physical measurements [29]. On the other hand, the internal exposome is associated with genetics and blood biochemistry [30]. Other works use MRI data for the diagnosis task [31]. Furthermore, the development of non-ML risk scores is also in a very initial state in comparison to other diseases such as CVDs [32]. For this reason, Parkinson prediction using ML represents a very promising research opportunity.

One of the first risk scoring system for Parkinson was developed by Alastair J. Noyce et al. [33]. They build a risk score (PREDICT-PD) based on information collected by surveying healthy individuals (age, gender, smoking status, coffee intake, PD family history, etc). To calculate the risk, they first compute the age-related risk as described in [34], and they increase/decrease the final output by multiplying the added risk for each of the other variables, as reported in prior studies. Then, they compare the PREDICT-PD scores in the studied data set to other medical risk scores, that measure variables known to be predictors for PD: the University of Pennsylvania Smell Identification Test (UPSIT) score, a smell identification test; REM Sleep Behavior Disorder Screening Questionnaire (RBDSQ); and the BRAIN test, bradykinesia akinesia incoordination test, for measuring keyboard tapping speed. They showed through statistical analysis that individuals with higher PREDICT-PD scores had worse results in these medical tests. However, they do not provide performance results as the ground truth (PD development) was yet to be observed.

In 2020, Jacobs et al. [35] continue with this line of work and build polygenic risk scores, a type of risk estimation method based on genetic information, with logistic regression models, for predicting PD. They also evaluate the original PREDICT-PD algorithm with UK Biobank data set, reporting an Area Under the Receiver Operating Characteristic curve (AUC-ROC) of 0.76. Although they present a methodology where they study gene-environment interactions, their proposed method does not significantly improve prior-work performances.

As we can see, the literature for developing ML risk scores is scarce and it mostly targets the diagnosis task, for a heterogeneity of data types.

2.2.2 Primary Hypertension

Hypertension is a chronic medical condition characterized by elevated blood pressure levels persistently exceeding the normal range. It is estimated that over 1.3 billion people have this disease, and in fact, it is one of the major causes of premature death [36](according to the WHO). Hypertension is diagnosed if, when it is measured on two different days, the systolic blood pressure readings on both days is $\geq$ 140 mmHg and/or the diastolic blood pressure readings on both days is $\geq$ 90 mmHg[36]. Primary Hypertension, in particular, refers to high blood pressure that develops gradually over time without any identifiable cause or underlying medical condition. It is often influenced by a combination of genetic, lifestyle, and environmental factors.

Silva et al. [37] present a review of the prior works on Hypertension prediction. The authors report AUC-ROCs between 0.766 and 1, although some of the included works have been criticized for potential data leakage. The best performing algorithms are XGBoost (XGB), Random Forest (RF) and Support Vector Machines (SVMs). In this study, there is a variety of data sets with different sizes and imbalance ratios.

Liu Yu et al. [38] perform a study to use nutritional ingredients intake as predictors for Hypertension. They report an AUC-ROC of 0.904 with the XGBoost model evaluated on the China Health and Nutrition Survey data set, consisting of 28 features, including age, gender and dietary information. These results are computed in a balanced scenario, where there is an equal proportion of healthy and hypertensive subjects. Renuka Patnaik et al. [39] combine systolic blood pressure measurements from past 10 years together with medical known risk factors to train a Support Vector Machine. They achieve an AUC-ROC of 0.9. Again, the data set used for evaluation is balanced.

Other works take into account the natural class imbalance and develop ML models with a slightly lower performance. For example, AlKaabi et al. [40] compare logistic regression, random forest and decision trees on Qatar Biobank data set, that presents a 15% of positive instances. Their models used a total of eleven variables, consisting of seven non-clinical and non-invasive factors (age, sex, education, employment, tobacco use, physical activity, and adequate consumption of fruits and vegetables), and four easily obtainable clinical variables (maternal history of Hypertension, history of diabetes, history of cholesterol, and abdominal obesity). The best performing model was Random Forest with an AUC-ROC of 0.87. However, the authors acknowledge that this study is limited due to the small data set size (over 1000 subjects) that is not representative enough of the Qatar population and the inherent selection bias induced by the volunteer selection of study subjects.

To sum up, there are prior works in the area of Hypertension prediction with ML that experiment with a variety of data sets and models. There is a lot of heterogeneity in the data set sizes and imbalance ratios, as well as the nature and number of features taken into account. We did not find prior works that evaluate the prediction task on the UK Biobank data set, but other works with similar features achieve high performance.

2.3 Fair disease prediction

Historically, medical research has often focused primarily on male subjects, leading to a lack of understanding about how certain diseases, symptoms, and treatments manifest in women. This gender bias in research can result in delayed diagnoses, misdiagnoses, and inappropriate treatments for women [41]. We see similar consequences for historically oppressed groups, such as black people [42]. As white cisgender males were seen as the default, unprivileged groups were underrepresented in the data sets and biases in disease risk scoring systems where not quantifiable. Fortunately, in the last years, there has been a rising concern on developing fair machine learning models and some works evaluate and mitigate the biases found in their models.

On the one hand, some prior works focus on adapting the fairness framework to the healthcare use case. Jinying Xu et al. [43] analyze what biases can appear in the medical context, define fairness metrics and mitigation strategies in terms of patient outcome. The work by Obermeyer et al. [3] reveals that an algorithm deployed to manage medical resources based on health risk was biased favoring white patients. This model was in reality predicting healthcare costs, and due to unequal access to care between blacks and whites, black patients where considered less sick than their actual state.

One the other hand, other more practical works focus on evaluating and mitigating discriminatory biases. For instance Pfohl et al. [11] revealed that procedures penalizing differences between prediction distributions across groups led to a degradation of performance metrics within groups. Additionally, the impact on fairness measures varied across experimental conditions. The study suggests that researchers developing predictive models for clinical use should go beyond algorithmic fairness and critically engage with the broader sociotechnical context of machine learning in healthcare.

Camacho et al. [44] compare different machine learning models for schizophrenia spectrum disorders prediction, evaluating both regular performance metrics such as AUC-ROC or F1 and fairness metrics such as Statistical Parity Difference or Disparate Impact. They achieve a fair XGBoost classifier with an AUC-ROC of 0.82. Ngoc Dang et al. [45] evaluate and mitigate found biases in depression predictions. Authors from the same group present a prediction model for CVDs and diabetes type 2 where they explicitly evaluate fairness metrics on the model outputs [46].

The field of fairness in healthcare research presents an exciting research opportunity, emphasizing the necessity of incorporating evaluation and mitigation strategies into all risk prediction models as a default practice. As evident from the ongoing studies, it is crucial to seize this opportunity and integrate these findings into our existing frameworks.

Chapter 3

Methodology

In this chapter, we dig into the methodology we followed to ensure non-discrimination for disease prediction. By considering fairness evaluation during the different stages of the classical data science pipeline, we aim to produce fair machine learning models that are robust, interpretable and transparent.

3.1 Overview: fair disease prediction pipeline

In this section, we present an overview of the proposed fair disease prediction pipeline while in the following sections, we provide details on each of the steps that compose the pipeline itself. Figure 3.1 presents the proposed pipeline for disease risk prediction while ensuring fairness.

The process starts in the data collection phase, followed by a data wrangling step, where we extract and combine data from multiple files. Once that we obtain a feature matrix and a target label, we perform a data bias evaluation to both report possible model limitations due to data quality and to prevent the training procedure if we identify discrimination. After selecting the best model through grid search and cross validation, we perform model evaluation in terms of performance, fairness and explainability. If the performance is not good enough, we may improve the model selection phase. If the model does not satisfy a set of fairness criteria, we apply post-processing bias mitigation strategies. Again, we evaluate these fairness criteria, and we might consider training the model with bias mitigation in-processing techniques. Once that we achieve a model that satisfies all our evaluation constraints, we can proceed with the deployment of this model.

3.2 Data

In this section we dive into the details of the data used in this project. We go from the data description, to the technical issues regarding the data wrangling and we explain our proposed method to evaluate fairness in the input data, prior to the development of any machine learning model.

3.2.1 UKBiobank

UK Biobank is a large-scale biomedical research database. It is one of the most significant health research initiatives worldwide and aims to improve the prevention, diagnosis, and treatment of various diseases.

The UK Biobank project started in 2006 and involves the collection and analysis of extensive health-related data from around 500,000 volunteers, aged between 40

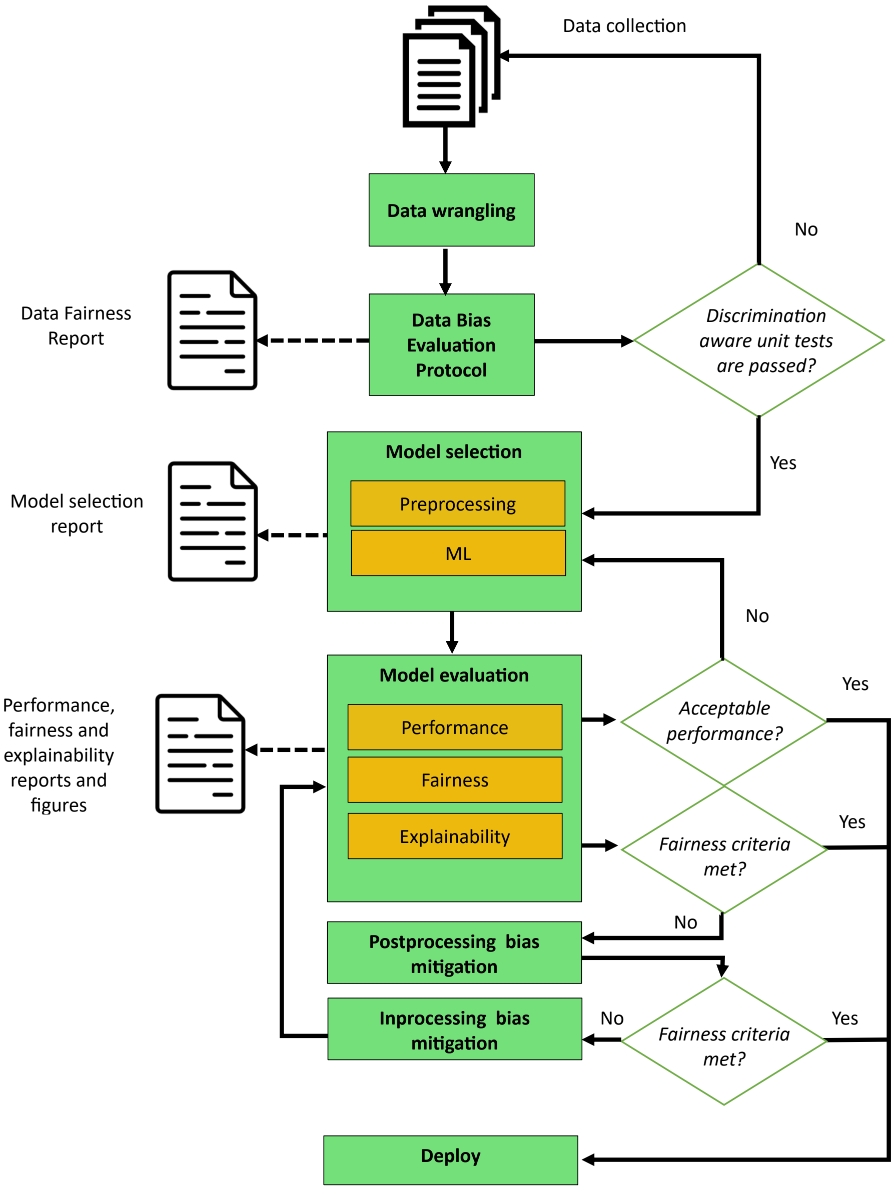


FIGURE 3.1: Fair disease prediction pipeline

and 69 years, from across the United Kingdom. Participants have undergone a comprehensive assessment that includes providing biological samples (such as blood, urine, and saliva) and completing detailed questionnaires about their lifestyle, medical history, and behaviors.

The primary goal of UK Biobank is to facilitate scientific research by providing an extensive resource for studying the complex interplay between genetics, environment, and lifestyle factors in the development of diseases. The project focuses on a broad spectrum of diseases, including cancer, cardiovascular disorders, neurodegenerative conditions, mental health issues, and many others. The data collected by UK Biobank are made available to approved researchers globally, enabling them to investigate a wide range of health-related questions. However, this data set is not open to the general public due to data privacy concerns. For more information, visit the official web page [47].

This data set is specially well-suited for studying disease risk as it allows to follow up the evolution of participants through time. UK Biobank has the potential to contribute significantly to medical research, improve disease prevention strategies, and lead to the development of more effective treatments. However, it is worth mentioning some of the limitations of this data set. First of all, the participants in the UK Biobank study are aged between 40 and 69 years. This **narrow age range** may limit the generalizability of findings, particularly for diseases or conditions that primarily affect other age groups, such as pediatric diseases or age-related conditions. Secondly, there is a **self-section bias** as individuals who volunteer to participate in research studies like UK Biobank may differ from the general population in terms of their health status, lifestyle, and socioeconomic factors. Moreover, there is a **lack of ethnic diversity**, as we will see later on. The UK Biobank cohort consists mainly of individuals of European descent. While UK Biobank covers a broad range of diseases and health-related factors, some specific diseases or conditions may have limited information available.

3.2.2 Data wrangling

The UK Biobank is a complex and large scale database. For the purpose of this project, I was granted access to a total of 326 csv files containing tabular medical information from the application 65769. In order to be able to analyze this data and build a machine learning system, we needed to rearrange it. We can divide the data wrangling process into two big steps: creating the feature data set and creating the target disease label.

Creating the feature data set

A data-field in the UK Biobank repository serves as the basic unit of data and represents the outcome of a single question, measurement, or result. Each data-field is stored in a tabular file, where each row represents a patient and multiple columns represent multiple measurements in different moments of time, or **assessment visits**. Not all fields have the same number of columns. In this project, we have considered the data measurements collected during the first visit to the assessment center, where the participant completed personal questionnaires and provided biological samples.

