WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Master Thesis

Detecting and Mitigating Gender Bias in AI-Driven Cardiovascular Disease Diagnosis: A Use Case Analysis and Legal Perspective

Patricia Haumer, BSc. (WU)

Subject Area: Artificial Intelligence, IT Law

Supervisor: Assoz. Prof PD Dr. Sabrina Kirrane

Date of Submission: 23. December 2025

Department of Information Systems & Operations Management, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria

DEPARTMENT FÜR WIRTSCHAFTSINFORMATIK
UND OPERATIONS MANAGEMENT DEPARTMENT
OF INFORMATION SYSTEMS AND OPERATIONS
MANAGEMENT

Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

ADV - Adversarial Debiasing

AMA - American Heart Association

AIA - Artificial Intelligence Act

AI - Artificial Intelligence

BA - Balanced Accuracy

CNN - Convolutional neural network

CVD - Cardiovascular Disease

DI - Disparate Impact

DL - Deep Learning

DPD - Demographic Parity Difference

DT - Decision Tree

ECG - Electrocardiogram

EHRs - Electronic Health Records

EOD - Equal Opportunity Difference

EU - European Union

FNR - False Negative Rate

FPR - False Positive Rate

GDPR - General Data Protection Regulation

HF - Heart Failure

IQR - Interquartile Range

KNN - K-Nearest Neighbors

LIME - Local Interpretable Model-Agnostic Explanations

LSTM - Long Short-Term Memory

ML - Machine Learning

MLP - Multilayer Perceptron

PPV - Positive Predictive Parity

RF - Random Forest

SHAP - SHapley Additive exPlanations

SMOTE - Synthetic Minority Oversampling Technique

SPV - Statistical Parity Difference

SVM - Support Vector Machine

TNR - True Negative Rate

TPR - True Positive Rate

Acknowledgements

In the first place, I want to express my deepest gratitude to my supervisor, Assoz. Prof PD Dr. Sabrina Kirrane, for her invaluable guidance and support through the process of this thesis.

Nun möchte ich mich bei meinen Eltern, meinen Geschwistern und meinem Partner für die andauernde mentale Unterstützung bedanken. Besonders möchte ich mich bei meinen Eltern bedanken, die mich immer daran erinnert haben, dass ich alles schaffen kann, und mir nie gesagt haben, dass ich etwas nicht kann. Ihr habt mir stets vorgelebt, dass sich harte Arbeit auszahlt. All dies hat maßgeblich dazu beigetragen, dass ich heute an diesem Punkt stehe. Danke für alles!

Abstract

This study explores the presence of gender bias in Artificial Intelligence-assisted diagnosis systems for cardiovascular diseases, considering legal and technical aspects. Driven by new regulatory requirements, the Artificial Intelligence Act, and the growing deployment of Artificial Intelligence in medical applications, this thesis examines how gender bias manifests in the identification of cardiovascular diseases and how to effectively identify and mitigate it. Utilizing three cardiovascular datasets, each evaluated under multiple gender composition scenarios, this work employs a range of Machine Learning methods to predict cardiovascular disease outcomes. The Fairlearn and FairMLHealth toolkits are used to assess bias through the implementation of well-established fairness metrics. Subsequently, Fairlearn and AIF360 are deployed to implement distinct mitigation approaches. The results reveal that both model performance and fairness outcomes are significantly influenced by gender imbalance in training data. Despite the efficacy of certain mitigation strategies in reducing gender bias, a clear trade-off persists between prediction accuracy and fairness. The thesis underscores the paramount importance of integrating fairness into the development of Artificial Intelligence to promote its effective and lawful application in healthcare contexts.

1 Introduction

In recent years, our society has placed greater emphasis on health and well-being [67], a development also noted by the World Health Organization (WHO), which emphasizes the growing global recognition of health promotion and well-being [106]. Since the 1950s, the average lifespan has increased considerably as a result of medical advancements and healthier lifestyles. However, the increase in overall longevity does not necessarily imply an equivalent increase in the health span of individuals [50]. In fact, the probability of developing a disease increases with advancing age [93], as more than 80% of individuals over 65 experience at least one chronic illness [47]. Consequently, progress in the medical and healthcare fields is essential for improving disease prevention, early detection, and treatment [88]. This is particularly critical in cases where diseases may result in sudden death, such as cardiovascular diseases (CVD). This extends beyond the mere severity of this disease, as CVD is the most common cause of death worldwide [14]. One such technological advancement that holds great promise for enhancing the detection of CVD is Artificial Intelligence (AI) [88]. Nevertheless, the identification of CVD is complicated by the fact that symptoms manifest in a gender-dependent manner, displaying notable variability between genders [51]. Therefore, this results in the potential for decisions to be biased towards one gender, which could be further exacerbated through the implementation of AI [73]. As datasets serve as the foundation for training AI systems, another concern emerges [84]: historically, more data regarding heart health has been documented for men than for women [13]. In turn, this can affect the performance of algorithms for females and exacerbate the tendency for misinterpretation [73]. Beyond these fundamental issues, the implementation of AI also raises significant concerns regarding data protection and responsible use [56]. In response to these concerns, the European Union (EU) developed the Artificial Intelligence Act (AIA), a regulatory framework intended to ensure the safe and ethical deployment of AI [105]. Its objective is to govern the application of AI in the European market by promoting the implementation of safe, human rights-respecting AI that preserves the environment, public health, and safety [11]. Within this scope, the AIA should prevent discriminatory outcomes, such as gender-based discrimination, and subsequently promote equal treatment among genders in medical diagnosis supported by AI [105]. Otherwise, the utilization of AI in the context of CVD diagnosis holds the potential to result in life-threatening consequences [101]. Current works offer a wide range of bias detection and mitigation methods; however, without considering the new legal foundation, the AIA. Therefore, this work establishes a connection

between the AIA obligations and a real-world use case regarding AI-assisted CVD predictions. Thus, the present thesis is intended to be aligned with the AIA in identifying and mitigating gender-related disparities in AI-assisted CVD diagnosis. Moreover, this study offers a legal analysis of a real-world use case, as well as a fairness evaluation and mitigation using multiple tools applied to three CVD datasets. Therefore, adherence to the recent legal framework is facilitated, and a thorough comparative analysis of various gender bias detection and mitigation techniques is conducted.

1.1 Motivation of the Thesis

AI is on its way to change and revolutionize every aspect of our lives [53]. Rapid breakthroughs in hardware speed and software algorithms are driving the rapid emergence of AI. This is also evident in healthcare and medicine, where AI has dramatically improved diagnostic accuracy, enhanced medical imaging analysis, and optimized disease prediction, resulting in more accurate clinical decisions and greater efficiency [89]. However, there are also risks and disadvantages for individuals and society that can result from decisions made by an AI-driven system. To prevent risks such as biased decisions by AI systems, ethical and legal guidelines must be established and followed [81]. With the first publication of the EU AIA on July 12, 2024, the first steps have been taken to regulate the use of AI in the EU [12]. Using a risk-based approach to classify AI systems, the AIA establishes a consistent framework for all EU nations, dividing risk into four categories: minimal risk, limited risk, high-risk, and unacceptable risk [48]. As this thesis is specifically concerned with a use case in the medical field, a high-risk scenario is applicable in accordance with Article 6(2) of the EU AIA [4].

The EU AIA categorizes the domain of AI-driven healthcare systems as high-risk due to concerns pertaining to ethics and regulations, in addition to their potential implications for individual health. The presence of bias in medical AI systems is a significant issue that can lead to discrimination, errors in diagnosis, and unfair treatment of patients. Accordingly, the EU AIA includes provisions aimed at safeguarding the interests of patients and facilitating equity in healthcare. One such provision is the stipulation in Article 10 that high-risk AI systems must adhere to the principles of fundamental rights of individuals. A reason why such fairness measures are necessary is the gender bias in the diagnosis of CVDs [51]. There is evidence that AI models are trained on datasets containing predominantly male patients, contributing to severe underdiagnosis of women in this regard [73]. This presents significant ethical and legal issues regarding compliance with

medical regulations and anti-discrimination laws [57], and conflicts with the principles of fairness outlined in provisions of the AIA. Therefore, the elimination of bias in AI-based cardiovascular diagnostics is a primary focus of this research, as it is both a scientific need and a regulatory requirement.

Nonetheless, the existence of bias in AI technologies deployed in healthcare is evident despite the scientific advances in this field that have been made to that point. It is worth noting that the majority of algorithms do not consider gender and its influence on the individual's health or any disease disparities, leading to gender bias [36]. Gender bias is already a well-documented issue, especially in the identification and treatment of CVD. Research shows that women are underdiagnosed and underrepresented in clinical trials, and thus experience inequality in treatment [13]. The significance of this topic in the healthcare domain is indisputable as well as in society in general, as CVDs are known as *"silent killers"* [56] in medicine because they often go unnoticed [56]. Historically, the prevailing perspective on CVDs was that they primarily affected males. Nevertheless, this persists in impacting the treatment and diagnosis of CVDs for women. Indeed, despite women sharing equivalent risk factors for CVDs with men, they are observed to be 50% more prone to receiving a misdiagnosis of a heart attack [18]. Additionally, statistical evidence indicates that women are more likely to die from CVD than men [39]. Especially in healthcare-related implementations, bias in AI algorithms has become a serious concern. As a result of learning from data, these algorithms often reinforce or exacerbate biases in the data, producing unfair or incorrect results [98]. Inaccurate risk prediction, delayed or missed diagnoses are possible outcomes for certain patient groups [73]. However, the consequences go beyond the individual patient and include loss of confidence in healthcare institutions, legal and moral implications, misallocation of resources, and a slowdown in innovation [35].

A number of studies have previously examined the detection and mitigation of bias in the prediction of CVD. For example, the work by Sufian et al. [98] using cardiovascular imaging with a focus on mitigation, as well as the work by Li et al. [65] and Karim et al. [60], but using electronic health records (EHRs). However, this thesis specifically aims to address gender bias in the diagnosis of CVDs. From a medical standpoint, this matter was also analyzed by Al Hamid et al. [18] in a review of various studies. Their findings revealed that, contrary to men, women had less access to diagnostic testing and cardiovascular medications [18]. Medical studies such as the review by Desai et al. [39] and the work of Suman et al. [100] emphasize the differences in symptoms experienced by people of different genders when

they have CVD. These differences may be of hormonal, vascular, biological, or biochemical origin. Both papers identify risk factors for CVD, with Desai et al. [39] highlighting risk factors specific to women as a first step forward [39] [100].

In the initial phase of this study, an examination of a use case from a legal perspective related to a high-risk AI system for the diagnosis of CVD will be conducted. This use case involves the implementation of an AI-powered diagnostic tool in medical facilities, with the objective of assisting healthcare professionals in the detection of CVD. A more thorough description of the use case can be found in Section 3.1. The results should establish guidelines and objectives. These will serve as a foundation when reviewing bias detection and mitigation strategies applicable to the medical field. Furthermore, the outcomes of the legal analysis will serve as guidelines when implementing the bias identification and mitigation techniques in a practical setting using different models based on healthcare data predicting CVDs.

1.2 Research questions

To address the issue of gender bias in the diagnosis of CVD, both from a legal and technical standpoint, the following research questions were defined. For an examination of the complexity of this healthcare issue, these research questions will guide us in proposing effective bias detection and mitigation techniques.

Main research question: *How can gender bias in AI-based diagnosis of cardiovascular disease be effectively detected and mitigated in the healthcare sector?*

- RQ 1 (Legal Analysis) What are the legal implications of the EU AIA for the employment of bias detection and mitigation methods to detect gender bias in medical data-driven AI systems for predicting CVD?
- RQ 2 (Detection and Mitigation Strategies) How can gender bias in cardiovascular disease be identified and addressed?
- RQ 3 (Technical Analysis) How does using gender bias detection and mitigation strategies in AI-based cardiovascular disease diagnosis affect model performance and fairness?

These research questions will provide a clear legal perspective on the aforementioned use case and a solution approach to minimize the risks of using AI in a practical setting in such a sensitive area as healthcare.

1.3 Structure of the Thesis

Chapter 1 outlines the purpose and motivation of this thesis, emphasizing the importance of investigating gender bias in AI-driven cardiovascular disease diagnoses. Therein, the research questions that are intended to be answered in this work are proposed. Chapter 2 delineates extant knowledge about the subject in literature. Initiating a discussion on the state of the art in AI-assisted CVD diagnosis, the most successful CVD algorithms are reviewed. Furthermore, bias detection and mitigation techniques, as well as the tools that facilitate these methods, are introduced. Additionally, some background knowledge regarding the regulatory landscape before and after the AIA enforcement is described, thereby exploring the impact of its implementation. Overall, this chapter builds a foundation for the theory applied in the subsequent chapters of analysis. Next, in Chapter 3, the methodological approaches employed are outlined, including the Design Science and IRAC methods, and their application is elaborated. In addition, it incorporates the introduction of the use case, thereby establishing the foundation for the legal analysis. Chapter 4 is concerned with the legal perspective of the use case, on the basis of the AIA. Therein, legal and ethical concerns are highlighted, followed by the regulatory assessment of the use case by applying the IRAC method. Subsequently, Chapter 5 initiates the technical implementation involving data preparation and model development. First, the characteristics of the datasets are presented, followed by a discussion of the preprocessing procedure. Subsequent to this, ML models are established and applied to predict CVD, thus serving as a foundation for the identification and mitigation of gender bias. Next, Chapter 6 continues the technical implementation with a delineation of fairness tools, followed by a presentation of the results. Chapter 7 provides a reflection on the work that has been done and offers a critical discussion of it. This chapter also provides answers to the research questions. Furthermore, the critical perspective yields the recognition of the limitations of this thesis. Finally, Chapter 8 synthesizes the findings and underscores the contributions of this study to the academic literature, along with its implications for the medical field in terms of CVD prediction.

2 State of the Art

The underlying causes of gender bias in the diagnosis of CVD are historical misconceptions and medical negligence [39]. CVD has traditionally been thought of as a disease that primarily concerns men, thus overlooking its substantial impact on women, even though it is the leading cause of death in women [18]. Part of the reason for this bias was the idea that estrogen significantly prevented heart disease and delayed its occurrence in women by 8-10 years, compared with men. As a result, women's symptoms were often mislabeled, delaying diagnosis and limiting treatment options. Accurate diagnosis in women is complicated by biological characteristics e.g. such as smaller coronary vessels and a higher prevalence of small vessel disease [39].

2.1 Current State of CVD Detection using AI

A considerable rise in interest regarding the utilization of AI within the biomedical domain has been documented over the past two decades [89]. AI, especially Machine Learning (ML) and Deep Learning (DL), is seeing this growing interest, particularly in cardiology [41]. The emphasis is directed towards leveraging AI to augment analysis and patient care, with the overarching aim to achieve maximum efficacy in the healthcare sector [89]. However, it can be observed that AI is becoming increasingly proficient in other professional disciplines, and the challenge lies in achieving an equivalent level of proficiency in the medical domain [41]. By leveraging AI, healthcare professionals can deliver better treatments through an earlier and more precise detection of many types of diseases. Disease diagnosis and prediction are essential application domains within the broader field of biomedicine. Thus, it is also the area of biomedicine where artificial intelligence is most needed [89]. As AI is being used in a variety of ways to enhance CVD care, prevention, and diagnosis, Addissouky et al. [14] declared the subsequent distinct categories in which AI has been applied in relation to CVD in clinical practice:

- **Risk prediction:** AI systems are capable of evaluating substantial volumes of patient data, including imaging reports, medical records, and genetic data, to identify trends and predict the likelihood of a CVD [14]. Dorado-Díaz et al. [41] mention in particular the prediction of cardiac arrhythmias, ischemic heart disease, and heart failure in this context [41]. Briganti and Le Moine [27] have also discussed that AI leveraging EHRs has the capacity to more accurately assess the

potential for cardiovascular disease, including acute coronary syndrome and heart failure, in comparison to conventional methods [27].

- **Image analysis:** AI algorithms are capable of investigating medical images, such as cardiac ultrasounds or scans, to identify indicators that could lead to CVD [14]. Therefore, with ML models that effectively predict cardiovascular mortality and differentiate between abnormal and physiological heart conditions using echocardiographic data, AI is revolutionizing the diagnosis of CVD [41].

The author also highlights virtual assistants and drug discovery as two other applications, with AI helping to find new therapeutic targets for CVD and assisting in patient care by providing individualized virtual support [14].

It is evident that ML has already proven to be a significant positive transformation in healthcare applications [82]. In particular, decision support systems driven by ML that examine a clinical characteristic of a patient offer a valuable way to diagnose heart disease, as early discovery is crucial to minimizing its severe consequences [86]. Deep learning approaches also hold particular promise in the prediction of heart diseases due to their sophisticated real-time prediction abilities and great accuracy [111]. In recent years, a wide range of AI-driven techniques have been applied to the diagnosis of CVD [82].

Therefore, a thorough review of the literature on AI-based CVD prediction from the past five years showed the diverse range of algorithms applied to the task of predicting CVD. As illustrated in Figure 1, apart from traditional ML models, the employment of ensemble models, often in combination with traditional and deep learning approaches, has been observed in various works. While traditional models are barely applied exclusively for the prediction of CVD in the explored papers, they are frequently compared with ensemble approaches. More recent studies have demonstrated a recurring representation of deep learning models.

Table 1 provides an extensive overview of models applied for the prediction of CVD. Decision Trees (DT), Support Vector Machines (SVM), and K-Nearest Neighbors (KNN) were the prevailing traditional ML algorithms, yielding promising results. In multiple studies [20], [49], [45], [52], [25], DT demonstrated the strongest performance, achieving 98.6% accuracy, for example, in the work by Gao et al.[49]. Moreover, Saboor et al. [90] and Li et al. [66] underscored the efficacy of SVM, highlighting its superior performance, with the former achieving 96.72% accuracy and 98% F1-score.