Each data-field is encoded by a unique identifier, and each measurement follows specific data coding. In order to keep track of the variable names, as well as the possible categories for each variable and other relevant information, it was crucial to build together with the feature data set, a metadata data set where we could store relevant information for the data processing. In order to do so, we retrieved metadata from two sources: a public showcase data dictionary available in UK Biobank web page, and a scrapper specially built to retrieve relevant information from the searching UK Biobank engine ([48]). Finally, we also manually annotated the category of each of the fields, resulting in a total of 19 categories (for example, *Physical activity*, *Diet summary*, *Physical measure summary*), following the guidelines for data categorization.

This metadata was also used to transform categories encoded with numbers into their original string label. Also, we performed a basic data cleaning for removing columns and rows with more than a 90% of missing values. So, in the end we were able to assemble a data set with information for 477182 patients distributed into 156 features. We include a brief description of these features, together with its data type and medical category in the Appendix A.

Creating the target label

In order to create the target vector indicating the future presence of the disease, we first gather all the ICD10 codes per patient. ICD-10 (International Classification of Diseases, 10th Revision) is a system of medical coding that is used to classify and code diagnoses, symptoms, and procedures in healthcare [49]. It is a standardized system developed by the World Health Organization (WHO) and is widely used internationally for statistical and billing purposes. Once we have all the disease codes per patient, we select the patients with a positive diagnosis for the studied disease and we filter out all patients that have been diagnosed prior to the first assessment visit. We do this in order to ensure that we are tackling the disease prediction task rather than the diagnosis task, as the feature data set is a picture of the patient information during the first assessment visit. In the end, we end up with a boolean vector indicating for each patient if they have developed the target disease in the future.

3.2.3 Protected Attributes

The first step in performing any fairness analysis is identifying the protected attributes and the unprivileged groups that are potentially subject to discrimination. In Table 3.1 we summarize the features identified as protected and the respective privileged groups. Most of these attributes coincide with the ones recognized by Great Britain's Equality and Human Rights Commision (EHRC) [50]. Townsend deprivation index is not a personal characteristic, but a measure of material deprivation within a population. Socioeconomic status is not legally considered as a protected characteristic but we will take it into account in this study because it is the cause of social inequalities. In the same manner, we will report the biases associated with Obesity.

Regarding the feature **sex**, it might sometimes refer to biological sex and sometimes to gender identity. According to the UK Biobank description, this feature is defined as: *Sex of participant. Acquired from central registry at recruitment, but in some cases updated by the participant. Hence this field may contain a mixture of the sex the NHS had recorded for the participant and self-reported sex.* Also, in relationship to the variable **race**, we have considered the three UK Biobank features that might encode it, according EHRC [50].

Finally, socioeconomic status, age and obesity are encoded in the form of continuous variables. As a result, in order to categorize a participant into the privileged or unprivileged group, it becomes necessary to set a threshold to divide the population in two. According to Joseph Rowntree Foundation[1], 20% of UK population is at poverty risk. For this reason, we chose the top 20% of the Townsend Deprivation indexes as an indicator of low socioeconomic status. For the age variable, we consider privileged those individuals less than 65 years old. This threshold is the mean retirement age in the UK, and it has been shown that elderly people might be subject to ageism [51]. Obesity is defined by the body mass index (BMI), which is calculated

[1] https://www.jrf.org.uk/data/overall-uk-poverty-rates

Protected Attribute	Feature name	Field ID	Privileged	Unprivileged
Sex or Gender Identity	Sex	31	Male	Female
Race	Ethnic background	21000	White (British, Irish, Any other white background)	Non-White Mixed (White and Caribbean, Black African, Asian, Any other mixed background) Asian or Asian British (Indian, Pakistani, Bangladeshi, Any other Asian background) Black or Black British (Caribbean, African, Any other Black background) Chinese Other ethnic group Do not know Prefer not to answer
Race	Skin Color	1717	Very fair, fair, light olive	Dark olive, Brown, Black, Do not know, Prefer not to answer
Race	Country of birth	1647	England, Wales, Scotland, Northern Ireland, Republic of Ireland	Elsewhere, Do not know, Prefer not to answer
Socioec. status	Townsend deprivation, index at recruitment	189	Below percentile 80%	Above percentile 80%
Age	Year of birth	34	Younger than 65 years old	Older than 65 years old
Obesity	BMI	21001	BMI over 30	BMI under 30

TABLE 3.1: Protected attributes and privileged groups

using one's weight (in kilograms) by the square of one's height (in meters). According to the BMI classification, a BMI equal to or greater than 25 is overweight, and a value equal to or greater than 30 is obese. For this reason, we have considered the last threshold to differentiate the unprivileged group.

3.3 Data Bias Evaluation Protocol

Following the guidelines proposed by d'Alessandro et al. [52], we decide to implement **discrimination aware unit tests**. The idea is to evaluate and document a series of metrics related to the protected attributes, and impose fairness criteria, before training any machine learning model. In this project, we aim to adapt this general fairness framework to the specific use case of disease prediction, and as we will see, this will influence the metrics chosen.

The two discrimination tests that we suggest are the following:

1. Measure the Disparate Impact Ratio (DIR) between privileged and unprivileged groups, and check that it is in the neighborhood of the theoretical DIR for the specific disease. In medical context, it can be the case that privileged

and non-privileged groups have different disease incidence rates based on biological or socioeconomic reasons. For example, the risk of developing Parkinson's disease (PD) is twice as high in men compared to women. As a result, we expect that the DIR is in the neighborhood of 0.5.

2. Calculate the support for each level of every protected attribute, in order to identify which protected groups are more prone to statistical estimation errors. The support represents the proportion within the overall population and aids in understanding the potential vulnerability of specific protected groups to such errors. For example, it might be the case that there are no positive samples for a minority unprivileged group. This represents a risk, as the model might not be able to make predictions for this specific group and more importantly, we will not be able to evaluate it. In order to pass the test, we can impose a minimum number of samples per group.

If any of these criteria is not met, we can decide whether to proceed with the ML pipeline applying mitigation strategies or stop the process until more data is gathered.

The last step in the protocol is to note any features that display a moderate to strong correlation with the protected attribute. As we saw, fairness through unawareness is not enough. It can be the case that other highly correlated features with the protected attributes (or combinations of them) become proxies. As a result, it becomes crucial to keep track of these features when analyzing the model.

The result of the data bias evaluation protocol is generated automatically so that the user can decide what steps to take before implementing any pre-processing or machine learning model.

3.4 Model selection

In this project, we refer to model as a combination of pre-processing steps and a machine learning model. This is not the typical definition, as traditionally data is first cleaned, scaled and transformed and later, the best machine learning model is chosen. However, this approach is introducing extra assumptions by generating an intermediate data representation. Moreover, for some models different pre-processing steps might be better than others. By introducing the pre-processing inside a sklearn pipeline, we can cross-validate pre-processing parameters such as for example the data imputation strategy or the data scaling method.

Regarding the pre-processing step, we designed two pipelines depending on whether the machine learning model applied was tree-based or not:

- Tree-based models: we encode string categories with integers, and unseen categories in testing times are encoded as missing values. Categorical missing values are replaced with the most frequent category and numerical missing values with the mean.
- Non tree-based models: instead of numerically mapping categories with integers, we perform one hot encoding producing as many new features as many categories minus one, per categorical column. With one-hot encoding, the expanded feature space can exacerbate the curse of dimensionality. High-dimensional spaces require more data to effectively cover the feature space, making it harder for tree models to find meaningful patterns and relationships. That is why we do not apply this transformation for tree models. Apart from

the data imputation strategies, we also perform data min-max scaling. Decision tree-based models, such as random forests or gradient boosting machines, are invariant to monotonic transformations of the features and do not necessarily require scaling. However, for most other non tree-based models, scaling the data is beneficial for better model performance and stability.

Regarding the machine learning model, the system is ready to accept any classifier that implements the standard sklearn fit-predict interface.

Finally, we encapsulate the pre-processor and ML model together in a pipeline that is the input for a grid search with 5 fold cross validation strategy. For each model, we store the one that achieves higher f1 score. The results of the grid search, together with a detailed report on performance evaluation metrics is documented automatically. Also, the top-10 most important features are listed in the model selection report.

3.5 Model evaluation

We refer to model evaluation to the process of both measuring the prediction power of the model and the algorithmic bias that might produce. In this section, we summarize the metrics used for predictive performance and bias evaluation.

3.5.1 Performance analysis

In order to evaluate the classifier, we measure traditional performance metrics such as the AUC-ROC and metrics derived from the confusion matrix, such as precision, recall and F1-score. In prior works [37], the usual reported metric is the AUC-ROC, but we realized that models with high AUC-ROC were achieving close to zero precision. In imbalanced scenarios where the majority class heavily outweighs the minority class, the classifier may achieve a high AUC-ROC by simply predicting the majority class correctly. However, the F1 score takes into account both precision and recall, providing a balanced evaluation that considers the performance on both positive and negative instances. For this reason, we decided to use F1 score both to select the best hyperparameters through cross-validation and to select the probability threshold for declaring a sample as positive.

Moreover, we decided to evaluate the performance in terms of the imbalanced specific metrics such as the **balanced accuracy (BA)** and **Matthews correlation coefficient (MCC)** [53] [54].

Balanced accuracy is a metric that takes into account the imbalanced nature of the data set by calculating the average of the class-wise accuracies. It is defined as the average of the sensitivity (true positive rate) for each class. The Matthews Correlation Coefficient (MCC) takes into account true positives, true negatives, false positives, and false negatives and produces a value between -1 and 1, where 1 represents a perfect prediction, 0 indicates a random prediction, and -1 represents a completely inverse prediction.

3.5.2 Model Fairness analysis

In order to evaluate the classifier in terms of fairness, we must decide which statistical criteria we want to focus on, as different metrics encode different fairness definitions that might be incompatible. In the task of risk assessment, we propose defining fairness as equal performance in terms of true and false positive rates. In the disease prediction application, different population might experiment different incident rates because of biological or socioeconomic reasons. As a result, in this context we are not so interested in achieving **independence**. For fairness evaluation, we decide to focus on **separation** metrics, which cannot be achieved simultaneously with independence, and in some specific scenarios with sufficiency. The concept of separation draws attention to a crucial question: Who bears the consequences of misclassification? The violation of separation underscores the reality that distinct groups face disparate costs resulting from misclassification. Concern arises when higher error rates align with historically marginalized and disadvantaged groups, exacerbating the additional harm inflicted upon them. In particular, we evaluate the model in terms of:

- **Equal Opportunity Difference (EOD)**. Our goal is to achieve close to zero, or at least positive EOD to ensure True Positive Rate (TPR) parity.
- **False Positive Rate Difference (FPRD)**. Our goal is to achieve close to zero, or at least negative FPRD to ensure False Positive Rate (FPR) parity.
- **Average Odds Error (AOE)**. This metric gives an average TPR and FPR disparity, in absolute value. We target close to 0 values.

These metrics are determined by the chosen threshold. For this reason, we inspect the desegregated ROC curve per group, in order to check if there is any intersection between the curves for privileged and unprivileged groups that would automatically ensure separation. Moreover, we compute the disparate impact per group in order to compare that the model is producing predictions at the same rate as observed in training data.

3.5.3 Model Explainability

Finally, the last step of the evaluation consists of inspecting the explainability of the machine learning (ML) disease prediction model. Explainability plays a crucial role in healthcare applications as it enhances the trust and acceptance of the model by clinicians and patients alike. By understanding the underlying factors and features that contribute to a particular prediction, clinicians can gain valuable insights into the decision-making process of the model, aiding in their clinical judgment and treatment planning. Furthermore, explainability allows patients to comprehend the reasoning behind the predictions, enabling them to make informed decisions about their health and potential interventions. Therefore, in this evaluation step, we aim to assess the explainability of the ML disease prediction model and its impact on the overall usefulness and acceptance of the system.

In particular, we rank the feature importance and we analyze the categories that the features belong to (primary demographics, test results, etc). Moreover, we use a state-of-the art explainability method, the SHapley Additive exPlanations (SHAP) method [55].

3.6 Bias mitigation

If the bias evaluation phase determines that there are significant biases, the first step would be to apply post-processing techniques and re-evaluate the model. These techniques are the less computationally expensive, as they only require access to model prediction or scores. In large scale data sets, we want to avoid retraining the model from scratch, however, if post-processing techniques do not achieve good performance or do not mitigate biases, then we consider retraining the model including in-processing mitigation methods.

3.6.1 Post-processing techniques

For the purposes of this work, we explored the available post-processing techniques from the AIF360 sklearn interface. In particular, we experimented with Rejection Option classifier and Calibrated Equalized Odds classifier. The first technique requires defining a decision threshold and a margin, in order to define the neighborhood where uncertain samples will favor the unprivileged group. These parameters can be previously determined (for example we can use the threshold that maximizes F1 score) or they can be selected following a cross-validation strategy to select the model with better fairness metrics.

Moreover, we developed a post-processing strategy inspired by the work of Solon Barocas et a. [5], where we use group-specific thresholds to achieve a target acceptance rate. To do so, given a target average FPR rate, the algorithm computes the FPR rates for a grid of thresholds per group and applies for each group the threshold that achieves the input FPR. This algorithm could be extended to achieve parity in other metrics, such as TPR parity (Equal Opportunity). We will refer to this method as Separation Classifier.