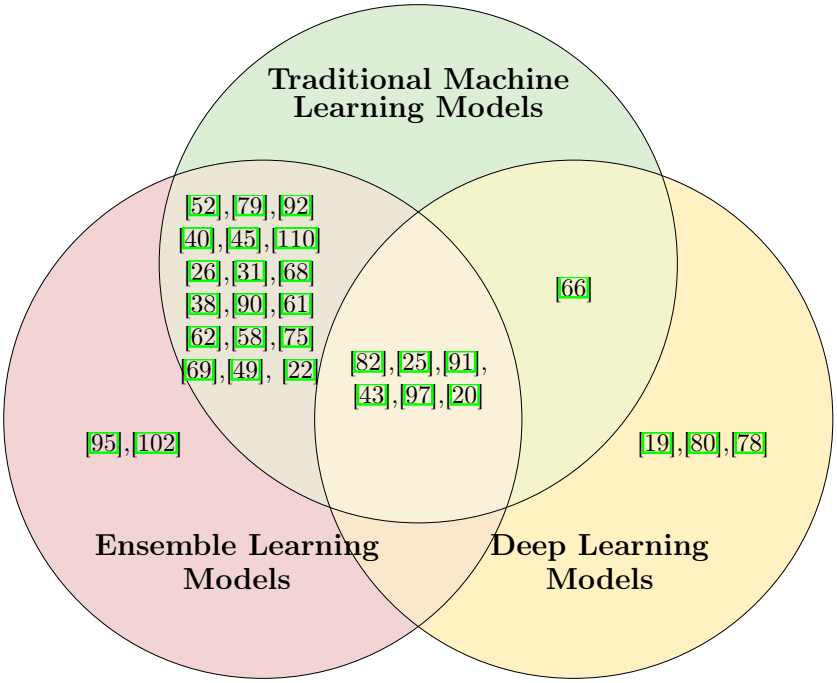


Figure 1: Distribution of Papers Reviewed across Traditional Machine Learning, Ensemble Models, and Deep Learning

In numerous investigations [20], [38], [78], [75], [69], [49], [82], Random Forest (RF) emerged as the most effective ensemble learning technique. Collectively, all these studies documented outstanding performance, with accuracies and F1-scores of at least 90% or higher. Other ensemble models that demonstrated great results included XGBoost [79], [45], [68], [95] and LightGBM [110]. With respect to DL, the most successful outcome was achieved through the combination of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) network, as demonstrated in Ali et al. [19] with an accuracy of 98,86% and F1-score of 99,83%. Apart from CNNs, multi-layer perceptrons (MLPs) are frequently incorporated from the deep learning domain, also showing promising performances in the works of Ali et al. [20] and Bhatt et al. [25].

Overall, it can be stated that the range of models employed in the context of CVD prediction is quite diverse. In accordance with the results of this examination of the relevant literature, the traditional models KNN and DT will be employed in the further proceedings of this work. As an ensemble model, the RF model will be utilized. In order to implement

CNNs or LSTMs, it is necessary to have time series data. Therefore, it was determined that the implementation MLP, would be the most suitable approach to analyze the EHRs that depict the patient's health status at a specific point in time.

Table 1: Overview of ML Approaches for CVD Prediction

Authors	**Year**	**Dataset**	**Traditional ML**	**Ensemble Learning**	**Deep Learning**	**Performance**
Al Reshan et al. [19]	2023	Cleveland dataset; Combination of five datasets: Cleveland, Switzerland, Statlog, Hungarian, Long Beach VA	SVM, DT, KNN	RF	ANN, CNN, LSTM, CNN-LSTM, Hybrid CNN-LSTM	In this study, the hybrid approach combining CNN-LSTM surpassed all other algorithms, achieving an F1-score of 0.9983 and accuracy of 98.86%.
Ali et al. [20]	2021	Heart Disease UCI dataset (sourced from Kaggle)	LR, DT, KNN	RF, AdaBoost	MLP	The algorithms KNN, RF and DT showed the highest levels of performance, each reaching an F1 score of 1.0 and an accuracy of 100%, indicating that no classification errors were encountered.
Baghdadi et al. [22]	2023	Heart Disease datasets from: Cleveland; Hungarian; Stalog; Long Beach VA; Switzerland	LR, LDA, SVM, DT, KNN	RF, AdaBoost, GB, CatBoost, XGBoost, LightGBM		The tuned CatBoost model achieved the highest accuracy of 90.94% and F1-score of 0.9231.
Bhatt et al. [25]	2023	Cardiovascular Disease Dataset Kaggle	DT	RF, XGBoost	MLP	The MLP demonstrated the highest F1 score, with a result of 86.71% and 87.28% accuracy, as determined by cross-validation.

Continued on next page

Table 1 – continued from previous page

Authors	Year	Dataset	Traditional ML	Ensemble Learning	Deep Learning	Performance
Bhowmik et al. [26]	2024	Cleveland Heart Disease	LR, SVM	RF		For predicting CVD, LR reached the best results in this study, showing an overall F1 score of 0.72 and an accuracy rate of 71.76%.
Chandrasekhar et al. [31]	2023	Cleveland Heart Disease; IEEE Dataport Heart Disease	LR, NB, KNN	RF, AdaBoost, GB, Soft Voting Ensemble		According to Chandrasekhar, LR yielded the best outcome for predicting CVD with an overall F1 score of 0.72 and the highest ROC-AUC score, 0.7810.
DeGroat et al. [38]	2024	CIGT (Clinically Integrated Genomics and Transcriptomics)	SVM, KNN	RF, XGBoost, Soft Voting		The Random Forest algorithm demonstrated the strongest performance, with a 95% accuracy rate, a F1 score of 0.96, and an ROC-AUC of 0.95.
Doppala et al. [40]	2022	Cleveland Heart Disease; Comprehensive Dataset aggregated from IEEE DataPort and other public sources; Mendeley Heart Disease Dataset	LR, NB, SVM, DT	RF, GB, XGBoost, Soft Voting Ensemble		The suggested soft voting ensemble model produced the best results, achieving an F1-score of 0.96 and an accuracy of 96.75% on the Mendeley dataset.

Continued on next page

Table 1 – continued from previous page

Authors	**Year**	**Dataset**	**Traditional ML**	**Ensemble Learning**	**Deep Learning**	**Performance**
Dritsas & Trigka [43]	2023	Dataset derived from prior study	LR, NB, KNN	RF, Rotation Forest, Bagging Classifier, AdaBoost, Voting, Stacking	MLP	Using Random Forest and Naive Bayes as base classifiers and Logistic Regression as the meta-classifier, the Stacking Ensemble Model showed the best performance, achieving an accuracy of 87.8%, a precision of 88%, and a recall of 88.3%.
El-Sofany et al. [45]	2024	Cleveland Heart Disease Dataset; Private Heart Disease Dataset	LR, NB, SVM, DT, KNN	RF, Bagging Classifier, AdaBoost, XGBoost, Voting		The XGBoost algorithm performed most successfully in the study, outperforming the other ten ML models that were examined with an accuracy of 97.57%.
Gao et al. [49]	2023	Heart Disease Dataset (Kaggle)	NB, SVM, DT, KNN	RF		The study revealed that the Decision Tree, utilizing Bagging and PCA-based feature selection, performed the best with 98.6% accuracy.
Ghosh et al. [52]	2021	Heart Disease datasets from Cleveland, Hungarian, Switzerland, VA Long Beach, Statlog	DT, KNN	RF, AdaBoost, GB, DTBM, KNNBM, GBBM, ABBM, RFBM		The RF with Bagging Method emerged as the most effective model in the study, yielding an accuracy of 92.63% and an F1-score of 0.92.

Continued on next page

Table 1 – continued from previous page

Authors	**Year**	**Dataset**	**Traditional ML**	**Ensemble Learning**	**Deep Learning**	**Performance**
Jindal et al. [58]	2021	Cleveland Heart Disease	LR, KNN	RF		In Jindal's work, the KNN model performed best and predicted heart disease with 88.52% accuracy.
Kavitha et al. [61]	2021	Cleveland Heart Disease	DT	RF, Hybrid RF+DT		The study revealed that the hybrid model integrating DT and RF achieved the highest accuracy, with 88% accuracy.
Khan et al. [62]	2023	Data obtained from Lady Reading Hospital & Khyber Teaching Hospital	LR, NB, SVM, DT	RF		The greatest accuracy of 85.01% and F1 score was achieved by the RF approach, which outperformed all other models examined.
Li et al. [66]	2020	Cleveland Heart Disease	LR, NB, SVM, DT, KNN		ANN	The SVM with a linear kernel, using features chosen by FCMIM, was the best-performing classifier in this investigation, achieving an accuracy of 92.37%.
Mahmud et al. [68]	2023	Cardiovascular Disease Dataset Kaggle	LR, SVM, DT, KNN	RF, XGBoost, Voting, Stacking, Bagging XG-Boost		The regular XGBoost model, the proposed Hybrid Voting Ensemble, and the Bagging XGBoost model all had the highest F1 score of 85%. The proposed Hybrid Voting Ensemble model demonstrated the most accurate performance with an accuracy of 84.036%, surpassing all other evaluated models.

Continued on next page

Table 1 – continued from previous page

Authors	Year	Dataset	Traditional ML	Ensemble Learning	Deep Learning	Performance
Maini et al. [69]	2021	Anonymized medical records were collected from a tertiary hospital in South India	LR, NB, KNN	RF, AdaBoost		Among the five evaluated models, RF performed the best. It achieved the highest accuracy 93.8% and F1 score of 0.93 when applied to the Indian population.
Motarwar et al. [75]	2020	Cleveland Heart Disease	Gaussian NB, SVM, Hoeffding Tree, Logistic Model Tree	RF		This study revealed the RF algorithm as a top performer, exhibiting the highest F1 score and an accuracy of 95.08%.
Nagavelli et al. [79]	2022	Cleveland Heart Disease Dataset; MCG signals obtained from 574 subjects; Raw ECG signal from PhysioNet; Cleveland and Statlog datasets were merged	NB, SVM	XGBoost, Stacking (SVMs and XGBoost)		XGBoost outclassed all other models and obtained the most outstanding performance, obtaining 95.90% accuracy and 97.10% precision.
Naeem et al. [78]	2024	Self-Generated Dataset(s)		RF	ANN	Among the ML baselines, the Random Forest model showed the highest AUC of 0.94 and the highest F-measure of 0.89. Still, its accuracy of 90% was lower than the 97.9% accuracy of the ANN.

Continued on next page

Table 1 – continued from previous page

Authors	Year	Dataset	Traditional ML	Ensemble Learning	Deep Learning	Performance
Nancy et al. [80]	2022	Cleveland Heart Disease; Hungarian Heart Disease			LSTM, FIS + LSTM, FBiLSTM: combining fuzzy logic + Bidirectional LSTM	The proposed fuzzy inference system, when paired with a bidirectional LSTM model, demonstrated 98.86% accuracy in predicting CVD.
Ogunpola et al. [82]	2024	Heart Disease Cleveland Dataset; Heart Disease Dataset from Mendeley database	LR, SVM, KNN	RF, GB, XGBoost	CNN	The authors found that a fine-tuned XGBoost model is one of the most accurate, as it can predict CVD with 98.5% accuracy and an F1-score of 98.71%.
Saboor et al. [90]	2022	Cleveland Heart Disease Dataset; Z-Alizadeh Sani Dataset; StatLog Heart Disease Dataset	LR, LDA, NB, SVM, Classification Tree	RF, GB, XGBoost		Among the several models put to the test, the SVM achieved the best predictive performance for CVD with an accuracy rate of 96.7% and an F1 score of 98%.
Sadr et al. [91]	2024	Cardiovascular Disease Dataset Kaggle; Heart Disease Dataset Kaggle	LR, NB, SVM, DT, KNN	RF, AdaBoost, XGBoost, LightGBM	MLP, CNN, LSTM, CNN-LSTM, CNN-LSTM + KNN + XGB	The proposed hybrid model, which integrates CNN-LSTM, KNN, and XGBoost, yielded superior outcomes across all three datasets, reaching an accuracy of 95.85% and an F1 score of 92.63%.
Shah et al. [92]	2020	Cleveland Heart Disease	NB, DT, KNN	RF		In a study by Shah et al., NB surpassed DT, KNN, and RF, achieving an accuracy of 88.16%.

Continued on next page

Table 1 – continued from previous page

Authors	Year	Dataset	Traditional ML	Ensemble Learning	Deep Learning	Performance
Srinivas & Katarya [95]	2022	Cleveland Heart Disease Dataset; Heart Failure Prediction Dataset (Kaggle); Heart Disease UCI Dataset		XGBoost		Using the Cleveland dataset, the XGBoost model with OPTUNA produced the best results: an accuracy of 94.71% and an F1 score of 0.94.
Subramani et al. [97]	2023	Heart disease dataset combining Cleveland, Hungarian, and Switzerland, Long Beach VA and Stalog	LR, NB, SVM, DT, KNN	RF, GB, CatBoost, XGBoost, LightGBM	MLP	The stacking ensemble model achieved the highest overall accuracy of 96% using LR as the meta-learner alongside LR, RF, DT, MLP, and CatBoost.
Theerthagiri & Vidya [102]	2022	Cardiovascular Disease Dataset Kaggle	LDA, NB, DT, KNN	GB	MLP	By combining a GB classifier with recursive feature elimination, the authors' suggested model achieved maximum accuracy of 89.78% and an F1 score of 0.83.
Zhang et al. [110]	2022	Z-Alizadeh Sani Dataset	LR	LightGBM, Stacking (LightGBM + LR)		The Stacking model proposed by the author outperforms both LR and the standalone LightGBM model, combining LightGBM's feature generation with LR, achieving 91.4% of accuracy and a F1 score of 94.2%.

However, many obstacles remain, including the necessity of large datasets of high-quality and their integration into real-world contexts [41]. Furthermore, bias in training datasets, interpretability issues, absence of clinical validation, and high implementation costs are also among the major challenges. To achieve widespread use of AI in medicine, several barriers must be overcome [14]. Despite recent advances, the integration of AI in medical diagnostics is still in its early stages, and more advanced technologies are already on the starting blocks. In research, quantum AI technology is already being explored for its potential to fasten processing in algorithms and provide accelerated diagnostic systems [17].

2.2 Bias Detection Techniques

AI has a lot to offer the medical field, including precision medicine, task automation, and advancing research by identifying intricate patterns in medical data [41]. However, there are also drawbacks to its application, such as bias, which in the medical field refers to "*systematic errors leading to a distance between prediction and truth, to the potential detriment of all or some patients" [63].* Bias can be categorized in various manners, this work adopts the classification proposed by Mehrabi et al. [71] which distinguishes bias into three main types:

- **Data bias:** is present when the data from which the algorithm is learning contains bias [71]. Parate et al.[84] reviewed data bias specifically in healthcare applications and emphasized that it is a significant issue in data-dependent decision-making. Data bias may originate during collection or emerge during the subsequent stages of analysis and interpretation [84], consequently, data bias can be further categorized. In terms of cardiovascular health predictions, Sufian et al. [98] specifically mention sampling and measurement bias as bias types. While measurement bias is caused by systematic flaws in data collection, such as malfunctioning medical equipment or inaccurate collection techniques, sampling bias arises when the training data is not representative of the general population, resulting in poor model performance for underrepresented groups [98].

- **Algorithmic bias:** is described as embedded and recurring biases in models that yield unfair results and exacerbate existing health disparities [74]. Unlike data bias, algorithmic bias can occur even when the input data is free of bias [71]. It arises from the manner in which models employed in medical practice interpret data and learn over time [74]. Predictions for different subgroups can be affected by

unintentional biases introduced by design choices such as statistical estimators, regularization strategies, and optimization functions [71].

- **User Interaction bias:** can be created by the user, by choosing a biased action, or through the user interface [71]. Bias can be introduced into an AI system by the user or by the human-designed interface that incorporates the bias into the system. Any bias held by humans can be transmitted to AI systems, which may serve to reinforce specific behaviors [81].

Having identified the primary categories of bias, the subsequent task is to address the essential task of bias detection in healthcare AI systems to ascertain unbiased results and prevent inequalities. One of the most significant methods is statistical analysis, which detects bias in healthcare statistics through the employment of fairness metrics [35]. Within the existing literature on healthcare applications, the metrics used for fairness evaluation are specified in Table 2. Reviewing bias detection techniques in AI-supported healthcare applications, the paper by Hasanzadeh et al. [54] lists demographic parity, equal opportunity, equalized odds, and causal fairness as the primary available fairness metrics. Moreover, fairness is generally measurable through performance metrics such as accuracy, false positive rate, and true positive rate, when evaluated across distinct subgroups [33].

Metric	Description	References
Accuracy	Overall correctness of the model by assessing the proportion of total correct cases with respect to all observations	Straw et al. [96], Sufian et al. [98]
Average Odds Difference	measures the average difference in true positive and false positive rates between groups [109]	Sufian et al. [98]
Balanced Classification Accuracy	measures the disparity in balanced accuracy (average of sensitivity and specificity) between groups [104]	Sufian et al. [98]
Causal Fairness	assesses how independent the expected prediction is from the sensitive attribute [54]	Hasanzadeh et al. [54]
Demographic Parity	quantifies the statistical independence between the sensitive attribute and the predictive result [54]	Hasanzadeh et al. [54]
Disparate Impact	is a measure of the ratio of predicted favorable label percentage between the privileged and unprivileged groups [65]	Li et al. [65], Sufian et al. [98]
Equal Opportunity	ensures that Individuals with the true positive outcome have an equal chance of receiving a positive prediction across groups [35]	Hasanzadeh et al. [54], Chinta et al. [35]
Equal Opportunity Difference (EOD)	calculates the discrepancy in true positive rates between the privileged and unprivileged groups [65]	Karim & Asjad [59], Li et al. [65], Sufian et al. [98]
Equalized Odds	computes the absolute difference between FPR and the FNR for both demographics, measuring disparate mistreatment [96]	Hasanzadeh et al. [54], Straw et al. [96]
False Negative Rate (FNR)	Percentage of positive results falsely labelled as negative [96]	Straw et al. [96]
False Positive Rate (FPR)	Percentage of negative results falsely labelled as positive [96]	Straw et al. [96]
ROC-AUC	measures the area under the receiver operating characteristic curve, representing the model's ability to distinguish between classes [96]	Straw et al. [96]
Statistical Parity	ensures that the probability of a positive predicted outcome is equal across groups defined by a sensitive attribute [35]	Chinta et al. [35]
Theil Index	quantifies the disparity in the prediction outcomes distributed to various groups [37]	Sufian et al. [98]

Table 2: Fairness Metrics Used in Healthcare Applications

In particular, statistical parity, equal opportunity, and predictive equity were cited as the fairness measures applied to EHRs [35]. The study by Li et al. [65] refers to Equal Opportunity Difference (EOD) and Disparate Impact (DI) as fairness metrics to detect bias in data. Sufian et al. [98] used, in addition to EOD and DI, the Theil Index to quantify the disparity in the prediction outcomes distributed to various groups, incorporating factors such as smoking behavior and gender. A comparison of model outcomes based on patient demographics, such as age or gender, can be used to identify potential biases that may influence the outcomes [63]. To assess differences in prediction outcomes among demographic groups and evaluate the model's overall accuracy, performance-based metrics such as Average Odds Difference and Balanced Classification Accuracy were employed by Sufian et al. [98]. Aside from statistical analysis as a method for bias detection, Chinta et al. [35] also suggested end-user feedback and audit tools.

In the work of Sufian et al. [98], Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) are introduced as interpretability techniques. These techniques are designed to promote the explainability of AI models, thereby helping physicians and medical staff to understand, properly trust, and interact with AI [23]. SHAP illustrates the contribution of each feature to the model's output, while LIME provides insight into how specific features contribute to the prediction probabilities produced by the model. By providing such insights, LIME and SHAP help expose sophisticated biases [98].

2.3 Bias Mitigation Approaches

As bias has the potential to manifest during the various stages of model development [84], the mitigation of bias can also be initiated during these three distinct stages: pre-processing, in-processing, and post-processing [32]. An overview of common bias mitigation techniques used in healthcare applications is given in Table 3, with the techniques grouped by the stage of mitigation.

Pre-Processing: The strategy in this step is to ensure diverse demographic representation prior to model training by adjusting the input data [35], as there are often biases in the training data that models learn from [84]. A variety of pre-processing methods, including reweighting, sampling and optimum data transformation have been employed in the past to mitigate bias. Depending on the specific circumstances, these approaches are either basic data preparation methods or more complex strategies that increase

Mitigation Stage	Technique	Description	Used By
Pre-processing	Reweighting	Adjusts the weights to balance the influence of each group while maintaining the sample size [87].	Raza et al. [87], Sufian et al. [98], Karim & Asjad [60]
	Resampling	Alters the composition of the dataset by reducing the representation of dominant groups and increasing the presence of minority ones [84].	Li et al. [65], Straw et al. [96]
	Data Transformation	Modifies or re-labels to eliminate bias before training [84].	Sufian et al. [98], Straw et al. [96]
	Fairness by Ignoring	Excluding sensitive attributes prior training [65].	Li et al. [65]
In-processing	Adversarial Debiasing (ADV)	Trains the model alongside an adversary that attempts to predict sensitive features [87].	Raza et al. [87], Sufian et al. [98], Straw et al. [96]
	Constraint-based Optimization	Incorporates fairness constraints into the model's optimization process [81].	Sufian et al. [98], Straw et al. [96]
	Reweighting during training	Updates sample weights dynamically during training, putting emphasis on underrepresented groups [84].	Cenitta et al. [30], Li et al. [64]
	Transfer Learning	Uses a pretrained model trained on a larger dataset for better performance [84].	Cenitta et al. [30]
Post-processing	Threshold-based adjustment	Applies methods such as output calibration and threshold adjustment to modify decision thresholds or reconfigure expected probabilities.	Sufian et al. [98]
	Constraint-Based Methods	Adapts model outputs to align with fairness constraints encompassing equalized odds, calibrated equalized odds and reject option classification.	Sufian et al. [98]
	Counterfactual-Based Techniques	Ensures that predictions remain consistent when sensitive attributes are altered by using counterfactual examples.	Sufian et al. [98]

Table 3: Overview of Bias Mitigation Techniques Grouped by Mitigation Stage

model predictability while reducing bias [84]. Techniques such as resampling and reweighting address data imbalances and promote fairness in model development [35]. In order to equalize the distribution of classes, re-sampling and re-weighting algorithms change the sampling probability or loss weight for majority and minority samples [63]. Li et al. [65] compared three debiasing strategies: resampling by proportion, which balances positive class ratios across groups; resampling by group size, which oversamples the minority group; and eliminating protected attributes such as gender and race to reduce bias through fairness by ignoring protected variables [65]. In the review on data bias in healthcare applications, Parate et al. [84] state that resampling techniques, such as BayesBoost, SMOTE and AdaSyn, have proven useful in addressing class imbalances. All of these approaches use synthetic data generation as a mitigation strategy to rebalance the underrepresented class to ensure fairness [84]. Although SMOTE and AdaSyn balance classes, they may not maintain the actual data distributions. For this reason, BayesBoost is considered to be more powerful because it attempts to more accurately represent the distributions in the real world, rather than simply rebalancing the classes [42]. However, as a more effective method, they presented FairSMOTE, which avoids omitting sensitive attributes during the resampling process by taking sensitive features and class labels into account when balancing data [84]. Re-weighting balances demographic representation by modifying training sample weights prior to model training. By weighting samples based on frequency counts, Raza's study used this method to predict hospital readmissions for patients with diabetes while maintaining equity across gender, race, and age [87]. Data transformation is mentioned by Parate et al. [84] as a preprocessing technique for debiasing data. This technique involves modifying data to debias it prior to model training [84]. Basic data transformation or relabeling are useful strategies when the bias is due to historical discrimination. In cases where bias results from inadequate or irregular data collection, more advanced techniques, such as optimal data transformation, may be required [84]. However, Sufian et al. [98] employ data transformation by the reformulation of sensitive attributes in conjunction with feature engineering to mitigate representational bias prior to model training [98].

In-Processing: In-processing methods build fairness into the learning algorithm during training [35]. In-processing techniques, therefore, concentrate on modifying the learning procedure of the model. These methods aim to prevent the model from predicting only the dominant class or from reflecting the existing bias [84]. Ntoutsi et al. [81] explained that by

directly addressing the discriminative behavior of the model, in-processing techniques modify the classification process. Therefore, in-processing approaches adjust the classification procedure by incorporating fairness constraints, such as modifying decision tree splits and reducing indirect bias with regularizers [81]. By ensuring that protected characteristics, such as gender, do not affect the results, techniques such as constraint-based optimization and Adversarial Debiasing (ADV) diminish bias [35]. The ADV method trains a classifier to optimize its prediction accuracy while simultaneously minimizing an adversary's ability to infer sensitive characteristics from the predictions [87]. Constraint-based optimization is responsible for debiasing while maintaining prediction accuracy by incorporating fairness requirements into the learning method. In healthcare models, this approach helps ensure fair treatment advice by imposing constraints such as equality of opportunity or demographic parity [35]. Reweighting can also be implemented during the training of the model. This process adjusts the sample weights of training data in a manner that emphasizes underrepresented classes [84]. An in-processing method that is particularly useful for reducing bias in situations where there is little data to work with is transfer learning. Applying pretrained models to large datasets improves model performance by reducing bias and allowing the target model to absorb past knowledge [84]. Further in-processing methods that facilitate debiasing include limited capacity models, gradient starvation mitigation, invariant risk minimization, ensembling approaches, distributionally robust optimization, and invariant causal predictors. These approaches achieve bias reduction by updating the objective function or imposing constraints on the model [63].

Post-Processing: This technique involves modifying the model output after training to achieve fair results [35], whereby this modification of the output predictions is usually based on fairness constraints [63]. Especially, in cases where the model is considered a black box, with no control over the learning or training data, post-processing stands as the only feasible option [71]. By modifying the output, post-processing techniques can help correct biases generated during the model training [84]. For example, such strategies involve threshold adjustment and output recalibration, which align predictions with real medical outcomes [35]. Decision threshold selection, which modifies the categorization threshold, is another way to ensure fair results. For models used in the context of high-stakes decisions, where fairness is indispensable, this approach is highly beneficial [84]. Threshold adjustment determines the optimal cutoff for predictive algorithms to correctly classify outcomes, including determining patient

risk or diagnosing disease. One notable application is the prediction of CVD [35]. By changing the expected probability of a model, calibration eliminates bias in the expected probabilities. This is very valuable for models that rely on predicted probabilities to make decisions, such as medical diagnoses [84]. Recalibrating the output is essential for AI-driven healthcare to adapt predictive models to local conditions and to ensure that the expected probabilities are consistent with real medical outcomes when applying the model to a new patient population [35]. Razza [87] outlined the following algorithms for post-processing debiasing: In order to produce more fair results, the reject option classification modifies the output by giving labels with preference to underprivileged groups. The equalized odds algorithm adjusts the labels to maximize the equalized odds through linear programming, whereas the calibrated equalized odds algorithm optimizes the scores to find a probability to make the output labels fairer [87]. Sufian et al. [98] mention counterfactual fairness as another post-processing technique that recognizes and reduces bias through the creation of counterfactual instances. These are modified inputs that help to determine whether the predictions of the model are changing in an unfair way depending on sensitive features. This approach ensures that predictions remain unbiased and consistent by modifying outputs to conform to fairness constraints [98].

2.4 Tools for Bias Detection and Mitigation:

Recently, there have been some advances in tools that are capable of judging the fairness of a system [71]. Given that the responsibility for fair models rests with the developer [83], the software engineering community has lately begun to engage with fairness-aware approaches and, in particular, with fairness testing [87]. Such tools or libraries are capable of measuring and visually representing the degree of bias in the data. Others focus more on investigating the output of the algorithms using metrics that measure fairness [108]. Raza [87] developed an ML model to address health inequalities in hospital readmission. FairML[1], FairTest[2], themis-ML[3], and AIF360[4] were mentioned as popular fairness toolkits in Raza's paper [87]. Furthermore, in the review of algorithmic fairness in the context of computational medicine,

[1]https://github.com/adebayoj/fairml
[2]https://github.com/columbia/fairtest
[3]https://themis-ml.readthedocs.io/en/latest/
[4]https://aif360.res.ibm.com/

Xu et al. [108] name FairMLHealth[5], Fairlearn[6], and some Google tools[7], such as the ML fairness gym and fairness indicators, as common libraries used in fairness research [108]. These insights from these two works provided an initial grasp of the tools used within the health and medicine sector. Drawing from this evaluation of literature, the following selection of tools will be leveraged throughout bias detection and mitigation: Fairlearn and FairMLHealth will be utilized for the purpose of bias detection, thereby contrasting Fairlearn as a general fairness tool and FairMLHealth as a healthcare-specific tool that provides domain-specific insights into fairness in the prediction of CVD. As FairMLHealth is constrained to bias detection techniques, bias mitigation is performed by comparing approaches from AIF360 and Fairlearn, two of the most widely used and comprehensive toolkits. These three tools are elaborated upon in the following paragraphs.

IBM AI Fairness 360 has been listed by Cirillo et al. [36] as a recent development in the detection and mitigation of bias [36]. Developed by IBM, it provides a wide range of tools to facilitate bias mitigation throughout the lifecycle of an AI model [70]. AIF360 is equipped with over 71 bias detection indicators, which allow users to identify various forms of bias in models and datasets. The toolkit incorporates nine bias mitigation algorithms, divided into pre-, in-, and post-processing techniques. The toolkit includes an interactive online application that makes it simple for non-programmers to experiment with bias mitigation and detection. Its flexible architecture allows for ongoing expansion and the addition of new solutions. Furthermore, tutorials and documentation are available to help users optimize the toolkit's functionality to ensure fairness in AI systems [24]. Focusing on protected attributes such as gender and race, and measuring fairness using metrics, the study by Chen et al. [34] sought to reduce bias in ML classifiers by applying and evaluating various representative bias mitigation techniques on multiple benchmark datasets using the AIF360 toolkit [34]. The study by Martini and Berton [70] evaluated the post-processing techniques of AIF360, with a particular focus on the EqOddsPostprocessing method, through its application on a dataset pertaining to stroke prediction. Among other post-processing methods, this strategy was found to reduce bias but have the least impact on model accuracy and precision. [70].

[5]https://github.com/KenSciResearch/fairMLHealth
[6]https://fairlearn.org/
[7]https://github.com/google/ml-fairness-gym

Fairlearn offers metrics and techniques analogous to those of AIF 360 for assessing model equity [98]. It is a Python library developed by Microsoft. Fairlearn's ability focuses on decreasing unfair inequalities in model predictions [70]. As part of Fairlearn's fairness evaluation tool, the MetricFrame class evaluates performance across multiple demographic groups, allowing for a separate analysis of a model's performance for each group. The toolkit also includes a set of fairness metrics that facilitate the measurement of bias, such as equalized odds difference and demographic parity difference. Analogous to AIF 360, Fairlearn is equipped with bias mitigation strategies that comprise the entire AI lifecycle. These involve pre-processing methods such as Correlation Remover, in-processing methods like Exponentiated Gradient, and post-processing methods including Threshold Optimizer. Fairlearn's interoperability with prominent Python libraries, such as scikit-learn, TensorFlow, and PyTorch, facilitates the management of a diverse range of ML models [106]. In order to minimize algorithmic bias in cardiovascular predictive models, Sufian et al. [98] employed Fairlearn in conjunction with an algorithm known as Capuchin, as well as adversarial debiasing and equalized odds post-processing techniques [98]. Fairlearn was also integrated by Karim et al. [60] to build for a fairer heart disease prediction algorithm [60].

FairMLHealth has been referenced as a popular fairness library in the publication by Xu et al. [108] concerning algorithmic fairness of computational medicine [108]. Building on the fundamental ideas introduced by Ahmad et al. [16] in their paper on fairness in healthcare machine learning [16], the authors presented a tutorial of the FairMLHealth library at the Conference on Knowledge Discovery & Data Mining in 2020 [21]. The Python toolkit, developed by KenSci, has been specifically designed for healthcare applications. Its purpose is to audit bias in ML models. By comparing metrics such as accuracy, calibration, and error rates, the toolkit enables users to assess the predictive performance of the model across different patient subgroups. In contrast to general-purpose fairness libraries, the library has been developed to address the challenges posed by the complexity of healthcare data, including the presence of diverse patient groups and biased observational datasets. FairMLHealth employs a healthcare-focused and cohort-based reporting approach, thereby establishing a useful link between the practicalities of medical ML nd theoretical concerns regarding fairness [21].

These fairness toolkits offer practical aids for identifying and mitigating bias, fostering the establishment of responsible AI applications. Nonetheless,

the impact of these tools is contingent upon the prevailing legal and ethical framework within which they are employed. To comprehend the present legal framework in which these tools are employed, the subsequent section explores the regulatory environment concerning AI, both prior to and following the implementation of the AIA.

2.5 Regulatory Landscape before and after the EU AIA

Prior to the implementation of the EU AIA, pre-existing EU legislation served as safeguards against bias and discrimination in AI systems. Discrimination was prohibited under Directive 2004/113[8], and Articles 20 and 21 of the EU Charter of Fundamental Rights[9], but these are difficult to enforce because they require clear proof of unfair treatment. In addition, Article 22 of the GDPR[10] offers individuals the right to human intervention and transparency, and establishes protections against automated decision-making. A key legal consideration when incorporating debiasing methods is whether changes to data, as well as model changes, fall within the scope of any regulatory frameworks and laws. Until the AIA was in place, there was no specific legal frame in place to control the collection, selection, or modification of training data used for AI models to reduce bias. Under the GDPR, legal rules apply mainly when the training data contains personal data. These required a valid reason, such as informed consent, contractual requirement, or legitimate interest. Furthermore, while Article 9(2)(g) allowed for processing in the public interest, Article 9 of the GDPR required explicit consent, making it more challenging to mitigate bias in sensitive data [81]. Given this prior absence of an adequate regulatory framework which also captures the ethical requirements properly highlights the demand for AI to be overseen in its creation and application [35]. Kocak et al. [63] also stated in their paper about bias in artificial intelligence for medical imaging that, *"ultimately, the ethical use of AI, including the mitigation of bias, needs to be addressed in regulations to protect patients"* [63].

Therefore, the AIA seems promising, as it aims to regulate the entire lifecycle of AI. In fact, its risk-based classification of AI confirms that the fundamental rights and safety of patients are paramount when AI systems are used in such a critical area as medicine. The AIA, as a comprehensive legal framework for creating and deploying AI-driven systems, is intended

[8] https://eur-lex.europa.eu/eli/dir/2004/113/oj/eng
[9] https://eur-lex.europa.eu/eli/treaty/char_2016/oj/eng
[10] https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

to address these aforementioned limitations and reduce ethical and legal concerns in place [77]. Consequently, it is essential to comprehend the impact of the AIA within the regulatory context.