3.6.2 In-processing techniques

In the same manner, the choice of in-processing techniques was conditioned on the available implementations of AIF360 interface. In particular, there are two implementations of the Reductions Approach: Exponentiated gradient reduction and Grid Search reduction [20]. These methods allow to include fairness constraints during training in order to reduce biases. In particular, we are interested in including separation constraints, such as FPR and TPR parity, as explained previously. One advantage of these methods is that they allow considering multiple protected attributes simultaneously, as opposed to the tested post-processing techniques. The main disadvantage is the computational cost that add. We need to retrain from scratch our models, now taking into account constraints in the minimization problem.

If these techniques are not enough in order to achieve fairness, or they produce a considerable performance drop, we might consider the viability of deploying such models.

Chapter 4

Results

After defining the methodology to ensure fair disease prediction, we validate it by considering two specific diseases: Primary Hypertension and Parkinson's Disease.

4.1 Primary Hypertension (PH)

As highlighted in the Background section, Primary Hypertension has a significant global impact, affecting a large population of individuals worldwide. After performing data wrangling, we discover that it is between the top-10 more popular diseases in UKBiobank. Specifically, we consider a sample of 440K subjects from which 70K will suffer from this disease. This represents a **15% of positively labeled samples**. In order to train and evaluate the disease prediction model, we split the data into three sets:

- Training data set: it represents 75% of the initial data set (330561 subjects). It is used for grid search hyper parameter selection with 5-fold cross validation and for fitting the final best model.
- Validation data set: one third of the remaining 25% of data (36729 subjects). It is used for selecting the best model among the candidates and for selecting the best decision threshold.
- Testing data set: the remaining data (73459 subjects). We use it for evaluating the final performance of the model.

All three data sets where splitted using **stratified sampling**, a common approach when dealing with imbalanced data sets that ensures that the distribution of the target variable or class labels remains representative in the sampled data.

4.1.1 Data Bias Evaluation Protocol

Before training any model, we inspect the input data and we run discrimination aware unit tests. First, we measure the **disparate impact ratio** per group to compare it to the expected group prevalence. Figure 4.1 shows this metric[1]. We see that unprivileged groups for age and BMI have higher incidence, a fact that is coherent with expert knowledge as these characteristics are considered as risk factors by the National Health Service (NHS) [56]. Also, we see that men have higher base rates than women and people with low socioeconomic status have a higher prevalence, as expected [57]. Regarding race, we see a slightly higher incidence in unprivileged

[1] Colors used in the plots of this work have been selected to accommodate individuals with colorblindness, aiming to enhance accessibility and ensure a meaningful visual representation for all readers.

groups, which could be explained by the presence of people with black African or black Caribbean ascendance, who are at a higher hypertension risk [56], but a more disagregated study would need to be conducted in order to inspect the effect of other ethnicities. As a summary, we can say that the training data is broadly coherent with medical domain expertise.

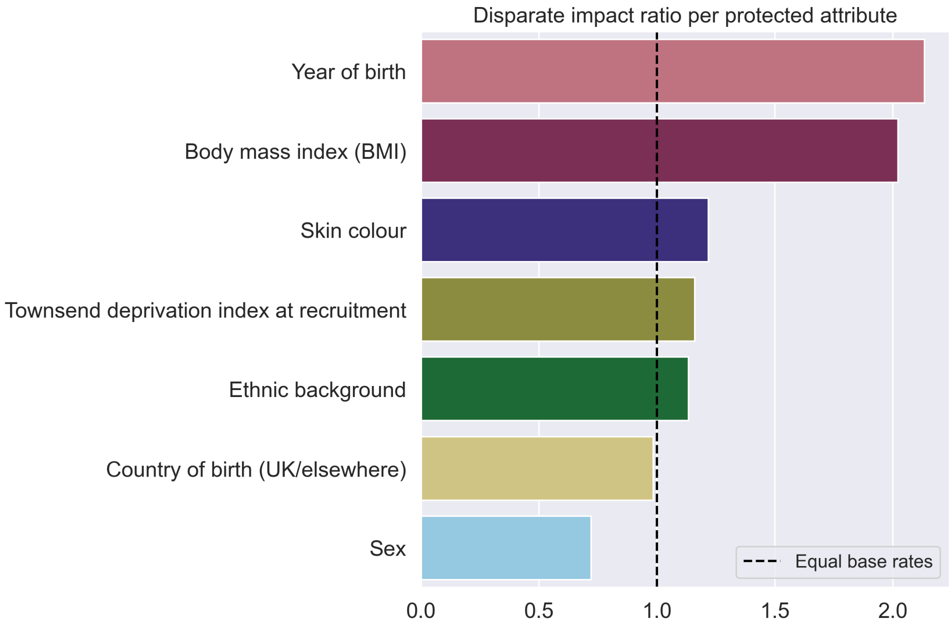


FIGURE 4.1: Disparate Impact Ratio in input data

Then, we inspect the **support** for each protected group and report the percentage of subjects that will develop the disease. In table 4.1, we learn that all protected groups except for females, are minorities, and hence they are more prone to statistical estimation errors. Furthermore, we observe that only 1% of the training samples belong to unprivileged subjects w.r.t. ethnic background that will develop hypertension. Up to this point, we should decide if this highly imbalanced data set regarding socially salient properties is good enough to proceed with the training or if we should stop the process and gather more data. As the second option is not possible in the scenario of this project, we will proceed with the training and bias evaluation for research purposes.

Field Name	**Support (%)**
Sex	44.98%
Ethnic background	94.46%
Skin color	93.26%
Country of birth	92.04%
TDI[2]at recruitment	79.88%
Year of birth	90.63%
BMI	77.40%

(A) Percentage of privileged class per protected attribute

Privileged?	**Future PH?**	**Support (%)**
No	No	4.56%
No	Yes	0.98%
Yes	No	79.76%
Yes	Yes	14.70%

(B) Percentage of total population w.r.t disease distribution by ethnic background

TABLE 4.1: Support for each level of protected attributes

[2] Townsend Deprivation Index

Finally, we report the top-10 most correlated features to the protected ones. For example, we see in figure 4.2 that body measurements are highly correlated to Sex, as expected. As a result, we should take into consideration these features as potential proxies for Sex. The rest of the plots for the remaining protected features are included in the the Appendix A.

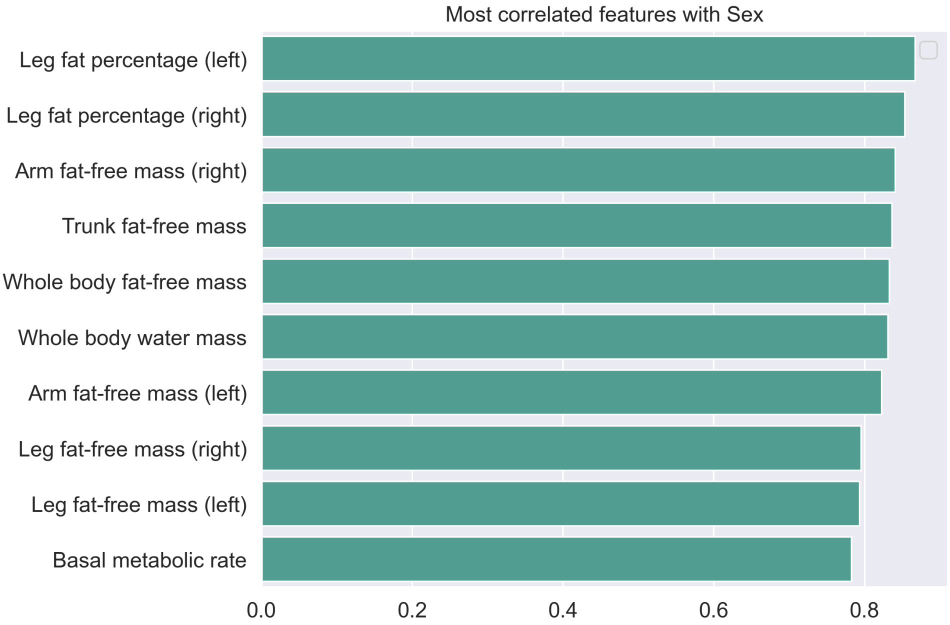


FIGURE 4.2: Highly correlated features to sex

4.1.2 Machine learning model selection

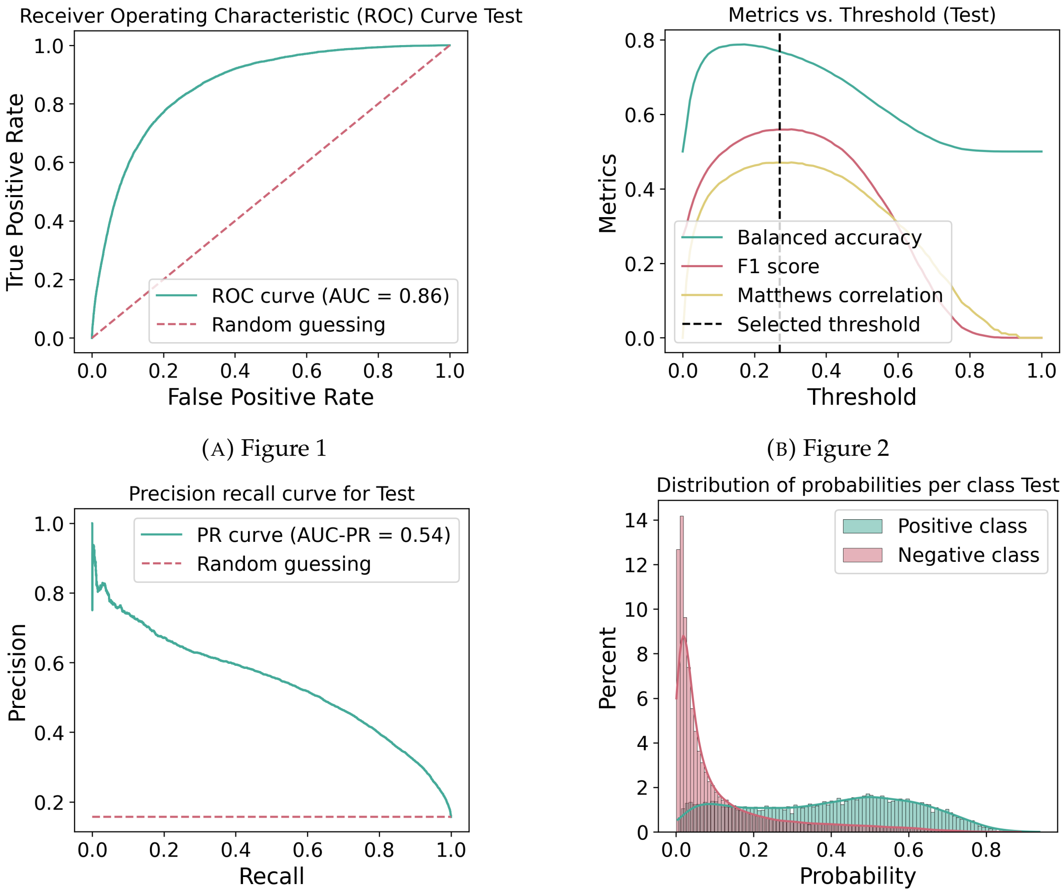
Table 4.2 presents the performance evaluation results for various ML models. Random Forest, SVM, and XGB have been recognized in previous works as state-of-the-art models for different data sets. Logistic Regression is a straightforward yet influential model extensively utilized in the medical field. It is noteworthy that all models generally exhibit comparable results. However, the CatBoost classifier stands out as the top-performing model, surpassing the others in all metrics, making it the preferred choice.

Model	AUC-ROC	F1	BA	MCC
Logistic Regression	0.859247	0.549077	0.758484	0.457635
Random Forest	0.851529	0.541325	0.754548	0.448032
SVM	0.8595	0.5509	0.7661	0.4605
XGB	0.839086	0.536963	0.742932	0.443135
CatBoost	**0.863682**	**0.556138**	**0.766806**	**0.466701**

TABLE 4.2: Primary Hypertension model performance metrics on UKBiobank validation set (15% of positive samples)

4.1.3 Performance evaluation

Now, we evaluate the best classifier more thoroughly on the test data. Figure 4.3 presents a summary of the main evaluation metrics. Subfigures 4.3a and 4.3c depict the ROC curve and the PR curve, respectively. Their behaviour is far from random guessing so we conclude that the classifier is indeed learning from data. Subfigure 4.3b presents different metrics as a function of the decision threshold. Also, we have marked the final selected threshold. We see that this threshold generalizes well for unseen test data as we also obtain the best F1 and MC possible. However, this maximum is not achieved for balanced accuracy, where lower thresholds would result in a higher metric. Finally, subfigure 4.3d shows the probability distributions for the two classes. We see that although the negative class is more concentrated towards close to zero probabilities, there are also negative samples all along the x-axis. Moreover, the positive class distributed quite uniformly along different probability regions. As a result, the two classes are not separable and the classification task is therefore limited.



(A) Figure 1

(B) Figure 2

(C) Figure 3

(D) Figure 4

FIGURE 4.3: Catboost classifier performance evaluation for Primary Hypertension prediction

Finally, in figure 4.4 we include the confusion matrix for the best classifier. We see that although the majority of the negative and positive samples are well classified, there are a significant number of test errors. In this particular case, with a precision of 48% and a recall of 67%, it indicates that the model can correctly identify a significant portion of the positive instances (recall), but it also generates a **relatively high number of false positives** (low precision). The specificity of 86% suggests that the model performs well in identifying negative instances.

It is important to note that the decision threshold for these results was chosen to maximize the F1 score, which balances precision and recall equally. Therefore, depending on the specific disease and with the advise of domain experts, adjustments to the decision threshold might be necessary to optimize the model's performance for the desired outcome.