In fact, the AIA is binding throughout the EU [46] and has come into force in each member state on August 1, 2024, with the provisions of the AIA being incorporated gradually by the implementation period [12]. While directives only set objectives that each Member State must meet, regulations such as the AIA translate the objectives into national law [46]. To assure compliance, the AIA requires each Member State to establish three key authorities. The Notifying Authority, based on Articles 28(1), is responsible for appointing and supervising the Conformity Assessment Bodies responsible for certifying AI systems, while the Market Surveillance Authority, based on Article 74, keeps an eye on AI goods to make sure they comply with EU standards. In addition, high-risk AI systems listed in Annex III are subject to fundamental rights obligations enforced by a national public authority on the basis of Article 77 (1) [6]. By November 2, 2024, each Member State should have established and publicly listed the authorities and bodies in charge of the protection of fundamental rights. Subsequently, they are required to designate notifying authorities and market surveillance authorities to submit the names of these authorities to the Commission and make their contact information available to the public by August 2, 2025. [12]. Austria has already taken the right steps in this direction. The Austrian Regulatory Authority for Broadcasting and Telecommunications (RTR[11]) has been appointed as the responsible body for the AI Service Desk [1]. It will serve as a central hub for citizens, organizations and businesses, providing guidance on the implementation of the AIA. Its goal is to increase legal certainty and transparency and to provide expert support for compliance with the law. Whereas both the market surveillance authority and the notification authority have not yet been appointed [6].

3 Methodology

In this section, the methods of the thesis are outlined to illustrate the process of this work. This thesis leverages two distinct methodologies. Firstly, the IRAC method [28] is employed to conduct a legal analysis on a use case of interest, applying the EU AIA. The subsequent section delineates this use case. Secondly, the Design Science Method [85] is used

[11] https://www.rtr.at/rtr/service/Startseite.en.html

to implement detection and mitigation strategies for gender bias, thereby enhancing fairness in the prediction of CVD.

3.1 Definition of High Risk Use Case

The foundation of this thesis is built on the following use case that illustrates a high risk of AI deployment. According to Recital 46 [9] of the EU AIA, AI systems are classified as high-risk when they *"have a significant harmful impact on the health, safety, and fundamental rights of persons"* [9]. The use case of interest is outlined below:

In this high-risk use case scenario, a hospital is preparing to implement an AI-powered diagnostic tool for cardiovascular diseases. Specifically, medical professionals utilize the AI system when diagnosing cardiac diseases based on various types of medical and patient data gathered during different medical examinations. The overall goal of the AI system is to improve the speed and accuracy of medical decision-making by supporting doctors. To guarantee reliable and fair performance of the algorithms used, data quality and efficient data governance must be prioritized in the development and implementation of the AI system. The system should be able to deliver consistent predictions across various patient demographics. Moreover, the system's transparency is a crucial factor, as it allows medical staff to comprehend the AI's decision-making process and, if necessary, counteract it in order to provide the best possible patient care.

The use case presented falls under the EU AIA as it involves an AI system used as a diagnostic tool in the medical field, directly affecting the health and safety of patients. AI systems used as safety features in products subject to the Medical Device Directive (MDD) according to Regulation (EU) 2017/745[12] are classified as high risk under Article 6(1) [4]. Its classification as high-risk is justified by the significant potential harm to safety and fundamental rights, such as the equal treatment of all patients.

A major problem, however, is that much of the training data in this domain could potentially be biased, often overrepresenting male patients. Because of the bias in the data, the AI system may be less accurate for different genders, which could lead to oversights or misdiagnoses. Such outcomes potentially exacerbate existing gaps in healthcare quality and access, and undermine the idea that everyone, regardless of gender, should

[12] https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng

have equal access to healthcare. Therefore, the AIA, as a regulation, needs to be investigated to determine its implications for the delineated use case.

3.2 IRAC Method

Issue, Rule, Application, and Conclusion, also referred to as IRAC, is a common way to approach legal issues [28]. This framework has been utilized for several decades in the context of legal research. It is seen as a process of thinking that is intended to guide and improve legal reasoning [29]. The *issue* formulates the key issue in the factual situation using the relevant legal rules and principles [28]. In the context of this work, the *issue* element represents the evaluation of the use case scenario. The *rule* in the framework implies identifying the relevant provisions of the applicable law and breaking them down into elements, as well as using the definitions given in the law. In this work, this is the identification of all rules applying to the high-risk AI system outlined in the use case. In the context of the IRAC method, the concept of *application* connects the defined use case to the theoretical law. This part aims to examine the impact and applicability of the legal rules concerning the specified use case. The schema's final step, the *conclusion*, should offer a resolution to the legal issues of the specified use case that is firmly grounded in legal statutes [28].

For the purpose of this legal analysis, it is essential to utilize official sources, including the document of the AIA itself and reports issued by the European Commission. Additionally, this thesis utilizes the Future of Life Institute's website[13], to access the full text of the AIA through their EU AI Act explorer to identify particular provisions and obtain current information on its implementation. It is crucial to always keep track of the most recent legislative developments, as the AIA is still evolving and is being implemented in stages. The results of the legal analysis provide a summary of the concepts that must be met to ensure full compliance in order to develop an AI model that minimizes gender bias in CVD prediction.

3.3 Design Science Method

The following section examines the Design Science Method, which describes the creation and analysis of artifacts in context [107]. Regarding this thesis, the concept of design science involves assessing and comparing approaches to identifying and mitigating bias in AI models, especially in diagnosing

[13]https://artificialintelligenceact.eu/

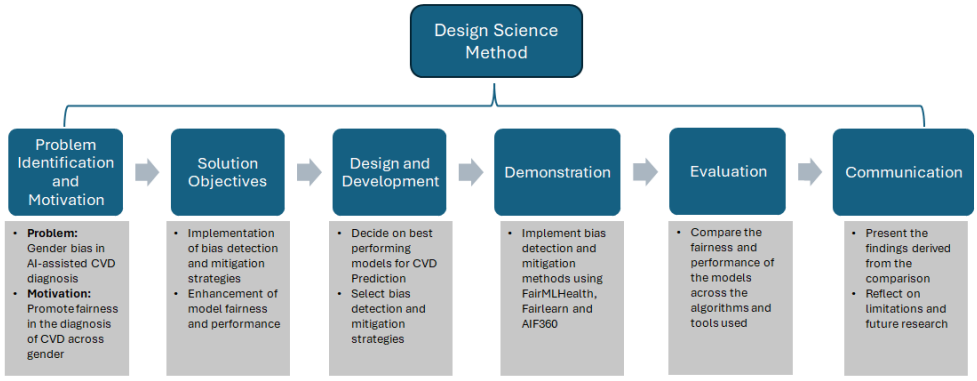


Figure 2: Design Science Method

CVD in the healthcare sector. The techniques for gender bias detection and mitigation are the artifacts in this regard. According the design science methodology of Peffers et al. [85], the following steps are employed, as illustrated in Figure 2:

Problem Identification and Motivation: The initial stage encompasses the identification of gender bias in AI models utilized for the prediction of CVD. To support the relevance and motivation of this work, a systematic literature review was conducted. The review focuses on the presence of gender bias in CVD prediction. To ensure both relevance and currency, the scope is limited to studies published within the last ten years, with an exclusive emphasis on the healthcare domain. The literature search was primarily conducted using Google Scholar[14] and databases from reputable medical and technical journals. To identify relevant papers, targeted keyword combinations were employed, incorporating the following terms: *gender bias in AI-assisted cardiovascular disease prediction, fairness in heart disease prediction, bias identification in CVD prediction, bias mitigation in CVD prediction, identifying and mitigating bias in heart disease prediction, and enhancing fairness in CVD prediction.*

Solution Objectives: In addressing the identified problem, the formulation of objectives for the solution, derived from the research, is crucial. In this particular instance, the solution objectives are defined as follows: the implementation of bias detection and mitigation, thereby enhancing the overall fairness and accuracy of algorithms.

[14]https://scholar.google.com/

Design and Development: This phase is oriented towards the exploration of approaches and instruments for the identification and mitigation of gender bias in ML models for the purpose of CVD prediction. Firstly, before starting technical implementation, it is necessary to acquire appropriate data. In the context of this thesis, it is imperative to ensure that the datasets leveraged contain characteristics that may introduce bias, in this particular instance, gender bias. Therefore, it is imperative that the patients' gender be specified. Furthermore, a dataset is required that comprises any data related to the health of the heart or associated risk factors, including diabetes, lack of physical activity [39], smoking and high blood pressure [18], that contribute to the development of CVD. Ultimately, the selection of an appropriate dataset for this work is determined by the size of the dataset, as well as its quality. Repositories such as Kaggle[15], the UCI Machine Learning Repository[16], HealthData[17] and Physionet[18] offer a substantial repository of medical data in this domain. As part of the design and development phase, a review focusing on identifying the most effective predictive models for CVD was conducted. A total of 30 studies published in the past five years were analyzed to determine which algorithms are most effective. These results are presented in Table 1, which summarizes the datasets used, the algorithms applied, and the corresponding performance metrics. Subsequently, the FairMLHealth tool, designed specifically for healthcare applications, is employed for the purpose of detecting bias. To facilitate a comparative analysis, Fairlearn, as a common bias detection tool, is also utilized. FairMLHealth is not equipped with any bias mitigation techniques. Therefore, mitigation is performed by leveraging Fairlearn and AIF360.

Demonstration: In accordance with the design and development step, the selected datasets, models, and bias detection and mitigation strategies determined are operationalized during the demonstration phase.

Evaluation: The previous processes are the foundation for a thorough examination of the artifacts created. This stage involves the evaluation and comparison of the effectiveness of different bias detection and mitigation techniques across various tools and algorithms applied. Furthermore, an evaluation is conducted to ascertain the impact on model performance and the extent to which the model has been enhanced in terms of fairness.

[15]https://www.kaggle.com/
[16]https://archive.ics.uci.edu/
[17]https://healthdata.gov/
[18]https://physionet.org/

Communication: The final step in the process is the presentation and documentation of the results. This involves thoroughly explaining the findings derived from the technical implementation, which includes assessing the fairness tools applied to the different models as well as reflecting on the shortcomings and future directions for further work in this field.

4 Legal Analysis

Advancements in AI are rapidly influencing the healthcare sector, augmenting efficiency and effectiveness through improved diagnostic accuracy. This development carries a significant positive impact on the domain of cardiology, especially with regard to the diagnosis of CVDs [35]. However, there are significant challenges associated with the application of these emerging solutions in clinical and patient care environments that demand thoughtful consideration of legal and regulatory issues [72].

In the context of this thesis, the objective is to establish a model that produces fair outcomes and handles the detection and mitigation of gender bias in AI-driven cardiovascular disease diagnosis. It is imperative to be aware of the regulatory landscape in place for compliance, as AI-driven decisions in healthcare fall under the high-risk category in the EU AIA [4]. Consequently, compliance with the corresponding provisions in place is compulsory, especially to ensure patient safety and prevent discrimination among different patient groups. For this purpose, this chapter employs the IRAC Method [28] to the previously defined use case scenario described in Subsection 3.1 for an assessment from a regulatory perspective. Beforehand, the ethical and legal concerns, as well as the regulatory efforts, that preceded the introduction of the AIA will be outlined in the next section.

4.1 Ethical and legal concerns in AI-driven healthcare

Not only are clinical risks a critical concern, but regulatory violations associated with anti-discrimination laws also arise when biased models are used. There is no question that ethical and legal requirements are closely intertwined when it comes to avoiding bias when implementing AI-driven healthcare [35]. AI systems can affect core values of humanity that may infringe fundamental rights such as human dignity, privacy and personal data protection, and non-discrimination. Consequently, they hold the potential to result in unforeseen harm to people's lives and health [44]. In

the medical context, the ethical key principles such as patient autonomy and the right to informed decision-making are of vital importance. Moreover, in healthcare ethics, justice refers to eliminating inequalities and ensuring universal access to healthcare resources, opportunities, and treatments [72]. Risks such as discrimination, privacy invasion, and opaque decision-making are of particular significance in the healthcare sector. Patients' vulnerability and reliance on accurate technology affect their independence, dignity, and trust. Their concerns are amplified by the life-threatening potential consequences of faulty AI, its opacity and its dependence on extensive health data [103]. From an ethical perspective, biased AI violates the core medical ethical principles of justice and non-maleficence, which call for equity and harm prevention, respectively [35]. Consequently, governments and public authorities should use ethical principles to guide the adaptation of regulatory frameworks to rapidly emerging AI technologies [72].

These concerns, with a particular emphasis on the realm of healthcare, underscore once more the necessity for an appropriate legislative framework to address the potential harms associated with the integration of artificial intelligence in healthcare. Prior to the enactment of the AIA, the existing array of European legislation offered only a limited and incoherent protection framework.

4.2 Regulatory Assessment of the Use Case

The foundation of this investigation is the AIA introduced by the European Commission. The aforementioned use case described in Section 3.1 is assessed by utilizing the IRAC method [85].

Issue The fundamental issue under consideration is to respond to the first research question in Section 3.1 in accordance with the use case of clinicians using an AI system to diagnose CVD. Of particular interest are the guidelines for processing such sensitive data, namely patient health data, and the corresponding patient rights in this context, such as fair treatment and transparent and explainable decisions.

Rule Looking at the structural framework of the AIA, especially Chapter III, which encompasses all aspects concerning high-risk systems, is of particular interest. The first section of the chapter establishes the classification rules for AI systems of high-risk. The subsequent provision is applicable to the previously defined use case:

Classification Rules for High-Risk AI Systems: According to *Article 6(2)* of the AIA, AI systems listed in Annex III should be classified as high-risk, even if the conditions listed in *Article 6(1)* do not apply. Furthermore, *Article 6(3)* refers to high-risk systems where the system in question "*poses a significant risk of harm to the health, safety or fundamental rights of natural persons*" [4]. *Annex III, paragraph 5a*, explicitly regulates AI systems, that are incorporated in health care services [2].

The second section of the AIA covers the requirements for AI systems posing high-risks. The section begins with *Article 8*, which states that a high-risk system, consistent with its purpose and the state of the art in AI, must meet all the requirements specified in this section [5]. This phase of the analysis aims to identify all articles pertinent to data management and governance, with a particular focus on the crucial aspect of transparency.

Data and Data Governance: The subject of data and its governance is only explicitly addressed in *Article 10* of the AIA [3]. Strict data governance and quality criteria must be followed for high-risk AI systems that use data for training, validation, or testing. To avoid discrimination and protect fundamental rights, safety, and health, providers must use data management procedures, such as bias detection and mitigation.

Transparency: The AIA establishes criteria aimed to achieve transparency on multiple levels. Thus, a number of articles are dedicated to addressing transparency in high-risk systems. Regulators, that is, e.g. the government and the designated authorities, must be able to trace and evaluate system performance and design according to *Articles 11 and 12*. Most importantly, *Article 13*, specifically entitled Transparency, ensures that users receive clear explanations of the AI's limitations and their intended use. Additionally, human oversight is a requirement specified in *Article 14* and is mandated to mitigate overreliance and preserve the user's ability to override the system when necessary. Together, these provisions aim to foster a more trustworthy environment, which is critical to the subsequent establishment of fair and transparent algorithms.

The third section delves into the responsibilities of providers,

users, and other stakeholders engaging with high-risk AI systems. The focus of this analysis is more on the implications for patients' rights, fair and transparent outcomes, and the performance of AI systems, and less on the hospitals and clinicians who deploy them. However, a more thorough examination of this chapter is beyond the scope.

Application According to *Article 6(2)* and Annex III of the AIA, a medical AI system that predicts CVDs is assigned to the high-risk category of AI systems. Consequently, the AI-powered CVD diagnostic system is required to adhere to the strict rules described in Chapter III of the AIA in order to establish fairness and trustworthiness for patients. Subsequently, the succeeding paragraphs elaborate on the application of each relevant article of the AIA to the defined use case.

Data and Data Governance (Article 10): As AI-powered medical diagnosis systems predominantly rely on data-driven learning models, the quality of the data is a significant factor in the algorithm's performance. Article 10 from the AIA addresses this factor and establishes rules for the use of data and data governance. Paragraph 3 requires the dataset utilized for the training, validation, and testing of the medical CVD diagnosis tool to be representative, relevant, and free of errors. The criteria of representativeness imply that different groups of patients of different sex or race must be equally represented in the health data. Furthermore, according to paragraph 4, the dataset employed must be aligned with the actual environment and consider the context in which the system will be implemented. Moreover, paragraph 5 of Article 10 allows the processing of *"special categories of personal data "* [3], such as health data, for the purpose of bias detection and mitigation techniques to the extent strictly necessary [3]. In the context of processing health data for the purposes of bias detection and mitigation, it is imperative to implement appropriate safeguards, including access control mechanisms, to ensure the protection of fundamental rights and the autonomy of patients.

Technical Documentation (Article 11): The provision of technical documentation is a prerequisite for the utilization of the system by clinicians for CVD diagnostic support. The technical documentation must include at least the aspects delineated in Annex IV and demonstrate compliance with the requirements established in Chapter III of the AIA. Annex IV delineates the following specifications for the Technical documentation: It should include explanations of the

system's development, training, validation, and performance across various demographic groups. The documentation is required to explicitly describe the provenance and characteristics of the data, as well as the techniques for detecting and mitigating potential bias. Additionally, it must detail the human oversight mechanisms that enable clinicians to interpret and, when necessary, override the outputs of this high-risk medical AI system. Continuous evaluation of performance indicators and equity for all patient groups must also be included in the post-market surveillance strategy.

Record-Keeping (Article 12): For the purpose of traceability and comprehensibility of the results of the CVD diagnostic tool, it should be capable of automatically recording the event logs throughout the system lifecycle. Records should document when the medical staff started and stopped using the diagnostic tool, as well as the input data used and the reference database being accessed. This procedure should support the continuous monitoring of the medical device in use. It should also facilitate the detection of any abnormal deviations from the system's performance and provide clinicians with the ability to verify and audit the AI's outputs, thereby ensuring accountability for the decisions made.

Transparency and Provision of Information to Deployers (Article 13): High-risk AI systems must be sufficiently transparent to allow clinicians to understand and properly use the results and subsequently the diagnosis made by the AI system. For the CVD diagnostic tool, this would mean giving medical professionals insight into the model's decisions and what data the models used to reach their final conclusion. Providers are obligned to supply clear, thorough and easily readable information regarding the AI system's accuracy, limitations, particularly those that could impact specific groups or other underrepresented populations, and foreseeable risks, as well as human oversight procedures. These instructions must also clarify how providers can manage record-capturing responsibilities, maintain the system, and interpret results to ensure proper operation and compliance. In addition, they should explain the technical features of the tool, such as input data standards and performance data by demographic group, to identify and address potential bias for underrepresented groups. Therefore, Article 13 promotes the appropriate use of high-risk AI in a clinical setting by ensuring that healthcare professionals are knowledgeable and competent enough to

critically evaluate the results of the AI.

Human oversight (Article 14): When implementing an AI-based CVD diagnostic tool, Article 14 ensures that human oversight is built into the system's architecture to protect patients' health and fundamental rights. This means that the system must have features and interfaces that allow physicians to view, analyze, and control the AI's output, especially when there are irregularities or unexpected results. Medical professionals need to be trained to prevent automation bias, and be fully informed about the system's capabilities and limitations, given the clinical situation. This will ensure that they continue to play a critical role in decision-making. The degree of autonomy of the AI system and the risk of misdiagnosis, especially for underrepresented patient groups, must be reflected in the oversight mechanisms. To preserve human control over patient care, clinicians must be able to interrupt, modify or override the diagnosis of the system.

Conclusion Based on the applicable regulations and how they apply to the use case, the following conclusions can be drawn regarding data governance and transparency in the context of the AI-assisted CVD tool. The AIA provides a comprehensive set of guidelines for the regulation of high-risk systems, such as the CVD diagnosis system from the defined use case. Additionally, the aspects of particular interest, the data used, data governance, and the transparency of the system are all regulated in some manner. As aforementioned, *Article 10* regulates data and data governance in high-risk AI systems. Requirements such as complete, accurate, and flawless data foster accuracy and fairness in predictions. Moreover, fundamental rights and non-discrimination are values that must be preserved. More crucially, this provision demands that providers implement bias detection and mitigation methods throughout their data management processes. The objective of enhancing transparency is generally a high priority as its explicitly subject to Article 13 and further supported by other articles. While *Article 13* promotes transparency through mandatory instructions to improve the comprehensibility of the diagnostic tool for the user, *Articles 11, 12 and 14* implicitly enhance transparency within the AI system. The requirement to provide technical documentation, the recording of each diagnostic session, and the establishment of human oversight measures all contribute to greater transparency. These provisions are all designed to provide greater insight into the processes

and decisions of the diagnostic tool. As a result, the improved comprehensibility of the AI system should lead to greater confidence in the diagnosis made by the system.

The legal assessment of the AIA employing the IRAC Methodology furnished an overview of the regulations that pertain to the specified use case. Although the IRAC Method incorporates an Application section, the specific actions that deployers and users of the system are obligated to implement remain vaguely defined within the regulatory framework. Nevertheless, it is possible to derive a set of guidelines for the technical implementation of this work. Initially, the data will be utilized exclusively if it is considered representative, relevant, and free of errors. The necessity of incorporating bias detection and mitigation techniques establishes the foundation for the subsequent technical section. Given the lack of clarity concerning the specific methods for detection and mitigation, this thesis will encompass a range of tools to ascertain the most efficacious. To ensure comprehensibility, extensive technical documentation will be conducted throughout the technical section as mandated.

5 Data Preparation and Model Development

The subsequent sections commence with an introduction to the datasets that were utilized in the study. This chapter outlines the retrieval, pre-processing, and exploration of the data. Additionally, the models used for CVD prediction will be presented.

5.1 Description of Datasets

This section delineates the description of the three Cardiovascular Disease datasets employed in this work: The Cardiovascular Disease Dataset[19] and the Heart Failure dataset[20], both from Kaggle, as well as the Mendeley dataset[21]. The latter comprises multiple smaller datasets aggregated into one. For the purpose of evaluating the generalizability and robustness of the findings, these datasets were selected on the basis of their varying gender dominance and distribution and overall sizes. The datasets were first examined to gain a better understanding of their structure, the distribution of important variables, and possible sources of bias before moving further with

[19]https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset
[20]https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data
[21]https://data.mendeley.com/datasets/dzz48mvjht/1

model building and fairness assessment. For this exploratory data analysis, the datasets were used in their clean, unprocessed state. This analysis focused on the distribution of gender, the protected attribute, and the target variable to find any imbalances that could potentially introduce model bias. In the context of this study, identical models were trained on all three datasets, and the same bias detection techniques were applied to each. For the purpose of maintaining consistency throughout this thesis, the datasets utilized are referred to as the Kaggle CVD dataset, the Composite Heart Failure (HF) dataset and the Mendeley dataset.

5.1.1 Description of the Kaggle CVD Dataset

This Cardiovascular Disease dataset is an extensive collection of patient data, providing information about their heart health. The comprehensive dataset encompasses 12 variables of health and medical data, including the target variable derived from a sample of 70,000 patients. Unfortunately, the source of the data and the date of its capture are not specified. The author of this dataset is Svetlana Ulianova and the data is available to the public on Kaggle [22]. The variables are illustrated in Table 4 alongside their corresponding data types and scale. The dataset is divided into three distinct categories: The objective features represent factual information about the patient. The subsequent category comprises data that has been collected during patient examination. Finally, subjective features constitute information provided by the patient [7].

[22] https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

Feature	Data Type | Scale
Objective Features	
age	Numeric (Continuous) | days
height	Numeric (Continuous)
weight	Numeric (Continuous)
gender	Categorical (Binary) | 1-2
Examination Features	
ap_hi	Numeric (Continuous)
ap_lo	Numeric (Continuous)
cholesterol	Categorical (Ordinal) | 1-3
gluc	Categorical (Ordinal) | 1-3
Subjective Features	
smoke	Categorical (Binary) | 0-1
alco	Categorical (Binary) | 0-1
active	Categorical (Binary) | 0-1
Target Variable	
cardio	Categorical (Binary) | 0-1

Table 4: Features & Data Types (Kaggle CVD Dataset)

The subsequent list offers a concise summary of the features present in the data, together with their corresponding descriptions.

Objective Features: The patient's `age` is recorded in days, thereby providing an exact numerical value. However, for the purpose of analysis and interpretation, it is more convenient when converted into years. The `height` of the patient is expressed in centimeters and represents the height of the individual determined during the examination. `Weight` is expressed in kilograms and represents the patient's bodily mass. `Gender` is commonly represented as a categorical variable, often with binary values, denoted by 1 for women and 2 for males in this dataset.

Examination Features: The primary two features that are measured in medical investigations are systolic and diastolic blood pressure (`ap_hi and ap_lo`), which are expressed in integers. The categorical variable of `cholesterol` is represented on a three-level scale. The numerical values 1, 2, and 3 correspond to normal levels, levels above normal, and levels significantly above normal, respectively. Analogous to cholesterol, glucose levels are also classified by the `gluc` variable on this three-point scale representing normal, high, and considerably elevated blood sugar levels.

Statistic	Age	Height	Weight	ap_hi	ap_lo
Unique Values	28	109	287	153	157
Mean	52.84	164.36	74.21	128.82	96.63
Min	29.00	55.00	10.00	-150.00	-70.00
Max	64.00	250.00	200.00	16020.00	11000.00

Table 5: Summary Statistics of Numerical Variables (Kaggle CVD Dataset)

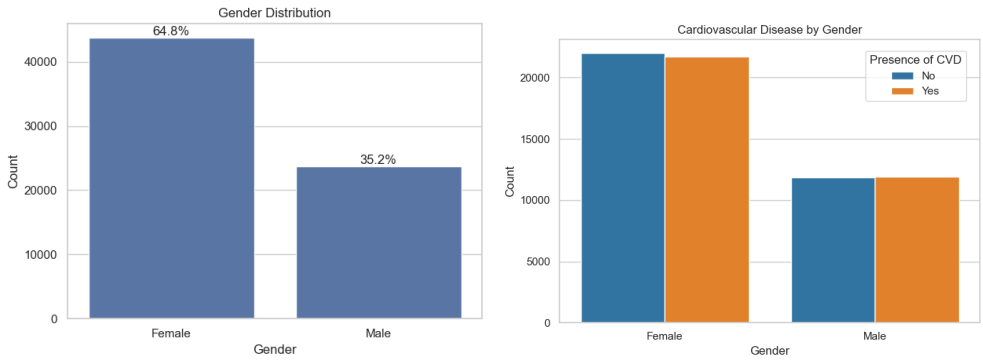
Group	Variable	Normal	Above normal	Well above normal
Categorical Variables	Cholesterol	50581 (75.0%)	9119 (13.5%)	7785 (11.5%)
	Glucose	57344 (85.0%)	4981 (7.4%)	5160 (7.6%)
Group	**Variable**	**No**		**Yes**
Binary Variables	Smoking	61520 (91.2%)		5965 (8.8%)
	Alcohol intake	63854 (94.6%)		3631 (5.4%)
	Physical activity	13287 (19.7%)		54198 (80.3%)
	CVD Presence	33865 (50.2%)		33620 (49.8%)

Table 6: Distribution of Categorical Variables (Kaggle CVD Dataset)

Subjective Features: The attribute of `smoke` serves as an indicator of an individual's smoking status, with a value of 1 denoting a smoker and 0 indicating a non-smoker. The feature `alco` comprises the presence of regular alcohol consumption. For the given variable, the integer 1 indicates regular alcohol consumption, while 0 signifies abstinence. The `active` attribute is indicative of the patient's exercise habits, with a value of 1 signifying an active lifestyle and a value of 0 indicating inactivity.

An overview of the summary statistics for numerical variables of the Kaggle CVD dataset is provided in Table 5. The data includes columns for `age` and `gender`, which pertain to demographic information. The remaining variables describe the individual's health status and lifestyle. The patients' ages range from 29 to 64 years, with an average of approximately 53 years. The patients exhibit a mean height of approximately 1.64 meters and an average weight of 74 kilograms. The majority of the variables employed to assess patients' health status are either binary or categorical in nature. Systolic and diastolic blood pressure are the only metrics that are presented in actual numerical values.

Table 6 presents, for each value type, the number of observations and their proportion relative to the total. The cholesterol and glucose levels are within the established normal range for the majority of the patients. The data also indicates that the majority of patients do not smoke or consume alcohol.



(a) Distribution of Gender

(b) Distribution of CVD by Gender

Figure 3: Gender Distribution and CVD Occurrence by Gender (CVD Kaggle Dataset)

According to the data derived most of the patients adhere to a relatively healthy lifestyle, as evidenced by the engagement in physical activity, absence of smoking and alcohol consumption. Despite this fact, nearly half of the patients are diagnosed with a CVD.

The protected attribute of the subsequent fairness evaluation, `gender`, is unevenly distributed, with about 64.8% females and 35.2% males illustrated in Figure 3a. With regard to the target variable, the gender-wise distribution of CVD cases is almost equal, as presented in Figure 3b. This dataset will serve as the foundation for the evaluation of fairness in the diagnosis of CVD with regard to gender.

5.1.2 Description of the Mendeley Dataset

The Cardiovascular Disease Dataset from Mendeley Data encompasses 14 columns of patient data, including the target variable, indicating the presence of CVD. Data from 1,000 anonymized patients was obtained from a multispecialty hospital in India. The dataset was published in 2021 by the authors Bhanu Prakash Doppala and Debnath Bhattacharyya. Furthermore, the dataset is licensed under a CC BY 4.0 license, thereby permitting its sharing, duplication, and modification, on the condition that the authors are adequately credited. This dataset is intended to facilitate the early detection of CVD and the establishment of predictive ML models [8].

Table 7 provides a concise overview of the data types and their respective scales. The succeeding paragraph will describe each feature in a concise

Feature	Data Type | Scale
age	Numeric (Continuous) | years
gender	Categorical (Binary) | 0-1
chestpain	Categorical (Nominal) | 0-3
restingBP	Numeric (Continuous)
serumcholestrol	Numeric (Continuous)
fastingbloodsugar	Categorical (Binary) | 0-1
restingrelectro	Categorical (Nominal) | 0-2
maxheartrate	Numeric (Continuous)
exerciseangia	Categorical (Binary) | 0-1
oldpeak	Numeric (Continuous)
slope	Categorical (Ordinal) | 1-3
noofmajorvessels	Numeric (Discrete) | 0-3
target	Categorical (Binary) | 0-1

Table 7: Features & Data Types (Mendeley Dataset)

manner to facilitate a more thorough understanding of the dataset.

The `age` is expressed in years. The `gender` is denoted in the binary categorical code 0 and 1, indicating 0 for female and 1 for male patients. The remaining variables are all related to the cardiovascular health status of the patient, starting with the `chest pain type`, which is differentiated in four distinct categories: typical angina, atypical angina, non-anginal pain, and asymptomatic. The `resting blood pressure` in this data set ranges from 94 to 200 mmHg, while the level of `serum cholesterol` varies from 126 to 564 mg/dL. The attribute of `fasting blood sugar` indicates whether the patient's measured blood sugar was above 120, thereby establishing a binary classification of 0 as false and 1 as true. The `resting electro` variable is encoded as 0, 1 or 2 to represent different diagnostic outcomes from an electrocardiogram when the patient is at rest. The `maximum heart rates` of patients have been recorded as actual values ranging from 71 to 202 beats per minute. `Exercise angina` is classified as a binary feature, with 1 indicating the presence of exercise-induced angina in a patient and 0 indicating its absence. The continuous numeric attribute `oldpeak` is defined in the range 0 to 6.2 as an indicator of ST-depression levels relative to rest, caused by exercise. The `slope` variable is of a categorical nature, comprising three levels: upsloping, flat and downsloping. Each of these levels is assigned a numerical value from 1 to 3 that corresponds to the slope of the peak exercise ST segment. The next feature is defined as a continuous variable spanning from 0 to 3, denoting the `number of major vessels` obtained

Statistic	Age	RestingBP	SerumChol	MaxHR	Oldpeak
Unique Values	61	95	344	129	63
Mean	49.24	151.75	311.45	145.48	2.71
Min	20.00	94.00	0.00	71.00	0.00
Max	80.00	200.00	602.00	202.00	6.20

Table 8: Summary Statistics of Numerical Variables (Mendeley Dataset)

through fluoroscopy. Finally, the collection of cardiovascular health data is concluded with the binary `target` variable indicating a CVD with 1 and no CVD with 0.

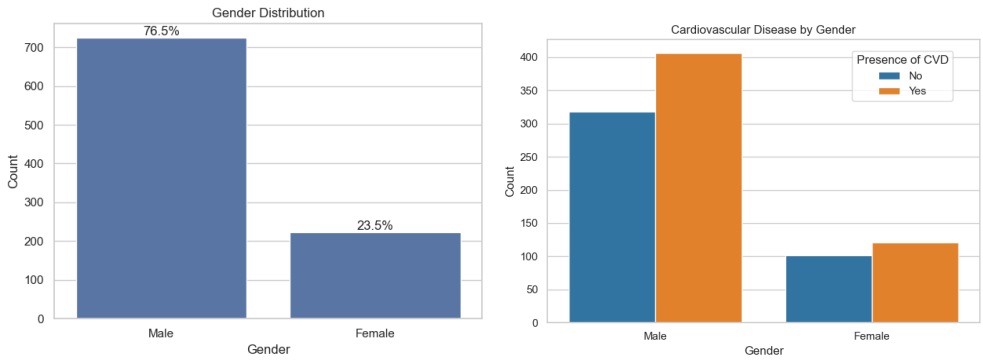
Table 8 illustrates some summary statistics of the numerical variables from the Mendeley Datasets. The mean age of patients in the Mendeley Dataset is 49 years, with the youngest patient being 20 years old and the oldest 80 years old. The resting blood pressure ranges from 94-200 mmHg, with a mean of 151.75 mmHg. According to the American Heart Association (AMA), such a measure is already considered to be hypertension, or high blood pressure [10]. The serum cholesterol levels range from 0 to 602, indicating either missing values or failed measurements, which must be addressed during the preprocessing stage. Given that the optimal cholesterol level is approximately 150 mg/dL, as indicated by the AMA, a mean level exceeding 300 mg/dL is associated with an elevated risk of CVD. Moreover, the maximum heart rate is between 71 and 202 beats per minute (bpm), with an average of 145.48. The mean value of the old peak variables is 2.71.

Table 9 presents insights into the categorical variables of the Mendeley Dataset. More specifically, the distribution counts for each categorical variable are displayed.