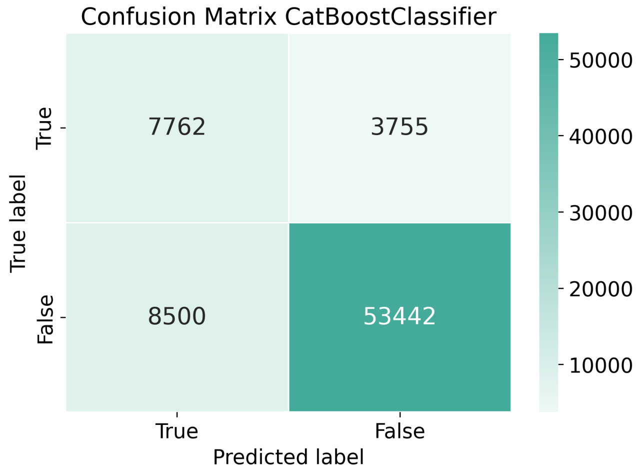


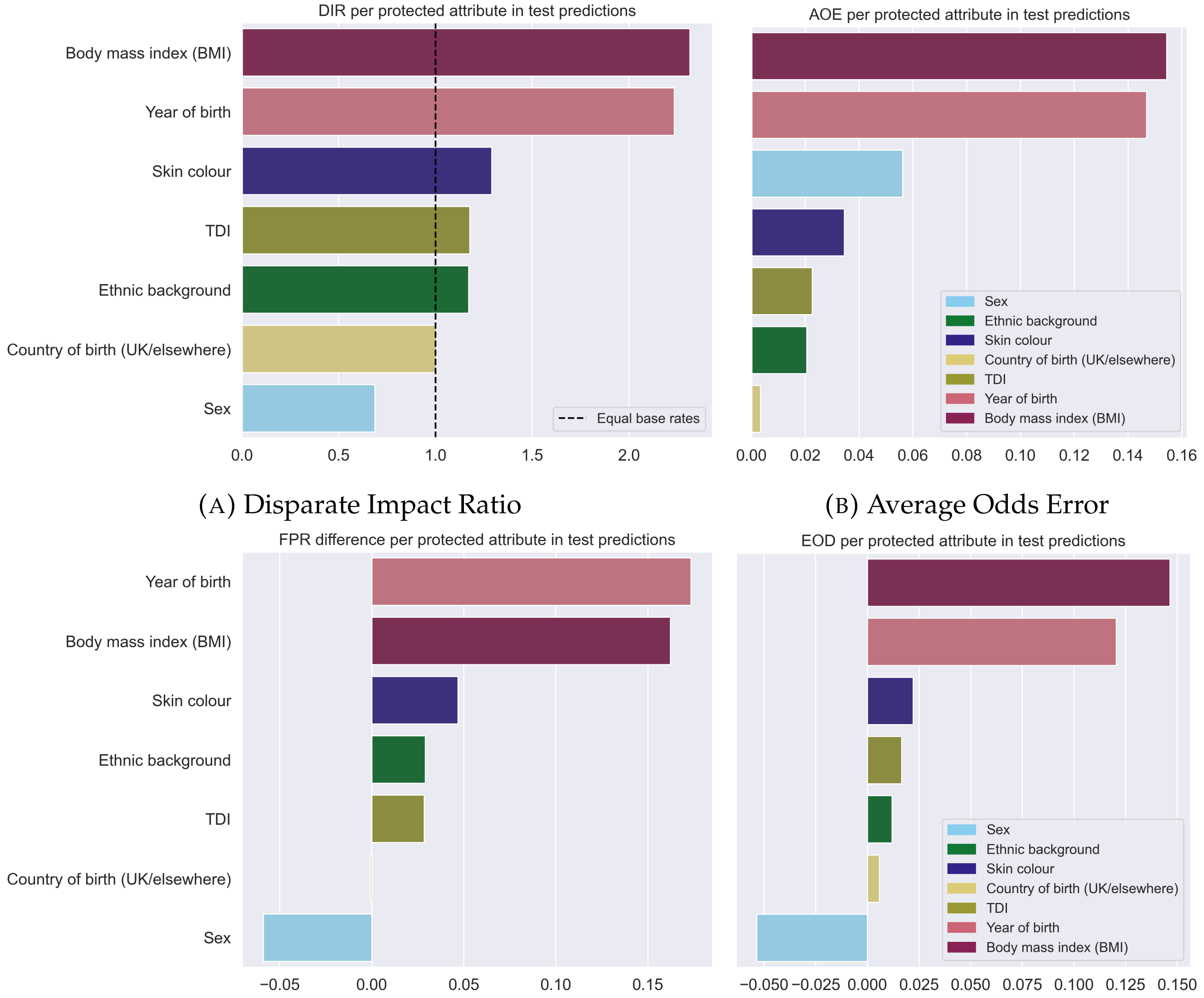
FIGURE 4.4: Confusion matrix for Catboost model for Primary Hypertension prediction

4.1.4 Model Fairness Evaluation

Figure 4.5 summarizes the mentioned model fairness metrics. First of all, we measure the disparate impact by groups in order to check that the predicted incidence rates (Subfigure 4.5a) are similar to those found in the data bias evaluation protocol (Figure 4.1). Effectively, we again see that both elderly people and people with obesity are the ones at a higher risk of developing PH according to Catboost, with respect to the privileged groups. Similar conclusions can be drawn for the rest of protected features.

Moving on to the rest of the metrics, that are related to the measure of separation, Subfigure 4.5b shows the Average Odds Error (AOE) per protected attribute in the test set. We see that those features that presented higher DIR are also the ones that achieve higher disparities in terms of FPR and TPR. These are *BMI*, *Year of birth* and *Sex*. However, notice that the maximum AOE is around 0.16. A value of zero indicates equality of odds and a value of one indicates maximum inequality of odds. For this reason, discrimination biases found in this model are rather small.

To gain insight on the nature of these disparities, we plot the FPR difference (Subfigure 4.5c) and the Equal Odds Difference or TPR difference (Subfigure 4.5d). We observe slight bias in terms of FPR difference for +65 individuals, and people with obesity, as a positive value indicates disadvantage for the privileged group. Also, we see slight bias in terms of Equal Opportunity Difference, where negative values indicate disadvantage for females.



(A) Disparate Impact Ratio

(B) Average Odds Error

(C) False Postive Rate Difference

(D) Equal Opportunity Difference

FIGURE 4.5: Model fairness evaluation metrics in test predictions

As separation fairness metrics are completely influenced by the decision threshold selected, we might wonder if it is possible to choose a threshold that instead of maximizing F1 performance, we would reach parity in terms of AOE. In order to do so, we inspect the disagregated ROC curves for the top-3 features with higher disparities (Figure 4.6). If we were able to find a point where both curves intersected, then, we would be able to select a threshold that achieves equal TPR and FPR performance for protected groups. However, as we see in this figure, neither of the ROC curves for privileged and unprivileged groups cross, so we will not be able to achieve complete separation without applying more sophisticated techniques. Later in this work, we apply bias mitigation methods with the objective of reducing disparities for *BMI*, *Year of birth* and *Sex*.

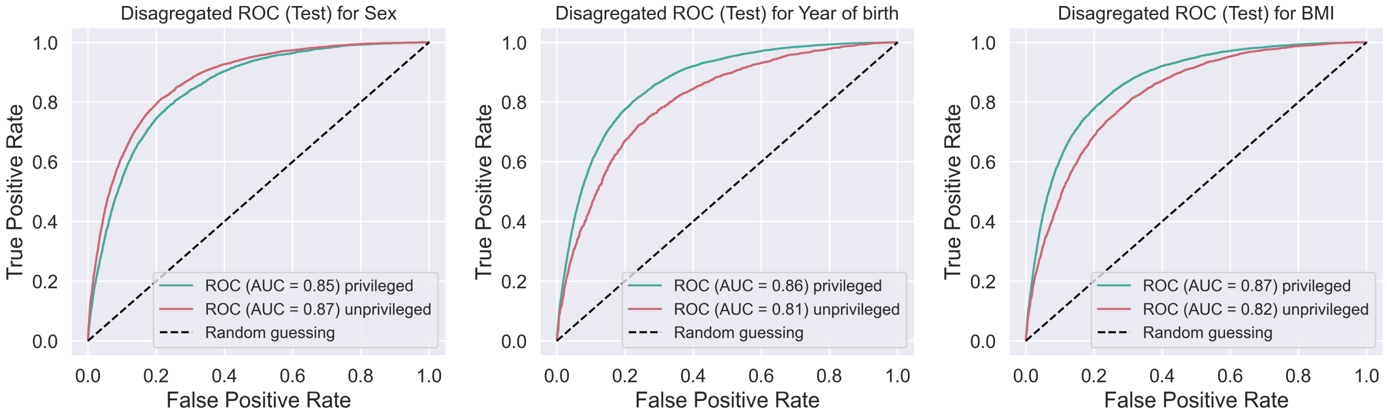


FIGURE 4.6: Disagregated ROC curves in terms of Sex, Year of birth and BMI for test data

4.1.5 Explainability

In this section, we provide an overview of the explainability techniques employed to interpret the results of disease prediction models. First of all, we analyze the **feature importance** provided by the Catboost model. Features with high importance scores indicate that they strongly influence the predictions made by the model. In Table 4.3 we present as summary of the Top-10 more important features in the best trained model. Moreover, we include the medical category to which these features belong to.

The first important result that we see is that diastolic and systolic blood pressure are between the most important features. These are automated readings taken to the subjects during the first assessment visit. High systolic or diastolic blood pressures can be very informative regarding the future development of the Primary Hypertension but at the same time, they are not sufficient for diagnosing it. Recall that according to the WHO, for the diagnosis of PH, repeated measurements in different days are required. This is because temporary high blood pressures can be caused by many reasons such as anxiety, stress, pregnancy or other non-chronic conditions. For this reason, we include these variables in the training set. In a similar fashion, the top most important variable is taken into consideration. This feature is self reported by the patient, and it is multicategorical. It can take values such as *angina*, *stroke* and *high blood pressure*. We do not have enough information with this data in order to assess if patients have been misdiagnosed with or without Primary Hypertension, or if these measurements correspond to other causes for high blood pressure. In the discussion section, we develop further this idea.

Moreover, we see other important results such as having *Year of birth* in the third place. This is coherent with the fact that age is considered a medical risk factor. Other important variables are highly correlated to protected features such as *Waist circumference* to BMI, and obesity is another risk factor for PH. Finally, it is interesting to comment that the top 10 most important features belong to a heterogeneous group of medical categories, making this disease being predicted by taking into account information from different data gathering procedures.

We can go deeper into the model explainability by computing the **SHAP values** for the test set (Figure 4.7). Beeswarm plots show in the x-axis the SHAP value. Higher values indicate a higher contribution to the model output. The color of the plot measures the original feature value. We also see that dots pile up to show density. In this way, we are able to observe that effectively high diastolic and systolic

Top-k	Feature	Medical Category
1	Vascular/heart problems diagnosed by doctor	Self-reported medical conditions
2	Systolic blood pressure, automated reading	Physical measure summary
3	Year of birth	Primary demographics
4	Number of treatments/medications taken	Self-reported medical conditions
5	Home area population density - urban or rural	Geographical and location
6	Waist circumference	Physical measure summary
7	Diastolic blood pressure, automated reading	Physical measure summary
8	Cystatin C	Biochemistry and Haematology
9	Biobank assessment centre	Primary demographics
10	Overall health rating	Self-reported medical conditions

TABLE 4.3: Feature importance of Catboost model

blood pressures increase the SHAP values. Similarly, taking many medications or having high Cystatin C values will increase the probability of developing future PH.

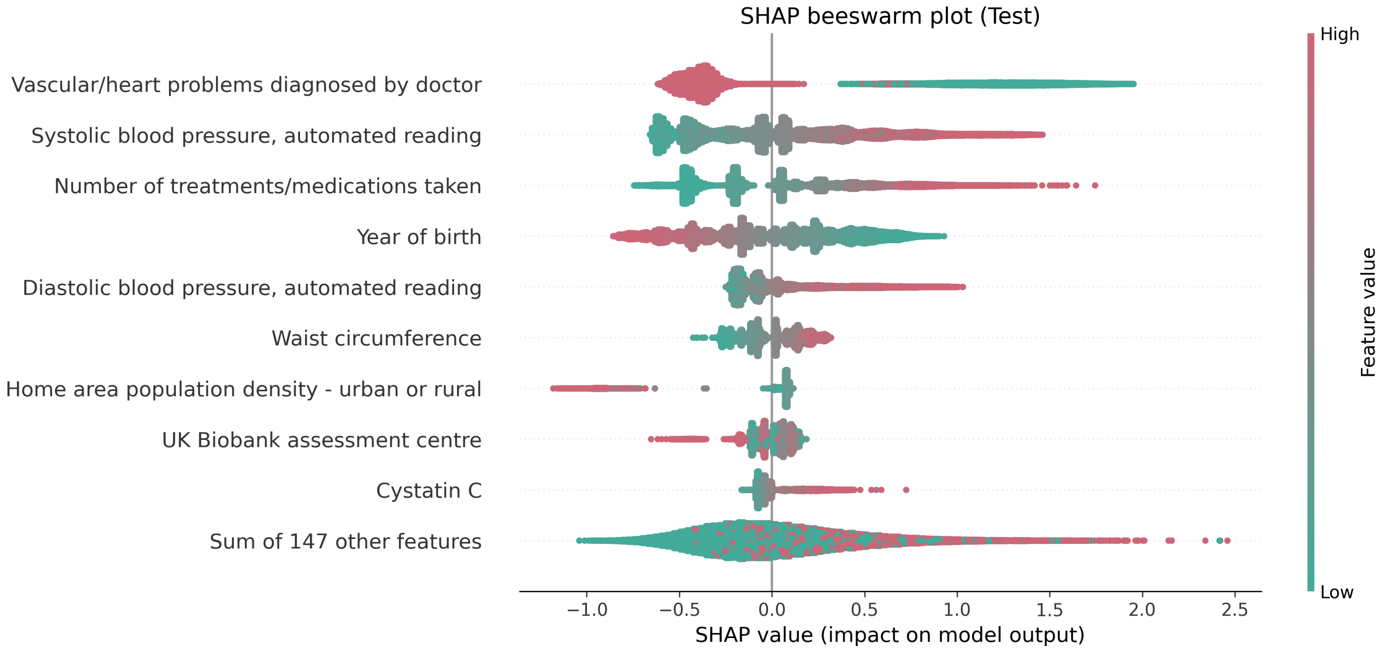


FIGURE 4.7: Beeswarm SHAP plot for test set

4.1.6 Bias mitigation results

As explained in Chapter 3, we first apply post-processing techniques as they are the less computationally costly. In Table 4.4 we include a summary of the main performance metrics, as well as average odd errors for the selected features, and particularly the metrics that we target to improve. This is, **for Sex, we aim to reduce Equal Opportunity differences**, and for **BMI and Age FPR differences**. We experimented with several techniques: our implementation of Separation Postprocessor for FPR difference parity both for age and BMI; AIF360 sklearn interface for Rejection Option Classifier using our previously selected F1-optimized threshold; AIF360 sklearn interface for Rejection Option Classifier with cross validation for threshold selection based on balanced accuracy and fairness metrics; and, AIF360 sklearn interface for Calibrated Equalized Odds postprocessor.