Group	Variable	0	1	2	3
Categorical Variables	Chest pain type	420	224	312	44
	Slope of ST segment	180	299	322	199
	Resting ECG results	454	344	202	-
Group	**Variable**	**No**		**Yes**	
Binary Variables	Fasting blood sugar	704		296	
	Exercise-induced angina	502		498	
	CVD Presence (target)	420		580	

Table 9: Distribution of Categorical Variables (Mendeley Dataset)

The most prevalent types of chest pain are typical angina (0) and non-anginal pain (2), which are experienced by 420 and 312 patients, respectively. Asymptomatic chest pain (3) was rarely observed in patients in the dataset. The slope of the peak exercise ST segment has three valid levels, labeled 1-3. The 180 entries indicating 0 are likely missing values, which will be addressed during the preprocessing phase. The majority of the patient slopes demonstrate a flattened or upsloping trend. The results of the resting electrocardiogram are within the normal range for the majority of patients, comprising 454 patients. The fasting blood sugar measurement revealed that the blood sugar level was below the threshold of 120 for the majority of patients. In approximately half of the patient cases, the presence of exercise-induced angina has been documented. Finally, the distribution of the presence of CVD is quite balanced in the Mendeley dataset, with 580 patients diagnosed with CVD.



(a) Distribution of Gender

(b) Distribution of CVD by Gender

Figure 4: Gender Distribution and CVD Occurrence by Gender (Mendeley Dataset)

While the previously examined Kaggle dataset exhibited a distribution skewed to females, the Mendeley Dataset predominantly consists of male individuals, with 76.5% of the patients being male, as shown in Figure 4a. Consequently, the higher representation of males in the dataset naturally leads to a greater number of male patients being diagnosed with CVD. However, both gender groups exhibited a higher occurrence of patients diagnosed with CVD compared to those without, as indicated in Figure 4b.

5.1.3 Description of the Composite Heart Failure (HF) Dataset

The HF Prediction Dataset is an aggregation of five distinct datasets. The Cleveland, Hungarian, Statlog, Longbeach VA, and Switzerland datasets have been consolidated into a composite dataset. The individual datasets can be found in the UCI Machine Learning Repository[23]. The aggregation resulted in 11 clinical features and target variables indicating the presence of heart failure. The data on Kaggle has already been cleaned of duplicates, leaving 918 instances in total. The composition of these five aforementioned datasets is credited to the Kaggle user named fedesoriano and was published in September 2021. Furthermore, the data is protected by the provisions of the Open Database License (ODbL) version 1.0[24].

Subsequently, Table 10 contains a representation of the data types and corresponding scales of the clinical features. Due to the similarity across the datasets in terms of features and their data types, as well as their

[23]https://archive.ics.uci.edu/

[24]https://opendatacommons.org/licenses/odbl/1-0/

Feature	Data Type | Scale
Age	Numeric (Continuous) | years
Sex	Categorical (Binary) | M/F
ChestPainType	Categorical (Nominal) | TA, ATA, NAP, ASY
RestingBP	Numeric (Continuous) |
Cholesterol	Numeric (Continuous) |
FastingBS	Categorical (Binary) | 0-1
RestingECG	Categorical (Nominal) | Normal, ST, LVH
MaxHR	Numeric (Continuous) |
ExerciseAngina	Categorical (Binary) | Y/N
Oldpeak	Numeric (Continuous) |
ST_Slope	Categorical (Ordinal) | Up, Flat, Down
HeartDisease	Categorical (Binary) | 0-1

Table 10: Features & Data Types (Composite HF Dataset)

Statistic	Age	RestingBP	Cholesterol	MaxHR	Oldpeak
Unique Values	50	67	222	119	53
Mean	53.51	132.4	198.8	136.81	0.89
Min	28.00	0.00	0.00	60.00	-2.60
Max	77.00	200.0	603.0	202.00	6.20

Table 11: Descriptive Statistics for Numerical Features (Composite HF Dataset)

scale, it is not necessary to elaborate on this overview table. Nevertheless, in contrast to the other datasets, the categorical features are provided as actual levels, labels, and not in numerical encoding.

Table 11 provides more detailed insights into the statistics of the numerical features of the Composite HF dataset. The youngest patient in the sample was 29 years of age, while the oldest was 77 years old. On average, the patient is approximately 54 years old in this consolidated dataset. The statistical analysis of the resting blood pressure and cholesterol variables indicates the presence of outliers on the lower bound, with a minimum of 0 observed. Both of these variables demonstrate a high mean value in terms of medical interpretation. The mean blood pressure at rest was found to be 132.4 mm/Hg, a value that is already considered to be high blood pressure. The serum cholesterol average in this sample is 198.8 milligrams per deciliter. The maximum heart rate has a mean of approximately 131 beats per minute and ranges from 60 to 202 beats per minute. The minimum value of the oldpeaks attribute is negative, suggesting the presence of outliers. This observation also clarifies the mean value to approximate zero. Furthermore,

Group	Variable	0	1	2	3
Categorical Variables	Chest Pain Type	46 (5.0%)	173 (18.8%)	203 (22.1%)	496 (54.0%)
	ST Slope	63 (6.9%)	460 (50.1%)	395 (43.0%)	-
	Resting ECG	552 (60.1%)	178 (19.4%)	188 (20.5%)	-

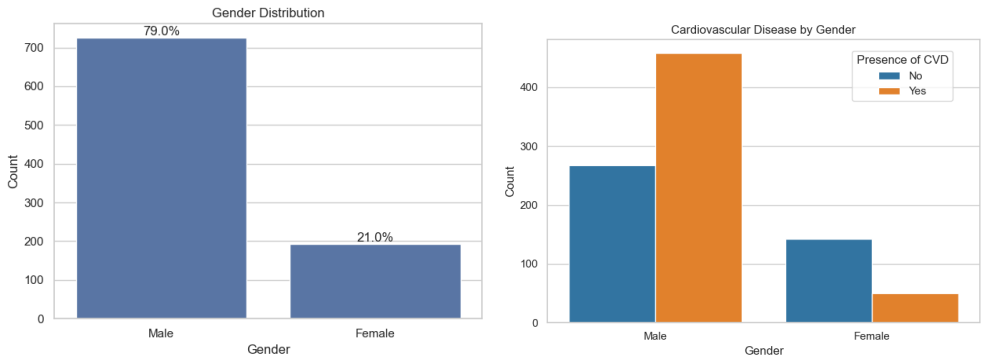
Table 12: Distribution of Categorical Variables (Composite HF Dataset)

Group	Variable	No	Yes
Binary Variables	Exercise Angina	547 (59.6%)	371 (40.4%)
	Heart Disease	410 (44.7%)	508 (55.3%)

Table 13: Distribution of Binary Variables (Composite HF Dataset)

Tables 12 and 13 below offer a refined interpretation of the distribution of the categorical variables in the consolidated dataset.

The most prevalent type of chest pain reported by patients is asymptomatic (ASY), followed by non-anginal pain (NAP) and atypical angina (ATA). Among the sample of patients, 50% exhibited a flat ST segment, while 43% demonstrated an upsloping segment. For the majority of patients, the electrocardiogram (ECG) at rest revealed normal results. In another fifth of patients, ECG examinations exposed wave abnormalities, and in the remaining fifth, the results indicated left ventricular hypertrophy. Exercise-induced angina is observed to be 40%, indicating that only 4 out of 10 individuals within the dataset experience this condition. The occurrence of heart disease is relatively balanced, with a ratio of 44.7% to 55.3%.



(a) Distribution of Gender

(b) Distribution of CVD by Gender

Figure 5: Gender Distribution and CVD Occurrence by Gender (Composite Heart Failure Dataset)

Figure 5a illustrates that this dataset is also male-dominant, with 79% of the sample consisting of males and 21% of females. The majority of male patients received a positive diagnosis of CVD, while only a minor fraction of the female patients were diagnosed with CVD, as depicted in Figure 5b.

5.2 Preprocessing of Datasets

The pre-processing stage aimed to unify the encodings of all three datasets to make the attributes comparable, while also ensuring that the measurements from examinations were clinically reasonable and treating missing and false data. During the preliminary inspection of summary statistics, implausible or false entries were identified and subsequently removed. Such entries included physiologically unrealistic blood pressure and heart rate values, as well as categorical codes that fell outside the established range of values. A set of preprocessing steps has been implemented across all data sources. The following paragraph delineates these steps. The processing of all datasets was conducted in Visual Studio Code using Python, leveraging the Pandas library[25] to ensure efficient data modification. Initially, each dataset was imported as a Pandas DataFrame to facilitate structured data processing.

As a first step, the identifier column was removed from the data collections because it served no analytical purpose. Furthermore, the gender attribute and the target variable were standardized to ensure consistent mapping. The gender attribute was set to `0 = female` and `1 = male`. The target variable was set to 0 for a negative CVD diagnosis and 1 for a positive CVD

[25] `https://pandas.pydata.org/`

diagnosis. To identify possible outliers and anomalies, descriptive statistics were explored to provide an overview of the variables' concentration, range, and overall distribution. In order to identify and manage implausible values, such as non-positive heart rate or blood pressure, zero serum cholesterol, or negative ST depression readings, plausibility checks were first performed. Consequently, quantile filters or the interquartile range (IQR) method were employed to construct robust ranges for continuous variables. Medically impossible values were either imputed or capped at acceptable thresholds using reliable sources, such as the AMA. Lastly, duplicate rows were eliminated. Subsequently, the data was stratified according to the target label and separated into an 80/20 train-test split. In all experiments, the test set was frozen and remained constant. According to the technical experiment, a variety of gender-ratio training subsets were developed, including balanced (50/50) and imbalanced (75/25 and 25/75) male-to-female ratios, for the purpose of fairness assessments. A balanced distribution is established globally for the target variable.

Finally, the preprocessed subsets were exported as a `.csv` file to enable efficient reuse in forthcoming analyses and modeling approaches. Although the datasets underwent the same sequence of preprocessing steps, some adjustments specific to each dataset were necessary due to their differing composition. The following subsections will shortly describe these steps.

5.2.1 Preprocessing of the Kaggle CVD Dataset

The Kaggle CVD Data was preprocessed according to the approach delineated by Bhatt et al. [25], with the objective of getting the most out of the model's performance. In the context of data conversion, the attribute age was converted into a value expressed in years. Patients between the ages of 30 and 65 were selected for further analysis. The gender variable originally encoded in 1 and 2 was mapped to 0 for the female patients and to 1 for the male patients. Extremes in height, weight, systolic blood pressure, and diastolic blood pressure were identified using the 2.5 and 97.5 percentile range. Bhatt's contributions also entailed the addition of two new variables to the data collection: mean arterial pressure (MAP) and body mass index (BMI). In addition, the authors employed binning on the variables of age, BMI, and MAP. The final dataset utilized for the modeling process comprises the following columns: gender, age, BMI_class, MAP, cholesterol, gluc, smoke, alco, active, and the target cardio [25]. For a more thorough exposition of the preprocessing process executed by Bhatt et al. [25], refer to the authors' paper. Furthermore, a dataset comprising a total of 30,000

observations, exhibiting a varying gender ratio, has been established for each experiment.

5.2.2 Preprocessing of the Mendeley Dataset

An initial examination of the descriptive statistics indicated an inconsistency in the variable slope. Based on the provided data description, the variable is comprised of three distinct levels ranging from 1 to 3, yet the minimum value documented was 0. Due to the limited data availability for this particular dataset, the invalid values were substituted with the most frequently occurring valid level. In regard of outliers, the following variables were subjected to the IQR approach: `resting blood pressure`, `serum cholesterol`, `maximum heart rate`, and `oldpeak`. An adjustment of the oldpeak variable turned out to be not required. However, with regard to the remaining variables, to ensure the comprehensiveness of the dataset, the data points that exceeded medical feasible thresholds were capped. For the purpose of the fairness evaluation experiments, the gender ratio was appropriately adjusted. Each subset contained 600 rows with target distributions balanced globally, while varying the male-to-female ratio.

5.2.3 Preprocessing of the Composite HF Dataset

For this aggregated dataset, the categorical attributes namely, `Chest Pain Type`, `RestingECG`, `Exercise Angina`, and `ST Slope` were transformed into numerical codes that reflected the corresponding label categories. In addition, the target variable was transformed into a binary variable. Subsequent to an initial examination of the data, it was determined that the columns labeled `RestingBP`, `Cholesterol`, `MaxHR`, and `Oldpeak` required outlier treatment. Employing the IQR method, the specified variables were examined, and the outliers were subsequently imputed with the median of the relevant observations. With regard to cholesterol, exceptionally high levels were also constrained at the upper threshold of the IQR. In the end, the cleaned data was used to create three subsets with different gender ratios while maintaining a total sample size of 600.

5.3 Application of ML Models

This subsection provides an overview of the models that achieved the highest predictive performance for CVD. Furthermore, it focuses on the systematic optimization of models, with the objective of ascertaining optimal model performance. To produce outputs that are consistent with the highest

performance documented for these models in related research.

The employment of bias detection and mitigation strategies is grounded in the use of CVD prediction on health data. Accordingly, four distinct algorithms are implemented: KNN and DT are used as traditional baseline models, RF is utilized as an ensemble model and as a deep learning algorithm: MLP. For a more thorough understanding of the algorithms employed, readers are directed to the work of Ali et al. [20], which contains a description of the utilized models. For purposes of experimentation, three distinct subsets were derived from each dataset, varying in gender distribution. The first subset demonstrates a male-dominant data collection consisting of 75% males and 25% females. In contrast, the second subset is skewed toward females, maintaining the same ratio. In each case, the overall balance of the target variable is ensured. Further, a balanced version of each dataset was created with an equal number of females and males in the sample. This setup of varying gender dominance allows us a systematic evaluation of its impact on model performance and, subsequently, the presence of gender bias. The following set of performance metrics is employed to assess the effectiveness of the distinct models:

- **Accuracy:** provides clarification regarding the overall correctness of the model by assessing the proportion of total correct cases with respect to all observations.
- **Precision:** concerns the quality of the positive predicted cases, indicating the extent to which the model successfully detects true positive cases.
- **Recall:** is considered as the ability to identify actual positive outcomes.
- **F1-Score:** is an indicator of the balance between precision and recall.

The performance metrics' mathematical equations are declared in the paper by Baghadi et al. [22]. With respect to the prediction of CVD, recall is of particular significance. As the most unfavorable outcome is the inability to diagnose CVD despite its presence. Additionally, the confusion matrix for each model is inspected to determine the actual number of true positives and negatives, as well as false positives and negative cases. Decisions regarding medical care, including those pertaining to the diagnosis of CVD, carry significant and potentially life-threatening implications. Consequently, relying on a single metric to assess algorithm performance could have severe consequences. Consequently, the optimal model is determined by prioritizing

the model recall. While maximizing recall is crucial, precision and accuracy should not undergo a substantial decline. Additionally, it needs to be ensured that the model does not become overly sensitive, thereby reducing the occurrence of false alarms. False alarms, in turn, also have the potential to result in erroneous patient treatments. Consequently, the selection of models for CVD diagnosis should exhibit high recall, as well as robust accuracy and precision.

For performance optimization, a systematic procedure was used to determine the optimal configuration of parameters such as the number of neighbors for KNN, the tree depth for DT, the number of estimators for RF, and the optimizer and architecture settings for MLP. Tuning yielded increased performance in the majority of cases. However, on some occasions, the baseline model produced the best outcome. As documented in the GitHub [26] accompanying the thesis, the details of the tuning can be found in the corresponding notebooks.

This process was implemented consistently for each dataset. The ensuing tables illustrate the performance of each utilized model across the scenarios of varying gender distribution. Furthermore, the bold values emphasize the most successful performance metrics across all models and data subsets.

[26] GitHub Repository

Metric	KNN			DT			RF			MLP		
	75/25	25/75	50/50	75/25	25/75	50/50	75/25	25/75	50/50	75/25	25/75	50/50
Accuracy	68.2%	70.6%	70.1%	70.8%	71.1%	71.3%	70.2%	70.8%	70.9%	71.4%	71.1%	**71.6%**
Precision	68.6%	70.9%	69.6%	70.1%	71.0%	71.9%	71.1%	72.1%	72.2%	**72.5%**	71.4%	72.3%
Recall	66.5%	69.6%	70.7%	**72.2%**	71.1%	69.5%	67.4%	67.1%	67.6%	68.6%	69.8%	69.6%
F1-Score	67.5%	70.2%	70.2%	**71.1%**	71.0%	70.7%	69.2%	69.6%	69.8%	70.5%	70.6%	70.9%

(a) Kaggle CVD Dataset

Metric	KNN			DT			RF			MLP		
	75/25	25/75	50/50	75/25	25/75	50/50	75/25	25/75	50/50	75/25	25/75	50/50
Accuracy	93.0%	89.0%	93.5%	90.5%	90.0%	94.0%	**95.5%**	92.5%	94.0%	92.0%	88.5%	92.5%
Precision	**97.2%**	91.0%	96.4%	90.8%	89.3%	93.3%	96.5%	94.7%	94.8%	92.4%	88.4%	95.5%
Recall	90.5%	90.0%	92.2%	93.1%	94.0%	**96.6%**	95.7%	92.2%	94.8%	94.0%	92.2%	91.4%
F1-Score	93.8%	90.0%	94.3%	91.9%	91.6%	94.9%	**96.1%**	93.4%	94.8%	93.2%	90.3%	93.4%

(b) Mendeley Dataset

Metric	KNN			DT			RF			MLP		
	75/25	25/75	50/50	75/25	25/75	50/50	75/25	25/75	50/50	75/25	25/75	50/50
Accuracy	**88.6%**	83.2%	87.5%	81.0%	82.6%	82.1%	88.0%	84.2%	83.2%	85.9%	82.1%	85.3%
Precision	90.9%	90.8%	**92.5%**	81.9%	79.7%	89.7%	87.0%	90.1%	85.9%	88.8%	87.1%	91.2%
Recall	88.2%	77.5%	84.3%	84.3%	92.2%	76.5%	**92.3%**	80.4%	83.3%	85.3%	79.4%	81.4%
F1-Score	**89.6%**	83.6%	88.2%	83.1%	85.5%	82.5%	89.5%	85.0%	84.6%	87.0%	83.1%	86.0%

(c) Composite HF Dataset

Table 14: Model performance across gender ratio scenarios for the three datasets

5.3.1 Kaggle CVD Dataset

Table 14a illustrates the accuracy as a performance metric for each specific gender distribution scenario across all algorithms applied to the Kaggle CVD dataset. The KNN model demonstrates the strongest performance across all metrics for the female-dominated subset, except for the recall, which exhibits slightly higher performance for the balanced dataset. In regard to the DT model, the balanced subset demonstrates the highest levels of accuracy and precision, while the male-skewed subset exhibits higher values for recall and F1-score. Moreover, the DT as well as RF demonstrate stable results across all three distinct scenarios. Furthermore, the RF model consistently yielded better performance for the balanced dataset, with equal representation of female and male patients. The stability of performance can also be observed for the MLP with only slight differences in the metrics. While the balanced subset achieved the greatest accuracy and F1-score, the male-dominated subset exhibited the highest precision, and the female-dominated subset had the highest recall. Overall, the MLP model produced maximum accuracy and precision, while the DT reached peak recall and F1-score in the present experimental setting.