In terms of performance, we observe that our CatBoost baseline is better than the other methods, except for balanced accuracy, where Rejection Option Classifier with threshold selection for Sex Equal Opportunity and balanced accuracy maximisation achieves better results. In terms of fairness metrics, the method that more effectively reduces Equal Opportunity differences for Sex is surprisingly CalibratedEqualizedOdds adjusted for Year of birth. However, this method has a great performance trade-off, as F1, balanced accuracy and MCC are close to the metrics a random classifier would achieve. The rest of the postprocessing methods that improve Equal Opportunity in terms of Sex, experience the same limitations. It is worth highlighting the case of RejectOptionClassifier F1-threshold for Sex, as we see that this method is achieving the highest EOD and it is causing a disproportionate mitigation, where we go from a -0.05 to a 0.19 EOD. This is a common pattern for all variants of the Rejection Option Classifier, as this method is increasing the positive rate prediction for unprivileged groups. Regarding, FPR difference we see that Separation postprocessor successfully reduces disparities both for BMI and Age, while maintaining acceptable performance and reducing Average Odds Error. As a result, we claim that this method is the most effective for postprocessing bias mitigation in this application.

In addition to the aforementioned limitations, it is important to note that as of the current date, the AIF360 sklearn interface does not support bias postprocessing techniques that address multiple protected attributes simultaneously. Consequently, we are only able to address one feature at a time and rely on the hope that fairness metrics improve for the remaining attributes. We strongly believe that adopting a more intersectional approach would promote fairness on a broader scale.

In contrast, in-processing methods offer the advantage of mitigating bias across multiple protected levels. In Table 4.5, we present the results obtained from a sample of 30,000 patients. These methods rely on a base estimator, and we chose Logistic Regression as our base model due to its simplicity while achieving comparable results to Catboost. These decisions were primarily driven by limited computational resources. As part of our future work, we plan to conduct more extensive experiments to further explore these techniques. Upon reviewing the results, we observed similar limitations to those found in post-processing techniques. Although fairness improvements were achieved, they came at a significant performance cost, rendering the model ineffective.

Name	F1	BA	MCC	AOE Sex	AOE BMI	AOE Age	EOD Sex	FPRD BMI	FPRD Age
Catboost Baseline	**0.556**	0.767	**0.467**	0.058	0.168	0.146	-0.051	0.175	0.171
Separation Postprocessor BMI	0.545	0.757	0.452	0.061	0.063	0.174	-0.057	**0.0**	0.201
Separation Postprocessor Year of Birth	0.547	0.759	0.455	0.06	0.172	0.056	-0.053	0.179	**-0.001**
RejectOptionClassifier F1-threshold BMI	0.537	0.747	0.443	0.054	0.358	0.138	-0.054	0.333	0.148
RejectOptionClassifier F1-threshold Year of birth	0.541	0.738	0.449	0.049	0.149	0.362	-0.054	0.136	0.371
RejectOptionClassifier F1-threshold Sex	0.543	0.759	0.45	0.125	0.165	0.162	0.19	0.182	0.2
RejectOptionClassifier EO Sex	0.516	**0.782**	0.434	0.056	0.187	0.204	0.077	0.257	0.302
RejectOptionClassifier AO Year of birth	0.068	0.517	0.146	0.004	0.02	0.092	-0.008	0.004	0.022
RejectOptionClassifier AO BMI	0.083	0.52	0.154	0.004	0.064	**0.038**	-0.007	0.013	0.013
CalibratedEqualizedOdds BMI	0.269	0.576	0.266	0.016	0.294	**0.038**	-0.02	0.107	0.027
CalibratedEqualizedOdds Year of birth	0.171	0.544	0.212	**0.002**	**0.009**	0.346	**-0.001**	0.009	0.133
CalibratedEqualizedOdds Sex	0.349	0.608	0.323	0.167	0.093	0.106	-0.285	0.053	0.061

TABLE 4.4: Primary Hypertension model performance metrics on validation set after bias mitigation with postprocessing techniques

Name	F1	BA	MCC	AOE Sex	AOE BMI	AOE Age	EOD Sex	FPRD BMI	FPRD Age
LogisticRegression Baseline	**0.518**	**0.75**	**0.42**	0.073	0.21	0.179	-0.075	0.219	0.226
ExponentiatedGradientReduction FPR Parity	0.179	0.54	0.128	**0.004**	**0.006**	**0.01**	0.008	**0.001**	0.005
ExponentiatedGradientReduction TPR Parity	0.24	0.56	0.159	0.006	0.027	0.021	**0.005**	0.034	0.025
GridSearchReduction TPR Parity	0.218	0.524	0.042	0.502	0.26	0.028	-0.509	-0.218	0.044
GridSearchReduction FPR Parity	0.17	0.451	-0.074	0.634	0.308	0.033	0.586	-0.345	**0.003**

TABLE 4.5: Primary Hypertension model performance metrics on validation set after bias mitigation with in-processing techniques

4.1.7 Discussion

In this section, we will discuss and summarize the results of our study, examining their implications in the context of our research objectives. First of all, we have analyzed the results for the Data Bias Evaluation Protocol, where we spotted different **Disparate Impact Ratios coherent with previous medical knowledge**. Also, we observed low support of most of the protected groups. Precisely, those groups that experienced higher differences in terms of DIR, **people with obesity, older than 65 years old and females, were the ones the ML model predicted with higher Average Odds Error**. Surprisingly, we did not find significant model biases in those minorities in terms of race and socioeconomic status, having similar performance for privileged and unprivileged groups.

In terms of model performance, the **CatBoost classifier achieves a comparable AUC-ROC to those reported in previous studies** [37]. However, conducting a comprehensive comparison becomes challenging due to the lack of evaluations on imbalanced test sets in most of the referenced works, unlike our study. The exception is the work by AlKaabi et al. [38]. They reported an AUC-ROC of 0.87 and a F1 of 81.6% using Random Forest in a scenario with a 15% class imbalance using the Qatar Biobank data set. It is important to note that their work's primary limitation is the small sample size of only 1000 individuals, from which the majority belong to highly educated and affluent population, in contrast to the substantial population of almost 500,000 subjects in the UK Biobank. However, thanks to this work we acknowledge that it might be possible to improve our model in terms of precision and recall scores, maybe by using techniques such as feature selection.

During our analysis, we examined model explanations and identified that the model was utilizing **prior indicators of high blood pressure as predictors for the future development of Primary Hypertension** (PH). In this study, we constructed the target variable based on subjects diagnosed with PH subsequent to the initial assessment visit. Therefore, we excluded individuals with a positive diagnosis at the time of data collection, so in principle we discard data leakage. It is important to note that high blood pressure can have various causes other than Primary Hypertension and other studies presented in the survey by Silva et al. [37] also include these readings as predictors.

However, since Primary Hypertension typically lacks symptoms in its early stages, it is frequently under diagnosed. In fact, the World Health Organization estimates that less than half of adults with hypertension are aware of their condition and receiving treatment, thereby increasing the potential consequences of this disease [57]. Consequently, it is possible that some of the training samples included individuals labeled as normotensive (negative for PH) who are, in reality, unaware that they have the disease. This hypothesis could explain the relatively high false positive rate observed in our results. As Primary Hypertension cannot be diagnosed based on a single blood pressure measurement, and our analysis is based on the data available during the first assessment visit, we acknowledge the need for future work to remove individuals from the study who exhibit high blood pressure levels in subsequent medical checkups, even if they have not received a formal diagnosis from a doctor, in order to account for this possible under diagnosis bias.

Finally, we applied bias mitigation techniques to reduce *Sex*, *Age* and *BMI* disparities. However, we encountered challenges with certain in-processing methods and some post-processing mitigation techniques, such as Calibrated Equalized Odds or the Rejection Option classifier with threshold optimization. These methods exhibited a **significant fairness-accuracy trade off**, making the resulting models less effective. Interestingly, we observed that the Rejection Option classifier, even with an F1-optimized threshold, exacerbated biases. This occurred because the method assigned positive outcomes to the underprivileged class in uncertain probability regions, consequently increasing the false positive rate difference between the groups. Ultimately, we reached a consensus that **the most effective post-processing method was our own adaptation of the approach proposed by Barocas et al.** [5]. This method involved assigning different thresholds to different groups, achieving both high performance and effective mitigation of biases for the target protected groups.

4.2 Parkinson's disease

Moving on to Parkinson's disease results, we consider a sample of 475K subjects, where barely 1.5K will develop PD. As we see, this is highly imbalanced data set, with a proportion of **less than 0.5% positively labeled samples**. Dealing with such a highly imbalanced data set poses significant challenges in developing an accurate predictive model for Parkinson's disease. Traditional machine learning algorithms tend to favour the majority class, leading to poor performance in detecting the minority class.

To overcome this issue, we have applied **under and oversampling techniques** as well as specific ensemble models that target this problem. In Table 4.6 we present the performance results for different models:

- Two traditional ML models (Logistic Regression and Catboost) with balanced class weights.
- Balanced Random Forest Classifier: an adapted random forest that performs undersampling during training and uses balanced splitting criteria.
- Catboost classifier with different combinations of resampling techniques: Random Under Sampling (RUS), Random Over Sampling (ROS) and Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTENC).

We observe that all of the classifiers achieve State-of-the-Art AUC-ROC, and significantly high balanced accuracy (BA), taking into account the high class imbalance. However, these models achieve close to zero F1 and MCC scores. For PD, we have decided to optimize balanced accuracy in decision threshold selection as F1 improvement was limited.

Model	AUC-ROC	F1	BA	MCC
Logistic Regression	0.8114	0.0160	0.7422	0.0587
CatBoost	0.8157	**0.0367**	0.6190	0.0644
Balanced Random Forest Classifier	0.8149	0.0199	0.7432	**0.0660**
Catboost + RUS	0.8155	0.0132	0.7376	0.0532
Catboost + RUS + ROS	0.8175	0.0171	0.7456	0.0613
Catboost + RUS + SMOTENC	**0.8217**	0.0189	**0.7462**	0.0646

TABLE 4.6: Parkinson's Disease model performance metrics on UK-Biobank validation set (<1% of positive samples)

To gain insights into the causes of this low performance, we conducted a detailed analysis of Catboost with Random Under Sampling and SMOTENC. Examining the confusion matrix presented in Figure 4.8, we can easily identify the issue: the classifier is overpredicting positive PD cases, resulting in a 1% precision and a significant drop in the F1 score. Moreover, we see that we are achieving a recall of 62% and a specificity of 78%.

In Subfigure 4.9c, we illustrate the Precision-Recall curve, which clearly demonstrates that, at best, we can anticipate a precision of only 7%, at the expense of other metrics like recall. Furthermore, Subfigure 4.9b reveals that optimizing the threshold selection based on the F1 score yields marginal improvements compared to the considerable gains achieved in terms of balanced accuracy.