Despite efforts towards optimization, none of the models applied could achieve an accuracy level exceeding 80%. Therefore, in contrast to the results documented in the review of CVD algorithms, such superior performance could not be achieved to the extent reported in some prior studies. Consequently, further investigation was conducted into studies that have utilized the Kaggle CVD Dataset. Since Bhatt et al. [25] achieved more promising results with their preprocessing approach, this work adapts it with the objective of achieving similar good performance. However, these measures did not yield the expected enhancements. Nevertheless, many previous studies [76], [99], [91], [102] have reported similar results to those achieved in this work for the Kaggle CVD dataset.

5.3.2 Mendeley Dataset

For the Mendeley dataset, KNN achieved the best results for the subset with a balanced gender distribution, shown in Table 14b. However, only precision was slightly higher for the male-dominant data. While DT produced superior results across all metrics when trained on the balanced subset, RF consistently yielded better results with the male-dominant data collection. As was the case with the DT, the MLP model demonstrated the most favorable outcomes when trained on the balanced dataset. However, the recall was found to be higher when applying the MLP to the male-skewed data. While RF revealed the best accuracy and F1-Score, KNN achieved the highest precision, and DT yielded the best recall. Ultimately, the models that were trained on the Mendeley dataset performed exceptionally well across all scenarios with different gender distributions.

Generally, the four algorithms that were trained on the Mendeley dataset exhibited consistent superior performance, aligning with the performance metrics observed in the review of prior work on CVD algorithms.

5.3.3 Composite HF Dataset

For this dataset under consideration, KNN, RF, and MLP demonstrated higher performance for the male-skewed subset, as depicted in Table 14c. For each of the models, the precision values are higher for the balanced or female-dominated subset. In general, the outcome of the DT is slightly diminished for this specific dataset, while reaching its best results when applied to the female-dominated data. Upon evaluation, KNN demonstrated superior performance in comparison to other models when applied to the male-dominated dataset. However, DT exhibited a higher level of precision.

The achievements of the Composite HF dataset indicate a slightly worse performance of the algorithms compared to the Mendeley dataset. The achievements of the Composite HF dataset indicate a slightly worse performance of the algorithms compared to the Mendeley dataset. However, KNN demonstrated the highest level of accuracy, precision, and F1-score for the Composite HF dataset. Meanwhile, RF showed the strongest recall.

6 Fairness Analysis and Bias Mitigation

The objective of this subsection is to provide a synopsis of the outcomes of the fairness analysis conducted. To detect and mitigate bias, three distinct toolkits were applied to three CVD prediction datasets to derive a comparative analysis. Therefore, the implementation of FairMLHealth and Fairlearn, along with their employed fairness metrics in bias detection, is explained. Subsequent to the identification of bias, the mitigation is performed with Fairlearn and AIF360, as FairMLHealth is not equipped with the mitigation strategies. The subsequent sections detail the application of each tool.

FairMLHealth: Given the central focus of this thesis on the detection of bias within a healthcare application, it is of particular interest to explore a healthcare-specific tool for bias analysis: FairMLHealth [15]. A virtual environment with Python 3.10.18 has been established due to package dependencies to ensure compatibility with FairMLHealth itself and its dependencies. Minor adjustments to the setup.cfg file were necessary to address conflicts related to the scypi. The use of the established kernel named `fairml (Python 3.10.18)` in Visual Studio Code enabled the application of FairMLHealth to assess fairness with regard to gender in CVD predictions.

The initial step in the fairness evaluation process involved using the function `measure summary()`, developed by FairMLHealth, on each trained classifier with gender as the protected attribute. Apart from typical performance indicators such as accuracy, precision, recall, and the F1-score, this function calculates a wide range of group fairness metrics. A variety of metrics are provided to assess disparities, including the Disparate Impact Ratio (DIR), the Statistical Parity Difference (SPD), the Equal Odds Difference (EOD) and Ratio, and the Positive Predictive Parity Difference (PPV) and Ratio. Table 25 presents an overview of the fairness metrics utilized for the analysis

and a comparison of bias detection tools, including FairMLHealth and Fairlearn.

Fairlearn: Fairlearn is a general fairness toolkit that has been used as a secondary detection instrument, in contrast to the healthcare-specific FairMLHealth. The Fairlearns MetricFrame class was employed for the computation of performance and fairness-related metrics. Performance-related metrics, including accuracy, precision, recall, selection rate, false positive rate, and true positive rate, were distinctly measured for each gender category. Employing the Equalized Odds Difference (EOD) and the Demographic Parity Difference (DPD) as fairness metrics, Fairlearn quantifies discrepancies in error distributions and prediction rates between gender groups. In addition to the group-wise illustration of performance metrics, the quantification of gaps was performed. Therefore, the area under the ROC curve (ROC-AUC) and the Brier Score were employed for each gender. Consequently, this facilitates a detailed, group-wise analysis employing multiple metrics, thereby offering a comprehensive perspective on fairness across gender.

In addition to its application in bias detection, Fairlearn was also utilized for bias mitigation. The available mitigation methods included the Exponentiated Gradient and Grid Search as in-processing methods, and the Threshold Optimizer as a post-processing method. The Exponentiated Gradient reduction approach was employed in conjunction with Demographic Parity (DP) and Equalized Odds (EO) as fairness constraints. The objective of this strategy was to align the models' predictions with these fairness constraints, thereby reducing disparities across sensitive groups.

AIF360: is utilized in conjunction with FairLearn to facilitate a comparative analysis of mitigation strategies. As a preprocessing technique, Reweighting was integrated to reduce potential disparities between female and male patients by adapting the distribution in the training data. The ADV approach, classified as an in-processing strategy, represents a fairness-aware approach designed to reduce an adversary's ability to infer sensitive attributes from predictive outcomes. Furthermore, postprocessing mitigation was performed using the EO offered by AIF360. This method modifies the predictive labels to achieve balanced outcomes by ensuring that each gender is equally affected by the model's errors.

Category	Metric	Fairlearn	FairMLHealth
Group-wise Metrics	Accuracy	✓	✓
	Precision	✓	✓
	Recall	✓	–
	F1-Score	✓	✓
	True Positive Rate (TPR)	✓	✓
	False Positive Rate (FPR)	✓	✓
	True Negative Rate (TNR)	✓	–
	False Negative Rate (FNR)	✓	–
	ROC–AUC	✓	✓
	PR–AUC	–	✓
	Brier Score	✓	–
Overall Metrics	Demographic Parity Difference (DPD)	✓	–
	Equalized Odds Difference (EOD)	✓	✓
	Equality of Opportunity Difference	✓	–
	Selection Rate	✓	–
	AUC Difference	–	✓
	Balanced Accuracy Difference	–	✓
	Positive Predictive Parity Difference	–	✓
	Statistical Parity Difference	–	✓

Table 15: Comparison of Fairness Metrics - Fairlearn vs. FairMLHealth

As illustrated in Table 15, the fairness metrics employed by Fairlearn and FairMLHealth, which are two fairness evaluation systems that were implemented in this thesis for bias detection, are compared. There are two main categories in the table. Overall Metrics offer a general view of fairness across the whole dataset. Group-wise Metrics evaluate performance differences among the gender groups. Both toolkits incorporate core performance metrics, such as accuracy, precision, ROC-AUC, true positive rate (TPR), false positive rate (FPR), and others. Nevertheless, the latter provides a greater variety of metrics, including TPR/FPR and PPV ratios and differences, balanced accuracy metrics, and others that measure inequalities across the distinct gender groups and overall. While Fairlearn provides the DI Ratio, EOD, Equality of Opportunity Difference, and DPD, the absence of several key ratio-based and mean-based metrics found in FairMLHealth is a notable distinction. Contrary to Fairlearn, which concentrates on a core selection of generally accepted fairness measures, FairMLHealth offers a richer array of fairness evaluation metrics, particularly with regard to capturing nuanced differences between the gender groups.

6.1 Bias Detection

The evaluation setup was established based on three distinct gender distribution scenarios for each dataset: a balanced (50F/50M), female-dominant (25M/75F), and male-dominant (75M/25F) composition, in order to facilitate clarity and ensure comparability. To ensure the validity of the results, consistent model settings were maintained while creating each scenario from the corresponding dataset. The impact of varying gender ratios on model performance and fairness outcomes was assessed using the same classifiers and fairness evaluation methods. With respect to the assessment of fairness, smaller gender disparities are indicated by lower DPD and EOD values. Accordingly, the most equitable outcomes across all models and datasets are indicated by bolded values, whereas the values reflecting the biggest gender disparities are underlined. To provide further context regarding the interpretation of fairness metrics, particularly absolute difference metrics such as DPD and EOD, the determination of an exact threshold is challenging due to the high dependency on the context. Moreover, a review of relevant literature and the AIA reveals a lack of specific criteria necessary to establish one. Consequently, a threshold of < 0.05 has been established for this study, following the approach of other studies in this field [55] [94]. Accordingly, fairness metrics that fall below 0.05 are regarded as a tolerable bias. The ensuing results are consequently presented in a sectioned format, corresponding to the previously defined scenarios.

6.1.1 Bias Detection | Scenario I: (75M/25F):

This subsection presents the results of the bias detection using Fairlearn and FairMLHealth for male-dominated datasets. Therefore, the objective is to analyze the presence of gender bias across the different algorithms and datasets used in this work and to establish a comparison of the consistency of the outcomes obtained by the two fairness tools.

Fairlearn: When applied to Scenario I, with male-dominant data collections, the Fairlearn toolset offers valuable insights into how gender imbalance can affect the fairness of predicting CVD. As shown in Table 16, the Kaggle CVD dataset has the lowest levels of gender inequality overall, with both EOD and DPD values close to zero. The MLP model, for instance, yields the most balanced outcomes, producing a DPD of *0.0021* and an EOD of *0.0119*. The results in these fairness metrics for the Kaggle CVD dataset indicate that, in this case, the male-dominant composition does not translate into biased predictions. Although still within a tolerable range, KNN exhibits

	Kaggle CVD		Mendeley		Composite HF	
Model	**DPD**	**EOD**	**DPD**	**EOD**	**DPD**	**EOD**
KNN	0.0358	0.0425	0.0802	0.1752	0.3796	0.0521
DT	0.0114	0.0155	0.0387	0.0406	0.2549	0.0812
RF	0.0101	0.0267	0.0438	0.0688	0.4081	0.0833
MLP	**0.0021**	**0.0119**	0.0604	0.0709	0.3396	0.1562

Table 16: DPD and EOD by Fairlearn across three male-dominated CVD datasets

considerably higher disparities, indicating disparities in fairness metrics below 0.05. Nevertheless, concerns over fairness become more prominent when deploying Fairlearn on the CVD, Mendeley, and Composite HF datasets. The DT exhibited the lowest signs of bias with a DPD metric of *0.0387* and EOD of *0.0406* for the Mendeley dataset. This indicates that the DT is more consistent in terms of achieving equitable prediction outcomes across the distinct gender groups. However, KNN reflected considerable gender bias, with a DPD metric of *0.0802* and an EOD of *0.1752*. The gender imbalance has the largest impact on the Composite HF dataset. Therefore, RF records the largest DPD of *0.4081*, whereas MLP shows a comparatively smaller value for DPD of *0.3396*, while suffering a greater EOD of *0.1562*.

Apart from DPD and EOD, as specific fairness metrics analyzed using Fairlearn performance metrics by gender provide critical insights into potential unequal treatments of a gender group in the prediction of CVD. An extensive overview of the performance metrics for each gender is provided for each dataset in Table 17, which illustrate the metrics for each of the analyzed datasets. A more comprehensive understanding of the strengths and gender-based disparities of the models is provided by the emphasized top results, which indicate which models produce the best value per metric across all datasets.

In the context of the Kaggle CVD dataset, both male and female performance metrics appear to be relatively balanced, with minor deviations observed in terms of recall, accuracy, and precision. According to minimal disparities in selection rate and FPR, a slight tendency towards misclassification can be observed for females.

Model	Gender	Performance Metrics						Fairness / Error Rate Metrics			
		Acc.	Prec.	Recall (TPR)	F1	ROC–AUC	Brier	Sel. Rate	FPR	FNR	TNR
KNN	0	0.6846	0.6869	0.6794	0.6831	0.7248	0.2228	0.4950	0.3102	0.3206	0.6898
	1	0.6757	0.6829	0.6369	0.6591	0.7184	0.2292	0.4592	0.2868	0.3631	0.7132
DT	0	0.7086	0.7056	0.7169	0.7112	0.7637	0.1962	0.5084	0.2996	0.2831	0.7004
	1	0.7075	0.6922	0.7308	0.7110	0.7514	0.2005	0.5198	0.3151	0.2692	0.6849
RF	0	0.7085	0.7198	0.6835	0.7012	0.7638	0.1979	0.4751	0.2664	0.3165	0.7336
	1	0.6893	0.6952	0.6568	0.6754	0.7449	0.2079	0.4651	0.2792	0.3432	0.7208
MLP	0	0.7165	0.7306	0.6868	0.7070	0.7706	0.1946	0.4704	0.2537	0.3132	0.7463
	1	0.7106	0.7147	0.6860	0.7000	0.7622	0.1985	0.4725	0.2656	0.3140	0.7344

(a) Kaggle CVD

Model	Gender	Performance Metrics						Fairness / Error Rate Metrics			
		Acc.	Prec.	Recall (TPR)	F1	ROC–AUC	Brier	Sel. Rate	FPR	FNR	TNR
KNN	0	0.8261	0.9091	0.7692	0.8333	0.9365	0.1036	0.4783	0.1000	0.2308	0.9000
	1	**0.9610**	0.9884	0.9444	**0.9659**	0.9922	**0.0380**	0.5584	**0.0156**	0.0556	**0.9844**
DT	0	0.9130	0.9231	0.9231	0.9231	0.9058	0.0867	0.5652	0.1000	0.0769	0.9000
	1	0.9026	0.9032	0.9333	0.9180	0.9352	0.0830	0.6039	0.1406	0.0667	0.8594
RF	0	0.9565	0.9286	**1.0000**	0.9630	**0.9962**	0.0407	0.6087	0.1000	**0.0000**	0.9000
	1	0.9545	**0.9770**	0.9444	0.9605	0.9858	0.0400	0.5649	0.0312	0.0556	0.9688
MLP	0	0.8913	0.9200	0.8846	0.9020	0.9260	0.1087	0.5435	0.1000	0.1154	0.9000
	1	0.9286	0.9247	0.9556	0.9399	0.9819	0.0675	0.6039	0.1094	0.0444	0.8906

(b) Mendeley

Model	Gender	Performance Metrics						Fairness / Error Rate Metrics			
		Acc.	Prec.	Recall (TPR)	F1	ROC–AUC	Brier	Sel. Rate	FPR	FNR	TNR
KNN	0	0.8684	0.5556	0.8333	0.6667	0.9583	0.0830	0.2368	0.1250	0.1667	0.8750
	1	0.8904	0.9444	0.8854	0.9140	0.9086	0.1041	0.6164	0.1000	0.1146	0.9000
DT	0	0.7368	0.3571	0.8333	0.5000	0.7630	0.2383	0.3684	0.2812	0.1667	0.7188
	1	0.8288	0.8901	0.8438	0.8663	0.8445	0.1553	0.6233	0.2000	0.1562	0.8000
RF	0	0.8947	0.6000	**1.0000**	0.7500	0.9818	0.0984	0.2632	0.1250	**0.0000**	0.8750
	1	0.8767	0.8980	0.9167	0.9072	0.8941	0.1031	0.6712	0.2000	0.0833	0.8000
MLP	0	0.8947	0.6000	**1.0000**	0.7500	0.9688	0.0793	0.2632	0.1250	**0.0000**	0.8750
	1	0.8493	0.9205	0.8438	0.8804	0.9042	0.1180	0.6027	0.1400	0.1562	0.8600

(c) Composite HF

Table 17: Gender-stratified performance in the male-dominated scenario using Fairlearn

However, a notable disparity emerges in the Mendeley dataset for KNN and MLP. For these two models, females consistently achieve lower accuracy, recall, and F1-score compared to males. Consequently, predictive performance is shown to be biased against women.

In accordance with the performance figures of the Composite HF dataset, most models have a tendency to underperform for female patients. Women are significantly less likely than men to be categorized as positive instances, as evidenced by their significantly lower selection rates across all models. There is also consistently lower precision for female patients, indicating that predictions are less accurate. Nevertheless, women tend to have better recall than men, indicating that true cases of CVD are rarely missed. Taken together, the Mendeley dataset demonstrated the most optimal overall performance across all datasets, as indicated by the highlighted best results. In this particular instance, the majority of the best-performing metrics were identified for male patients, particularly in terms of accuracy, F1-score, and ROC-AUC, which indicate both more robust prediction and more equitable error rates. Furthermore, the selection rate does not exhibit any significant disparities between the two genders.