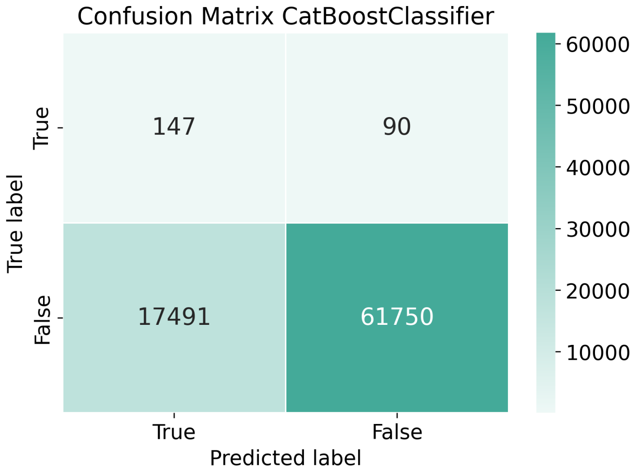
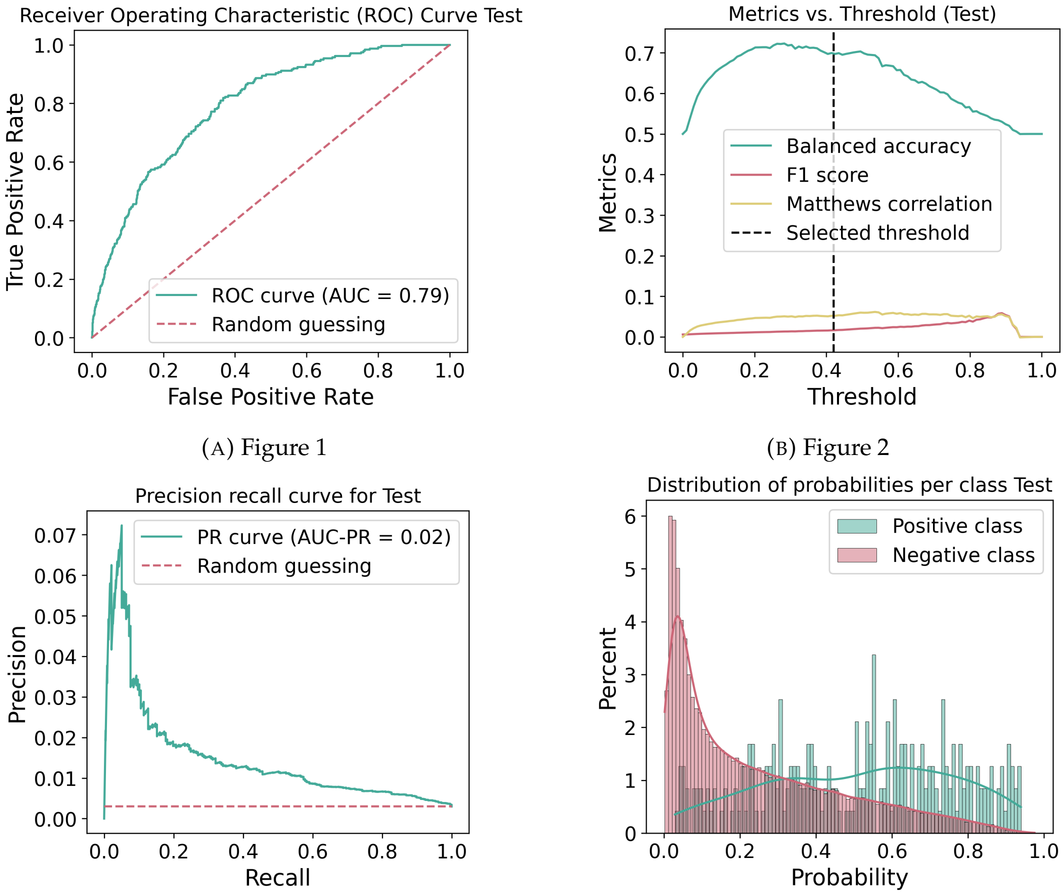


FIGURE 4.8: Confusion matrix for Catboost+RUS+SMOTENC model for Parkinson's Disease prediction

In summary, although precision may not be the primary metric of concern in the disease prediction application, the current results are considerably low. Therefore, we have not yet assessed the fairness of the predictions as future initial focus is on improving performance.



(A) Figure 1

(B) Figure 2

(C) Figure 3

(D) Figure 4

FIGURE 4.9: Catboost+RUS+SMOTENC classifier performance evaluation for Parkinson's Disease prediction

Chapter 5

Conclusion

This master thesis has explored the crucial topic of ensuring non-discrimination in disease prediction through the lens of fair machine learning. This research contributes to the ongoing efforts aimed at creating a more equitable healthcare system.

Throughout this study, we have **recognized the inherent biases present in a large-scale data set** such as UK Biobank, and we have analyzed the challenges in the disease prediction task. We have proposed a complete **ML Pipeline that goes from data wrangling to model deployment**, taking into account fairness considerations in different stages by design. We have **adapted prior statistical frameworks on fairness evaluation and mitigation for this particular application**, acknowledging the biological and social impact of protected features (sex, age, BMI, socioeconomic status and race) in the future development of diseases.

Although the methodology and technical implementation could be applied to potentially any disease in UKBiobank, our focus has been on two specific diseases: Primary Hypertension and Parkinson's Disease. For Primary Hypertension, we have successfully constructed a **prediction system that demonstrates performance comparable to the state-of-the-art**. Throughout the evaluation process, we have assessed model explanations and fairness metrics, allowing us to mitigate disparities for specific targeted groups. However, when it comes to Parkinson's Disease, we faced challenges in achieving acceptable performance. This difficulty may be attributed to the high data imbalance, with less than 1% of positive instances, or potentially due to the lack of pertinent information within the input features. It is worth noting that Parkinson's Disease is a complex disorder, and there remain numerous uncertainties in medical research concerning its underlying causes.

5.1 Future Work

Although our study has provided valuable insights into the challenges and opportunities associated with the disease prediction task, there are several avenues for future work that can expand upon our findings and address the limitations encountered.

In relation to the fairness framework, our future research aims to delve deeper into **addressing multiple protected attributes simultaneously**, recognizing that mitigating disparities for one group may inadvertently perpetuate biases for others. One avenue we intend to explore is the **expansion of our implementation of the Separation Postprocessor**, with a particular focus on identifying optimal thresholds for combinations of protected attributes. Additionally, we are keen on investigating the adaptability of the existing postprocessing techniques offered by AIF360 to address this particular issue. Also, we would like to **expand the bias mitigation with in-processing techniques** to a larger sample with more complex ML models. By doing so, we hope to assess whether we can mitigate the observed trade-off between fairness and accuracy, seeking a balance that optimizes both aspects. Lastly, we are eager to extend our research to incorporate **causal fairness frameworks**. By integrating causal fairness considerations into the medical domain, we aim to gain insights into the ability of such frameworks to adapt and provide fairness evaluations and interventions that account for the intricacies of causal relationships in disease development.

Regarding the disease prediction results in terms of performance, we target to improve model predictions by several means. First of all, we are aware of the **high data imbalance and we would like to further investigate this issue** by testing more combinations of under and over samplers for different target imbalance ratios. Secondly, we would like to evaluate the performance of different models that consider less input features. Currently, we experimented with all 156 variables in the training data but we suspect that many of these variables are adding noise and complexity to the model. We would like to apply **feature selection or dimensionality reduction techniques** to compare them with our actual model.

Appendix A

Appendix

The code used in this project is available in GitHub

Field	Name	Data Type	Medical category
F102	Pulse rate, automated reading	Integer	Physical measure summary
F1050	Time spend outdoors in summer	Integer, hours/day	Lifestyle
F1060	Time spent outdoors in winter	Integer, hours/day	Lifestyle
F1110	Length of mobile phone use	Categorical (single)	Lifestyle
F1160	Sleep duration	Integer, hours/day	Sleep
F1190	Nap during day	Categorical (single)	Sleep
F120	Birth weight known	Categorical single	Early life
F1200	Sleeplessness / insomnia	Categorical (single)	Sleep
F1220	Daytime dozing / sleeping	Categorical (single)	Sleep
F1239	Current tobacco smoking	Categorical single	Smoking
F1249	Past tobacco smoking	Categorical (single)	Smoking
F1279	Exposure to tobacco smoke outside home	Integer, hours/week	Smoking
F1289	Cooked vegetable intake	Integer, tablespoons/day	Diet summary
F1299	Salad / raw vegetable intake	Integer, tablespoons/day	Diet summary
F1309	Fresh fruit intake	Integer, pieces/day	Diet summary
F1329	Oily fish intake	Categorical (single)	Diet summary
F1339	Non-oily fish intake	Categorical (single)	Diet summary
F1349	Processed meat intake	Categorical (single)	Diet summary
F1359	Poultry intake	Categorical (single)	Diet summary
F1369	Beef intake	Categorical (single)	Diet summary
F137	Number of treatments/medications taken	Integer	Self-reported medical conditions
F1379	Lamb/mutton intake	Categorical (single)	Diet summary
F1389	Pork intake	Categorical (single)	Diet summary
F1408	Cheese intake	Categorical (single)	Diet summary
F1418	Milk type used	Categorical (single)	Diet summary
F1428	Spread type	Categorical (single)	Diet summary
F1438	Bread intake	Integer	Diet summary
F1448	Bread type	Categorical (single)	Diet summary
F1458	Cereal intake	Integer, bowls/week	Diet summary
F1478	Salt added to food	Categorical (single)	Diet summary
F1488	Tea intake	Integer, cups/day	Diet summary
F1498	Coffee intake	Integer, cups/day	Diet summary
F1528	Water intake	Integer, glasses/day	Diet summary
F1538	Major dietary changes in the last 5 years	Categorical single	Diet summary
F1548	Variation in diet	Categorical single	Diet summary
F1558	Alcohol intake frequency.	Categorical (single)	Alcohol
F1628	Alcohol intake versus 10 years previously	Categorical (single)	Alcohol
F1647	Country of birth (UK/elsewhere)	Categorical single	Early life
F1677	Breastfed as a baby	Categorical single	Early life
F1687	Comparative body size at age 10	Categorical (single)	Early life
F1697	Comparative height size at age 10	Categorical (single)	Early life
F1707	Handedness (chirality/laterality)	Categorical single	Early life
F1717	Skin colour	Categorical (single)	Lifestyle
F1727	Ease of skin tanning	Categorical single	Lifestyle
F1737	Childhood sunburn occasions	Integer, times	Lifestyle
F1747	Hair colour (natural, before greying)	Categorical (single)	Lifestyle
F1757	Facial ageing	Categorical single	Lifestyle
F1767	Adopted as a child	Categorical single	Early life
F1777	Part of a multiple birth	Categorical (single)	Early life
F1787	Maternal smoking around birth	Categorical (single)	Early life
F189	Townsend deprivation index at recruitment	Continuous	Baseline characteristics
F20023	Mean time to correctly identify matches	Integer	Cognitive function
F20116	Smoking status	Categorical single	Smoking

F20117	Alcohol drinker status	Categorical single	Alcohol
F20118	Home area population density - urban or rural	Categorical single	Geographical and location
F20160	Ever smoked	Categorical single	Smoking
F2040	Risk taking	Categorical single	Mental health
F2050	Frequency of depressed mood in last 2 weeks	Categorical (single)	Mental health
F2060	Frequency of unenthusiasm / disinterest in las...	Categorical (single)	Mental health
F2070	Frequency of tenseness / restlessness in last ...	Categorical (single)	Mental health
F2080	Frequency of tiredness / lethargy in last 2 weeks	Categorical (single)	Mental health
F2090	Seen doctor (GP) for nerves, anxiety, tension ...	Categorical (single)	Mental health
F2100	Seen a psychiatrist for nerves, anxiety, tensi...	Categorical (single)	Mental health
F21000	Ethnic background	Categorical single	Primary demographics
F21001	Body mass index (BMI)	Continuous	Physical measure summary
F21002	Weight	Continuous	Physical measure summary
F2178	Overall health rating	Categorical single	Self-reported medical conditions
F2188	Long-standing illness, disability or infirmity	Categorical (single)	Self-reported medical conditions
F2227	Other eye problems	Categorical single	Self-reported medical conditions
F2237	Plays computer games	Categorical single	Lifestyle
F2247	Hearing difficulty/problems	Categorical single	Self-reported medical conditions
F2267	Use of sun/uv protection	Categorical (single)	Lifestyle
F2296	Falls in the last year	Categorical single	Self-reported medical conditions
F23099	Body fat percentage	Continuous	Physical measure summary
F23100	Whole body fat mass	Continuous	Physical measure summary
F23101	Whole body fat-free mass	Continuous	Physical measure summary
F23102	Whole body water mass	Continuous	Physical measure summary
F23105	Basal metabolic rate	Continuous	Physical measure summary
F23106	Impedance of whole body	Continuous	Physical measure summary
F23107	Impedance of leg (right)	Continuous	Physical measure summary
F23108	Impedance of leg (left)	Continuous	Physical measure summary
F23109	Impedance of arm (right)	Continuous	Physical measure summary
F23110	Impedance of arm (left)	Continuous	Physical measure summary
F23111	Leg fat percentage (right)	Continuous	Physical measure summary
F23112	Leg fat mass (right)	Continuous	Physical measure summary
F23113	Leg fat-free mass (right)	Continuous	Physical measure summary
F23115	Leg fat percentage (left)	Continuous	Physical measure summary
F23116	Leg fat mass (left)	Continuous	Physical measure summary
F23117	Leg fat-free mass (left)	Continuous	Physical measure summary
F23119	Arm fat percentage (right)	Continuous	Physical measure summary
F23120	Arm fat mass (right)	Continuous	Physical measure summary
F23121	Arm fat-free mass (right)	Continuous	Physical measure summary
F23123	Arm fat percentage (left)	Continuous	Physical measure summary
F23124	Arm fat mass (left)	Continuous	Physical measure summary
F23125	Arm fat-free mass (left)	Continuous	Physical measure summary
F23127	Trunk fat percentage	Continuous	Physical measure summary
F23128	Trunk fat mass	Continuous	Physical measure summary
F23129	Trunk fat-free mass	Continuous	Physical measure summary
F2316	Wheeze or whistling in the chest in last year	Categorical single	Self-reported medical conditions
F2335	Chest pain or discomfort	Categorical single	Self-reported medical conditions
F24009	Traffic intensity on the nearest road	Integer	Geographical measures
F24014	Close to major road	Categorical single	Geographical and location
F24020	Average daytime sound level of noise pollution	Continuous	Geographical measures
F24021	Average evening sound level of noise pollution	Continuous	Geographical measures
F24022	Average night-time sound level of noise pollution	Continuous	Geographical measures
F24023	Average 16-hour sound level of noise pollution	Continuous	Residential noise pollution
F24024	Average 24-hour sound level of noise pollution	Continuous	Residential noise pollution
F2443	Diabetes diagnosed by doctor	Categorical (single)	Self-reported medical conditions
F2453	Cancer diagnosed by doctor	Categorical (single)	Self-reported medical conditions
F2473	Other serious medical condition/disability dia...	Categorical (single)	Self-reported medical conditions
F2492	Taking other prescription medications	Categorical single	Self-reported medical conditions
F30000	White blood cell (leukocyte) count	Continuous, 10^9 cells/Litre	Biochemistry and Haematology
F30020	Haemoglobin concentration	Continuous, grams/decilitre	Biochemistry and Haematology
F30080	Platelet count	Continuous, 10^9 cells/Litre	Biochemistry and Haematology
F30140	Neutrophill count	Continuous, 10^9 cells/Litre	Biochemistry and Haematology
F30150	Eosinophill count	Continuous, 10^9 cells/Litre	Biochemistry and Haematology
F30610	Alkaline phosphatase	Continuous	Biochemistry and Haematology
F30620	Alanine aminotransferase	Continuous	Biochemistry and Haematology
F3064	Peak expiratory flow (PEF)	Integer	Physical measure summary
F30640	Apolipoprotein B	Continuous	Biochemistry and Haematology
F30650	Aspartate aminotransferase	Continuous	Biochemistry and Haematology
F30670	Urea	Continuous	Biochemistry and Haematology
F30690	Cholesterol	Continuous	Biochemistry and Haematology
F30700	Creatinine	Continuous	Biochemistry and Haematology
F30710	C-reactive protein	Continuous	Biochemistry and Haematology

F30720	Cystatin C	Continuous	Biochemistry and Haematology
F30730	Gamma glutamyltransferase	Continuous	Biochemistry and Haematology
F30750	Glycated haemoglobin (HbA1c)	Continuous	Biochemistry and Haematology
F30770	IGF-1	Continuous	Biochemistry and Haematology
F30780	LDL direct	Continuous	Biochemistry and Haematology
F30870	Triglycerides	Continuous	Biochemistry and Haematology
F30880	Urate	Continuous	Biochemistry and Haematology
F31	Sex	Categorical single	Primary demographics
F34	Year of birth	Integer	Primary demographics
F4079	Diastolic blood pressure, automated reading	Integer	Physical measure summary
F4080	Systolic blood pressure, automated reading	Integer	Physical measure summary
F46	Hand grip strength (left)	Integer	Physical measure summary
F47	Hand grip strength (right)	Integer	Physical measure summary
F48	Waist circumference	Continuous	Physical measure summary
F49	Hip circumference	Continuous	Physical measure summary
F50	Standing height	Continuous	Physical measure summary
F54	UK Biobank assessment centre	Categorical single	Primary demographics
F6138	Qualifications	Categorical (multiple)	Education and employment
F6142	Current employment status	Categorical (multiple)	Education and employment
F6144	Never eat eggs, dairy, wheat, sugar	Categorical (multiple)	Diet summary
F6145	Illness, injury, bereavement, stress in last 2...	Categorical (multiple)	Mental health
F6150	Vascular/heart problems diagnosed by doctor	Categorical (multiple)	Self-reported medical conditions
F6152	Blood clot, DVT, bronchitis, emphysema, asthma...	Categorical (multiple)	Self-reported medical conditions
F6155	Vitamin and mineral supplements	Categorical (multiple)	Self-reported medical conditions
F6164	Types of physical activity in last 4 weeks	Categorical (multiple)	Physical activity
F6179	Mineral and other dietary supplements	Categorical (multiple)	Self-reported medical conditions
F864	Number of days/week walked 10+ minutes	Integer, days/week	Physical activity
F874	Duration of walks	Integer, minutes/day	Physical activity
F884	Number of days/week of moderate physical activ...	Integer, days/week	Physical activity
F904	Number of days/week of vigorous physical activ...	Integer, days/week	Physical activity
F924	Usual walking pace	Categorical (single)	Physical activity

TABLE A.1: Training features

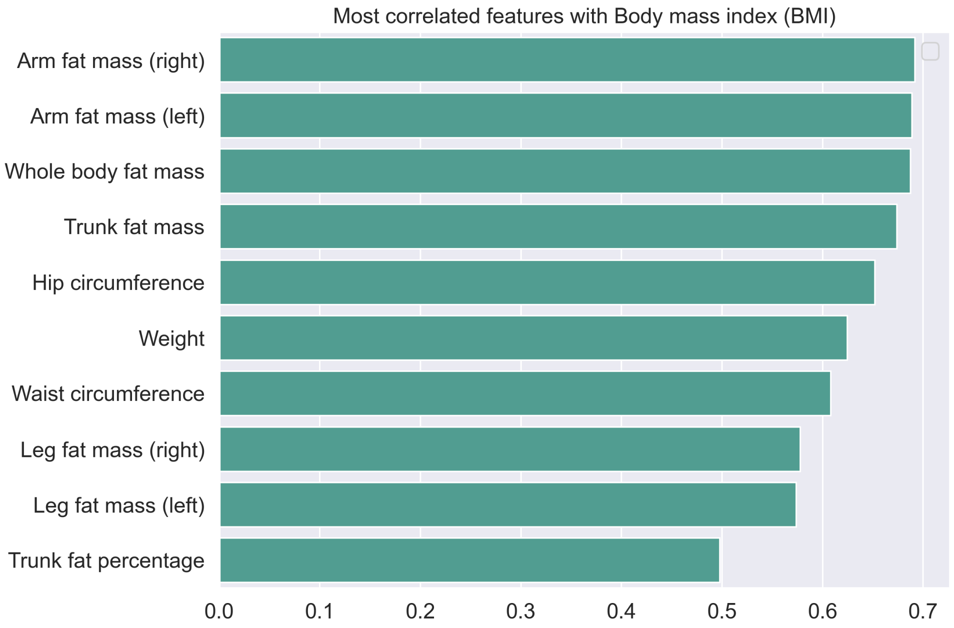


FIGURE A.1: Highly correlated features to BMI

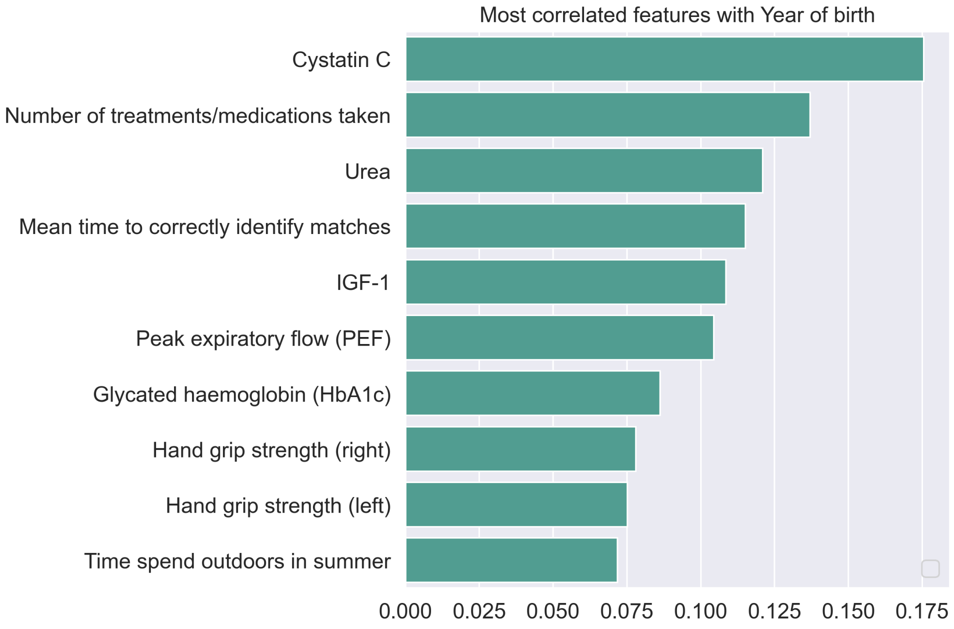


FIGURE A.2: Highly correlated features to Year of birth

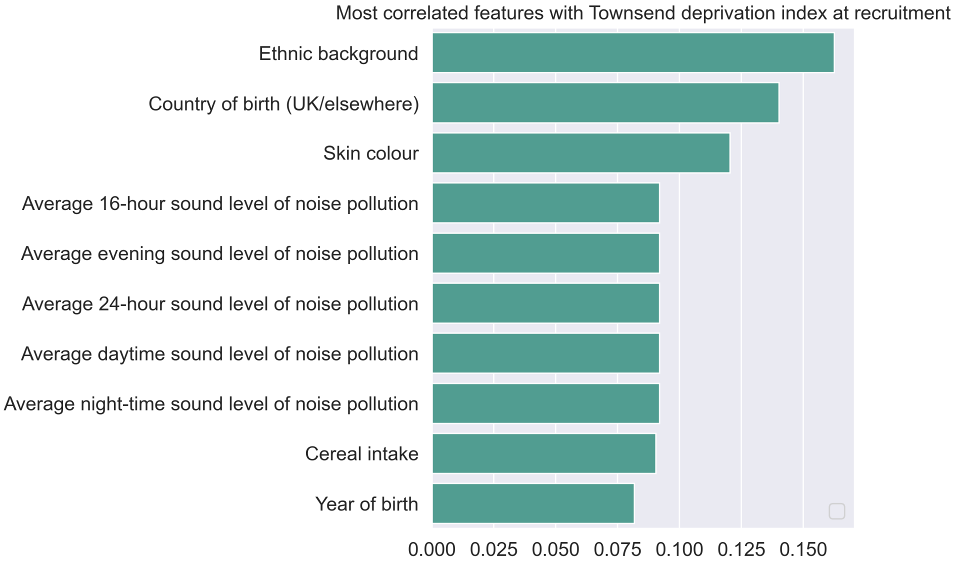


FIGURE A.3: Highly correlated features to Townsend deprivation index

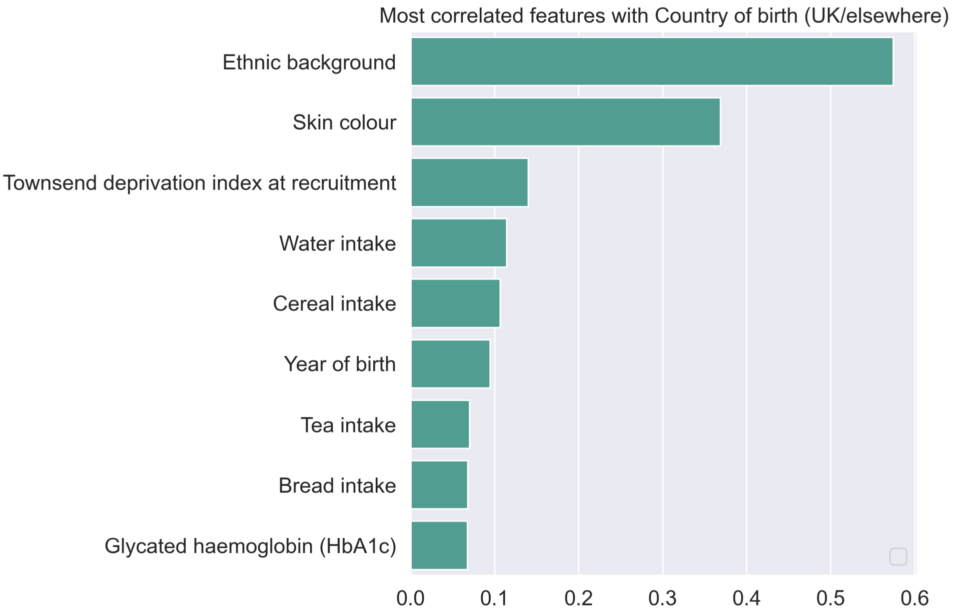


FIGURE A.4: Highly correlated features to Country of birth

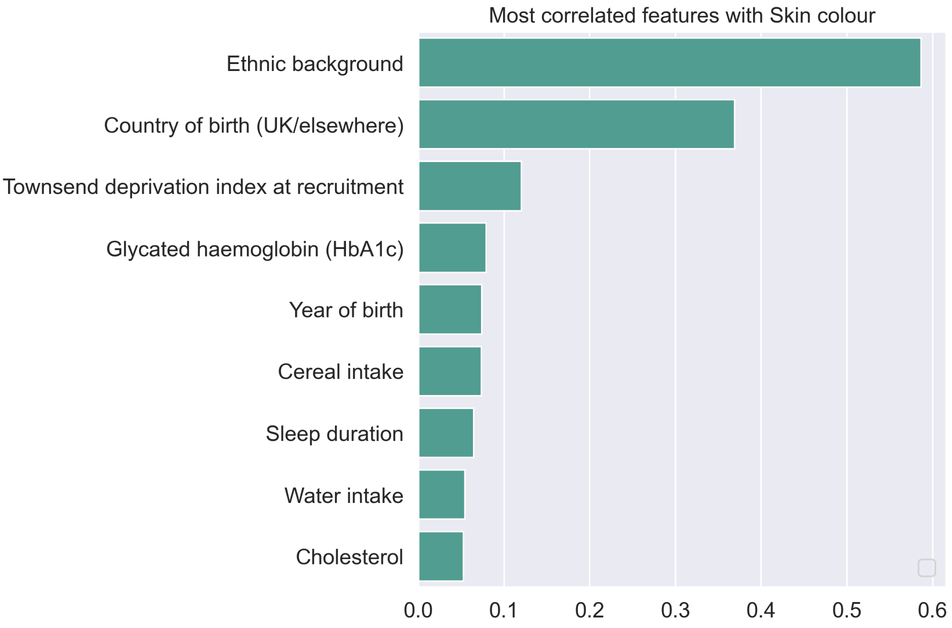


FIGURE A.5: Highly correlated features to Skin color

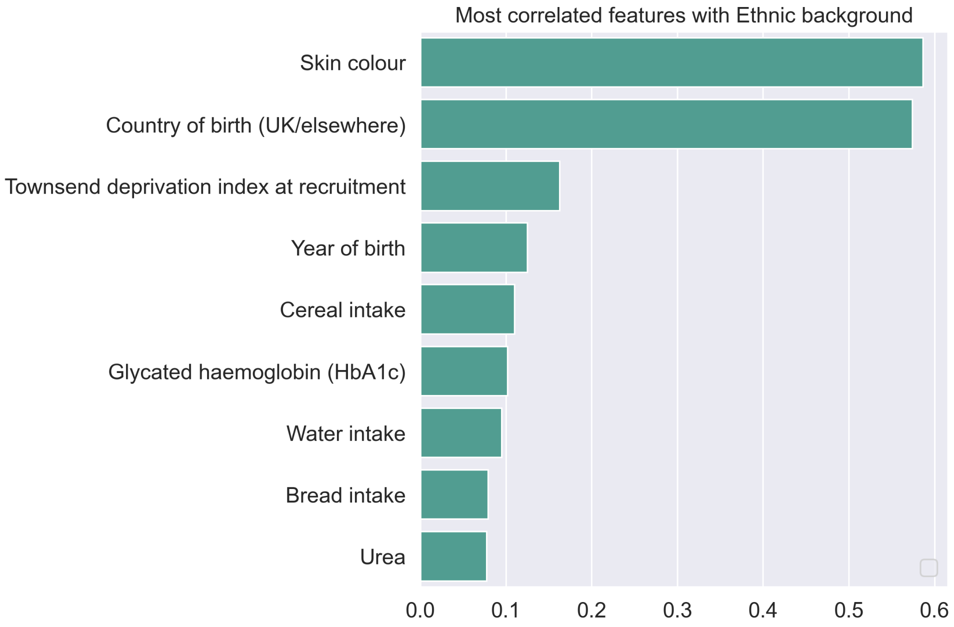


FIGURE A.6: Highly correlated features to Ethnic background

Bibliography

[1] P. Dworzynski, M. Aasbrenn, K. Rostgaard, and et al., "Nationwide prediction of type 2 diabetes comorbidities," *Scientific Reports*, vol. 10, no. 1, p. 1776, 2020. DOI: 10.1038/s41598-020-58601-7. [Online]. Available: https://doi.org/10.1038/s41598-020-58601-7.

[2] A. M. Alaa, T. Bolton, E. Di Angelantonio, J. H. F. Rudd, and M. van der Schaar, "Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 uk biobank participants," *PLOS ONE*, vol. 14, no. 5, pp. 1–17, May 2019. DOI: 10.1371/journal.pone.0213653. [Online]. Available: https://doi.org/10.1371/journal.pone.0213653.

[3] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019. DOI: 10.1126/science.aax2342.

[4] E. Commission, C. Directorate-General for Communications Networks, and Technology, *Ethics guidelines for trustworthy AI*. Publications Office, 2019. DOI: doi/10.2759/346720.

[5] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019, http://www.fairmlbook.org.

[6] A. N. Carey and X. Wu, *The fairness field guide: Perspectives from social and formal sciences*, 2022. DOI: https://doi.org/10.48550/arXiv.2201.05216. arXiv: 2201.05216 [cs.AI].

[7] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '17, Halifax, NS, Canada: Association for Computing Machinery, 2017, 797–806, ISBN: 9781450348874. DOI: 10.1145/3097983.3098095. [Online]. Available: https://doi.org/10.1145/3097983.3098095.

[8] Z. Ou, J. Pan, S. Tang, *et al.*, *Global trends in the incidence, prevalence, and years lived with disability of parkinson's disease in 204 countries/territories from 1990 to 2019*, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpubh.2021.776847/full.

[9] NCD Risk Factor Collaboration (NCD-RisC), "Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: A pooled analysis of 1201 population-representative studies with 104 million participants," *The Lancet*, vol. 398, no. 10304, pp. 957–978, 2021. DOI: 10.1016/S0140-6736(21)01330-1.

[10] A. Ward, A. Sarraju, S. Chung, and et al., "Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population," *npj Digital Medicine*, vol. 3, no. 1, p. 125, 2020. DOI: 10.1038/s41746-020-00331-1. [Online]. Available: https://doi.org/10.1038/s41746-020-00331-1.

[11] S. R. Pfohl, A. Foryciarz, and N. H. Shah, "An empirical characterization of fair machine learning for clinical risk prediction," *Journal of Biomedical Informatics*, vol. 113, p. 103 621, 2021, ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2020.103621. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046420302495.

[12] D. Pessach and E. Shmueli, "A review on fairness in machine learning," *ACM Comput. Surv.*, vol. 55, no. 3, 2022, ISSN: 0360-0300. DOI: 10.1145/3494672. [Online]. Available: https://doi.org/10.1145/3494672.

[13] R. K. E. Bellamy, K. Dey, M. Hind, *et al.*, "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, 4–4:15, 2019, ISSN: 0018-8646. DOI: 10.1147/JRD.2019.2942287.

[14] H. Weerts, M. Dudík, R. Edgar, A. Jalali, R. Lutz, and M. Madaio, *Fairlearn: Assessing and improving fairness of ai systems*, 2023. arXiv: 2303.16626 [cs.LG].

[15] F. Kamiran and T. Calders, "Classification with no discrimination by preferential sampling," Jan. 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*, ser. KDD '15, Sydney, NSW, Australia: Association for Computing Machinery, 2015, 259–268, ISBN: 9781450336642. DOI: 10.1145/2783258.2783311. [Online]. Available: https://doi.org/10.1145/2783258.2783311.

[17] F. Kamiran and T. Calders, "Data pre-processing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, Oct. 2011. DOI: 10.1007/s10115-011-0463-8.

[18] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International Conference on Machine Learning*, 2013.

[19] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Machine Learning and Knowledge Discovery in Databases*, P. A. Flach, T. De Bie, and N. Cristianini, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 35–50, ISBN: 978-3-642-33486-3.

[20] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, 2018, pp. 60–69. [Online]. Available: https://proceedings.mlr.press/v80/agarwal18a.html.

[21] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '18, New Orleans, LA, USA: Association for Computing Machinery, 2018, 335–340, ISBN: 9781450360128. DOI: 10.1145/3278721.3278779. [Online]. Available: https://doi.org/10.1145/3278721.3278779.

[22] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *CoRR*, vol. abs/1610.02413, 2016. arXiv: 1610.02413. [Online]. Available: http://arxiv.org/abs/1610.02413.

[23] 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. DOI: 10.1109/ICDM.2012.45.

[24] C. J. Kelly, A. Karthikesalingam, M. Suleyman, and et al., "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine*, vol. 17, no. 1, p. 195, 2019. DOI: 10.1186/s12916-019-1426-2. [Online]. Available: https://doi.org/10.1186/s12916-019-1426-2.

[25] A. Vellido, "The importance of interpretability and visualization in machine learning for applications in medicine and health care," *Neural Computing and Applications*, vol. 32, pp. 18 069–18 083, 2020. DOI: 10.1007/s00521-019-04051-w. [Online]. Available: https://doi.org/10.1007/s00521-019-04051-w.

[26] A. C. Dimopoulos, M. Nikolaidou, F. F. Caballero, and et al., "Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk," *BMC Med Res Methodol*, vol. 18, no. 1, p. 179, 2018. DOI: 10.1186/s12874-018-0644-1. [Online]. Available: https://doi.org/10.1186/s12874-018-0644-1.

[27] Parkinson's Europe, *About parkinson's - what is parkinson's?* https://www.parkinsonseurope.org/about-parkinsons/what-is-parkinsons/, Accessed: June 3, 2023, Content last reviewed: February 2018, 2018.

[28] World Health Organization (WHO), *Parkinson disease*, https://www.who.int/news-room/fact-sheets/detail/parkinson-disease, Accessed: June 3, 2023, 2022.

[29] P. Vineis, O. Robinson, M. Chadeau-Hyam, A. Dehghan, I. Mudway, and S. Dagnino, "What is new in the exposome?" *Environment international*, vol. 143, p. 105 887, 2020.

[30] S. M. Rappaport, D. K. Barupal, D. Wishart, P. Vineis, and A. Scalbert, "The blood exposome and its role in discovering causes of disease," *Environmental health perspectives*, vol. 122, no. 8, pp. 769–774, 2014.

[31] S. Bind, A. Tiwari, and A. Kumar, "A survey of machine learning based approaches for parkinson disease prediction," Aug. 2022.

[32] J. Winkler, R. Ehret, T. Büttner, and et al., "Parkinson's disease risk score: Moving to a premotor diagnosis," *Journal of Neurology*, vol. 258, no. Suppl 2, pp. 311–315, 2011. DOI: 10.1007/s00415-011-5952-x. [Online]. Available: https://doi.org/10.1007/s00415-011-5952-x.

[33] A. J. Noyce, J. P. Bestwick, L. Silveira-Moriyama, *et al.*, "Predict-pd: Identifying risk of parkinson's disease in the community: Methods and baseline results," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 85, no. 1, pp. 31–37, 2014, ISSN: 0022-3050. DOI: 10.1136/jnnp-2013-305420. eprint: https://jnnp.bmj.com/content/85/1/31.full.pdf. [Online]. Available: https://jnnp.bmj.com/content/85/1/31.

[34] J. A. Driver, G. Logroscino, J. M. Gaziano, and T. Kurth, "Incidence and remaining lifetime risk of parkinson disease in advanced age," *Neurology*, vol. 72, no. 5, pp. 432–438, 2009. DOI: 10.1212/01.wnl.0000341769.50075.bb.

[35] B. M. Jacobs, D. Belete, J. Bestwick, *et al.*, "Parkinson's disease determinants, prediction and gene-environment interactions in the uk biobank," *Journal of Neurology, Neurosurgery, and Psychiatry*, vol. 91, no. 10, pp. 1046–1054, 2020. DOI: 10.1136/jnnp-2020-323646.

[36] W.H.O., *Hypertension*, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/hypertension.

[37] G. F. S. Silva, T. P. Fagundes, B. C. Teixeira, and A. D. P. Chiavegatto Filho, "Machine learning for hypertension prediction: A systematic review," *Current Hypertension Reports*, vol. 24, no. 11, pp. 523–533, 2022, Epub 2022 Jun 22. DOI: 10.1007/s11906-022-01212-6.

[38] L. Yu, S. Li, H. Jiang, and J. Wang, "Exploring the relationship between hypertension and nutritional ingredients intake with machine learning," *Healthcare Technology Letters*, vol. 7, May 2020. DOI: 10.1049/htl.2019.0055.

[39] R. Patnaik, M. Chandran, S.-C. Lee, A. Gupta, C. Kim, and C. Kim, "Predicting the occurrence of essential hypertension using annual health records," in *2018 Second International Conference on Advances in Electronics, Computers and Communications (ICAECC)*, 2018, pp. 1–5. DOI: 10.1109/ICAECC.2018.8479458.

[40] L. A. AlKaabi, L. S. Ahmed, M. F. Al Attiyah, and M. E. Abdel-Rahman, "Predicting hypertension using machine learning: Findings from qatar biobank study," *PLoS One*, vol. 15, no. 10, e0240370, 2020. DOI: 10.1371/journal.pone.0240370.

[41] C. Criado Perez, *Invisible Women: Data Bias in a World Designed for Men*. New York, NY: Abrams Press, 2019, ISBN: 978-1419729072.

[42] K. M. Hoffman, S. Trawalter, J. R. Axt, and M. N. Oliver, "Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites," *Proceedings of the National Academy of Sciences*, vol. 113, no. 16, pp. 4296–4301, 2016. DOI: 10.1073/pnas.1516047113.

[43] J. Xu, Y. Xiao, W.-H. Wang, *et al.*, "Algorithmic fairness in computational medicine," *EBioMedicine*, vol. 84, p. 104 250, 2022. DOI: 10.1016/j.ebiom.2022.104250.

[44] M. Camacho, P. Gkontra, A. Atehortúa, and K. Lekadir, "Accessible and fair machine learning models for risk prediction of schizophrenia spectrum disorders," in *Empowering Communities: A Participatory Approach to AI for Mental Health*, 2022. [Online]. Available: https://openreview.net/forum?id=aYqwjL8CbH.

[45] V. N. Dang, A. Cascarano, R. H. Mulder, *et al.*, *Fairness and bias correction in machine learning for depression prediction: Results from four different study populations*, 2023. arXiv: 2211.05321 [cs.LG].

[46] A. Labrador, P. Gkontra, M. Camacho, *et al.*, "Cardiometabolic risk estimation using exposome data and machine learning," *SSRN Electronic Journal*, Jan. 2023. DOI: 10.2139/ssrn.4367352.

[47] *UK Biobank*, Retrieved from https://www.ukbiobank.ac.uk, 2022.

[48] L. Edwards, *Uk biobank scraper*, https://github.com/lwaw/UK-Biobank-scraper, Accessed: June 25, 2023, 2021.

[49] *International statistical classification of diseases and related health problems 10th revision*. [Online]. Available: https://icd.who.int/browse10/2019/en.

[50] E. H. R. Comission, *Protected characteristics*. [Online]. Available: https://www.equalityhumanrights.com/en/equality-act/protected-characteristics.

[51] J. O. Allen, E. Solway, M. Kirch, *et al.*, "Experiences of everyday ageism and the health of older us adults," *JAMA Netw Open*, vol. 5, no. 6, e2217240, 2022. DOI: 10.1001/jamanetworkopen.2022.17240.

[52] B. d'Alessandro, C. O'Neil, and T. LaGatta, "Conscientious classification: A data scientist's guide to discrimination-aware classification," *Big Data*, vol. 5, no. 2, pp. 120–134, 2017, PMID: 28632437. DOI: 10.1089/big.2016.0048. eprint: https://doi.org/10.1089/big.2016.0048. [Online]. Available: https://doi.org/10.1089/big.2016.0048.

[53] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *2010 20th international conference on pattern recognition*, IEEE, 2010, pp. 3121–3124.

[54] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, 2020. DOI: 10.1186/s12864-019-6413-7. [Online]. Available: https://doi.org/10.1186/s12864-019-6413-7.

[55] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[56] NHS, *High blood pressure (hypertension)*. [Online]. Available: https://www.nhs.uk/conditions/high-blood-pressure-hypertension/.

[57] WHO, *Raised blood pressure, global health observatory (gho) data*, 2015. [Online]. Available: https://web.archive.org/web/20160808122609/http://www.who.int/gho/ncd/risk_factors/blood_pressure_text/en/.