According to the fairness analysis conducted across the three male-dominated CVD datasets, inequalities detected by Fairlearn using DPD and EOD are consistently reflected in the stratified performance metrics. DPD and EOD values, as well as the differences in performance, remain low in the Kaggle CVD dataset. There are only slight differences in the precision, recall, and selection rates of males and females, implying fair results. Larger gender gaps are evident through lower recall and selection rates for females in the Mendeley dataset, particularly for KNN. Despite this, DT shows nearly comparable group performance, consistent with its low DPD and EOD values. The Composite HF dataset has the most pronounced gender bias, with high DPD and EOD values corresponding to extremely low precision and selection rates for females, even though their recall is typically greater. This suggests that females are not as often and accurately categorized as positive cases. Selection rate, precision, and recall offer convincing justifications for the discrepancies identified by DPD and EOD, and parity-based fairness measures generally correlate well with group-level performance differences.

FairMLHealth: The results in Table 18 were derived from the application of the designated healthcare application tool, FairMLHealth. Thus, the variety of fairness metrics offered by the tools and their consistency in detecting bias are compared.

Model	AUC Diff	BA Diff	EOD	PPV Diff	SPD
KNN	0.0064	0.0095	0.0425	**0.0040**	0.0358
DT	0.0123	**0.0008**	-0.0155	0.0134	-0.0114
RF	0.0188	0.0198	0.0267	0.0246	0.0101
MLP	**0.0084**	0.0063	**-0.0119**	0.0159	**-0.0021**

(a) Kaggle CVD

Model	AUC Diff	BA Diff	EOD	PPV Diff	SPD
KNN	-0.0550	-0.0174	0.0431	-0.0793	0.0860
DT	0.0206	0.0100	-0.0287	-0.0199	-0.0387
RF	-0.0104	-0.0167	-0.0556	0.0484	0.0438
MLP	-0.0560	-0.0308	-0.0709	-0.0047	-0.0604

(b) Mendeley

Model	AUC Diff	BA Diff	EOD	PPV Diff	SPD
KNN	0.0497	-0.0385	-0.0521	-0.3889	-0.3796
DT	0.0579	0.0579	-0.0950	-0.3208	-0.3727
RF	0.0877	0.0792	0.0833	-0.2980	-0.4081
MLP	0.0646	0.0856	0.1562	-0.3205	-0.3396

(c) Composite HF

Table 18: FairMLHealth group fairness metrics differences across three CVD datasets in the male-dominated scenario

All fairness metrics evaluated based on the Kaggle CVD dataset indicate the absence of bias across all models utilized. On the contrary, the Mendeley dataset shows slight signs of bias in the case of the KNN and MLP algorithms. Both models demonstrate differences in AUC, indicating lower predictive performance for females. Furthermore, a gap in the error distribution is observed for the MLP and RF with EOD values of *-0.709* and *-0.556*, respectively. The metrics demonstrate that these models are more accurate in predicting CVD for males, producing more errors for females. The Composite HF Dataset has the most considerable differences, as almost all models show clear signs of bias. An SPD of *-0.3796* and a PPV Difference of *-0.3889* are both quite substantial for KNN, indicating persistent discrepancies in positive prediction accuracy and selection rates. A similar underrepresentation of the female patients, the disadvantaged group, in favorable outcomes

	Kaggle CVD		CVD Mendeley		Composite HF	
Model	**DPD**	**EOD**	**DPD**	**EOD**	**DPD**	**EOD**
KNN	0.0458	0.0564	0.0220	0.1062	0.4301	0.1146
DT	0.0038	**0.0078**	0.0017	0.0594	0.3439	0.0938
RF	0.0096	0.0228	0.0285	0.1000	0.4243	0.1288
MLP	0.0289	0.0421	**0.0003**	0.0231	0.4712	0.1888

Table 19: DPD and EOD by Fairlearn across three female-dominated CVD datasets

is indicated by the DT model's PPV Difference of *-0.3208* and SPD of *-0.3727*, and the RF model's PPV Difference of *-0.2980* and SPD of *-0.4081*. Furthermore, the MLP model's true positive rates are significantly biased in favor of the unprivileged group, as seen in the largest EOD difference at *0.1562*. Thus, the MLP shows bias against female patients in the selection rate, with an SPD of *-0.3396*.

6.1.2 Bias Detection | Scenario II: (75F/25M)

The following section presents the outcomes of the bias detection analysis obtained from Fairlearn and FairMLHealth for Scenario II, in which the datasets predominantly consist of female patients.

Fairlearn: In a manner analogous to the male-dominant scenario, the female-dominant scenario reveals that the levels of DPD and EOD persistently remain at low levels for the Kaggle CVD dataset across all models implemented. As demonstrated in Table 19, the DPD metrics approximate zero, reflecting balanced selection rates for both genders in the Mendeley Dataset. The EOD metric reveals imbalances in the KNN and RF models with values of *0.1062* and *0.10*, respectively. However, it is very close to zero for the DT and MLP models, indicating an equivalent error distribution across female and male subjects. Both fairness values show significant signs of bias across all models due to high DPD and EOD values for the Composite HF Dataset. Across all models, DPD shows substantial differences in selection rates among gender groups, ranging from *0.3439* to *0.4712*. Additionally, EOD reveals a bias in the form of significant differences in the distribution of errors, especially for the MLP, with a value of *0.1888*.

Model	Gender	Performance Metrics						Fairness / Error Rate Metrics			
		Acc.	Prec.	Recall (TPR)	F1	ROC–AUC	Brier	Sel. Rate	FPR	FNR	TNR
KNN	0	0.7109	0.7095	0.7153	0.7124	0.7661	0.1969	0.5045	0.2934	0.2847	0.7066
	1	0.6977	0.7072	0.6588	0.6821	0.7515	0.2030	0.4587	0.2645	0.3412	0.7355
DT	0	0.7111	0.7128	0.7079	0.7103	0.7709	0.1940	0.4970	0.2858	0.2921	0.7142
	1	0.7116	0.7036	0.7157	0.7096	0.7556	0.2001	0.5008	0.2924	0.2843	0.7076
RF	0	0.7108	0.7297	0.6705	0.6988	0.7697	0.1950	0.4598	0.2488	0.3295	0.7512
	1	0.7013	0.7062	0.6734	0.6895	0.7566	0.2014	0.4694	0.2716	0.3266	0.7284
MLP	0	0.7142	0.7251	0.6908	0.7075	0.7725	0.1936	0.4768	0.2624	0.3092	0.7376
	1	0.7042	0.6943	0.7131	0.7036	0.7592	0.2002	0.5056	0.3045	0.2869	0.6955

(a) Kaggle CVD

Model	Gender	Performance Metrics						Fairness / Error Rate Metrics			
		Acc.	Prec.	Recall (TPR)	F1	ROC–AUC	Brier	Sel. Rate	FPR	FNR	TNR
KNN	0	0.8478	0.8519	0.8846	0.8679	0.9279	0.0991	0.5870	0.2000	0.1154	0.8000
	1	0.9026	0.9310	0.9000	0.9153	0.9354	0.0914	0.5649	0.0938	0.1000	0.9062
DT	0	0.8696	0.8571	0.9231	0.8889	0.8365	0.1285	0.6087	0.2000	0.0769	0.8000
	1	0.9091	0.9043	0.9444	0.9239	0.9559	0.0853	0.6104	0.1406	0.0556	0.8594
RF	0	**0.9783**	**0.9630**	**1.0000**	**0.9811**	**1.0000**	**0.0385**	0.5870	0.0500	**0.0000**	0.9500
	1	0.9091	0.9419	0.9000	0.9205	0.9799	0.0632	0.5584	0.0781	0.1000	0.9219
MLP	0	0.9130	0.9231	0.9231	0.9231	0.9577	0.0745	0.5652	0.1000	0.0769	0.9000
	1	0.9026	0.9310	0.9000	0.9153	0.9700	0.0906	0.5649	0.0938	0.1000	0.9062

(b) Mendeley

Model	Gender	Performance Metrics						Fairness / Error Rate Metrics			
		Acc.	Prec.	Recall (TPR)	F1	ROC–AUC	Brier	Sel. Rate	FPR	FNR	TNR
KNN	0	0.9211	0.8000	0.6667	0.7273	0.9531	0.0737	0.1316	**0.0312**	0.3333	**0.9688**
	1	0.8082	0.9146	0.7812	0.8427	0.8899	0.1363	0.5616	0.1400	0.2188	0.8600
DT	0	0.7368	0.3571	0.8333	0.5000	0.7969	0.1961	0.3684	0.2812	0.1667	0.7188
	1	0.8493	0.8558	0.9271	0.8900	0.8539	0.1392	0.7123	0.3000	0.0729	0.7000
RF	0	0.9474	0.8333	0.8333	0.8333	0.9531	0.0694	0.1579	**0.0312**	0.1667	**0.9688**
	1	0.8151	0.9059	0.8021	0.8508	0.9038	0.1353	0.5822	0.1600	0.1979	0.8400
MLP	0	0.9211	0.8000	0.6667	0.7273	0.8646	0.0615	0.1316	**0.0312**	0.3333	**0.9688**
	1	0.7945	0.8750	0.8021	0.8370	0.8710	0.1675	0.6027	0.2200	0.1979	0.7800

(c) Composite HF

Table 20: Gender-stratified performance in the female-dominated scenario using Fairlearn

The performance of male and female patients in the Kaggle CVD dataset, shown in Table 20 is essentially the same, with only minor deviations in F1-score, recall, accuracy, and precision. Gender

differences in selection rates and recall values are minimal as well. Taken together, these figures suggest that subgroups are treated fairly equally in the Kaggle CVD dataset.

Despite women comprising the majority of the training data, performance findings from the Mendeley dataset indicate that male patients tend to have better outcomes across most models. A pattern can be observed in both KNN and DT, where all performance metrics are lower for female patients. The differences in performance metrics are modest across gender. The MLP model shows smaller disparities, with females achieving slightly higher accuracy, recall, and F1-score, while males record higher precision. There are also no significant gender differences in error rates.

Upon examination of the Composite HF Dataset, it is evident that there are substantial disparities. The significant differences in selection rates across all models highlight fairness concerns, with males consistently receiving more favorable predictions. In terms of accuracy and ROC-AUC, females perform better in KNN and MLP. However, they have poorer recall and F1, meaning more CVD cases are overlooked. In the context of DT, the recall, precision, and F1-score are much lower for female patients. In general, the gaps in the metrics are substantial, and females suffer from extremely low precision. The RF model performs better for females in terms of accuracy, ROC-AUC, calibration, and minimal false positives; however, it only predicts positives for a small percentage of them. Although the model has a higher F1-score for men due to good precision and recall, its accuracy and calibration are poorer. This leads to many patients being flagged unnecessarily because the model predicts positives much more frequently.

In the female-dominant scenario, the top overall performance across all models was obtained for female patients, as evidenced in the output of the Mendeley and the Composite HF dataset. In the context of the Mendeley dataset, the RF model demonstrated particularly notable results for females, achieving the highest output across all performance metrics. Furthermore, the output from the Composite HF dataset indicate that females tend to achieve better performance than males, with the lowest FPR and highest TNR values. Taking the top results from the female-dominant scenario, models tend to favor female patients in terms of both predictive accuracy and fairness.

The fairness metrics principally underpin the performance-based findings. In summary, the Kaggle CVD dataset exhibits minor gender-related differences in performance, which are consistent with the low levels of DPD and EOD. The differences in fairness are slightly greater in the Mendeley dataset. DT and MLP are closer to parity, while KNN and RF have higher EOD values, indicating recall gaps that favor males. Men are consistently more likely to receive positive predictions, as evidenced by the Composite HF dataset, which displays significantly higher DPD values across all models. Furthermore, EOD values are notably high for MLP and RF, which is confirmed through substantial gender disparities in true positive rates.

FairMLHealth: Following the bias evaluation conducted using Fairlearn, the FairMLHealth fairness metrics results are presented in Table 21.

Aligned with the Fairlearn results, the FairMLHealth tool confirms the absence of bias in the Kaggle CVD dataset, even in a scenario with female dominance in the training data. The only slight disparity appears in the application of KNN, where EOD reaches a value of *0.0564*. This represents a slight disparity in the detection of true cases among males, whereas the MLP shows a slight disadvantage for females in this context. The other metrics remain low, which indicates overall fair results. However, inequalities are more prevalent in the Mendeley dataset. Interestingly, these inequalities reveal themselves in a more mixed way, demonstrating bias against both genders. KNN shows different disadvantages for both genders. Female patients are advantaged by an EOD of *0.1062*, which suggests that they are more likely to receive accurate positive predictions. At the same time, the model appears to perform worse for females in terms of accurate prediction, as indicated by a BA difference of *-0.0608* and a PPV difference of *-0.0792*. Likewise, the DT shows a substantial drop in predictive performance for women with a negative AUC difference of *-0.1194*, despite the modest EOD advantage of *0.0594*. While KNN and DT demonstrate mixed inequalities, RF consistently shows a moderate preference for females. The MLP model produced the most balanced results, with only minor variations. Severe bias across all models is indicated by the highly elevated fairness metrics displayed for the Composite HF Dataset. As indicated by the consistent demonstration of lower positive prediction frequency and reduced prediction accuracy, as measured by negative statistical parity and positive predictive value,

Model	AUC Diff	BA Diff	EOD	PPV Diff	SPD
KNN	0.0146	0.0138	0.0564	**0.0023**	0.0458
DT	0.0153	**-0.0006**	**-0.0078**	0.0092	-0.0038
RF	0.0131	0.0099	-0.0228	0.0235	-0.0096
MLP	0.0133	0.0099	-0.0421	0.0308	-0.0289

(a) Kaggle CVD

Model	AUC Diff	BA Diff	EOD	PPV Diff	SPD
KNN	-0.0075	-0.0608	0.1062	-0.0792	0.0220
DT	-0.1194	-0.0404	0.0594	-0.0471	-0.0017
RF	0.0201	0.0641	0.1000	0.0211	0.0285
MLP	-0.0123	0.0084	0.0231	-0.0080	**0.0003**

(b) Mendeley

Model	AUC Diff	BA Diff	EOD	PPV Diff	SPD
KNN	0.0632	-0.0029	0.1146	-0.1146	-0.4301
DT	-0.0570	-0.0375	-0.0938	-0.4986	-0.3439
RF	0.0494	0.0800	-0.1288	-0.0725	-0.4243
MLP	**-0.0065**	0.0267	-0.1888	-0.0750	-0.4712

(c) Composite HF

Table 21: FairMLHealth group fairness metrics differences across three CVD datasets in the female-dominated scenario

across all models, women are consistently disadvantaged. Furthermore, women frequently face a disadvantage in terms of detection, as reflected by negative EOD values ranging from *-0.090* to *-0.19*. This indicates that the models exhibit lower sensitivity for females. It is evident that women are not treated fairly at the selection and prediction levels. This disparity is observed even when they reach higher BA difference and AUC Difference, as in the case of RF.

6.1.3 Bias Detection | Scenario III: (50F/50M)

The bias detection results for the gender-balanced scenario conducted with Fairlearn and FairMLHealth are outlined in the subsequent section.

Fairlearn: Consistent with the previous scenarios, DPD as well as EOD remain very close to zero for the Kaggle CVD dataset, as presented in

	Kaggle CVD		CVD Mendeley		Composite HF	
Model	**DPD**	**EOD**	**DPD**	**EOD**	**DPD**	**EOD**
KNN	**0.0071**	**0.0060**	0.0150	0.0688	0.4380	0.1875
DT	0.0254	0.0357	0.0113	0.0444	0.3306	0.1042
RF	0.0155	0.0215	0.0373	0.0667	0.3464	0.0237
MLP	0.0199	0.0267	0.0997	0.1368	0.3980	0.1562

Table 22: DPD and EOD by Fairlearn across three balanced CVD datasets

Table 22. Concerning the Mendeley dataset, the values imply parity for the majority of models. Only the MLP model displays moderate bias in both fairness metrics. Although in the case of gender balance, the Composite HF dataset yields elevated DPD and EOD values, implying gender bias. There are consistently high DPD values, ranging from *0.3306* to *0.4380*, particularly for KNN and MLP. Although RF achieves relatively balanced error distributions across genders through a low DPD of *0.0237*, it still produces large gaps with a high DPD of *0.3464*. Finally, KNN, DT, and MLP also display high imbalance in error rates through elevated EOD.

Table 23 shows the performance of each model applied to gender-balanced CVD datasets, stratified by gender. Equal performance is shown by men and women in the Kaggle CVD dataset, with similar accuracy, F1 scores, and selection rates. However, most models show that men have a slight advantage in recall, alongside higher false positive rates, while women have greater precision. DT exhibits the greatest discrepancies, while KNN appears to be the most gender-balanced model. When evaluated using the Mendeley dataset, the KNN, DT, and RF models exhibited a relatively balanced performance across both genders. The assessment of the MLP model reveals persistent inequalities in performance metrics. The most significant disparity emerges in the recall, where male patients are given priority, as evidenced by the considerably lower FNR for males and the lower TPR for females.

In accordance with the preceding scenarios, the Composite HF Dataset demonstrates inequalities across all models applied. The most significant disparities are evident in the selection rate, precision, recall,

Model	Gender	Performance Metrics						Fairness / Error Rate Metrics			
		Acc.	Prec.	Recall (TPR)	F1	ROC–AUC	Brier	Sel. Rate	FPR	FNR	TNR
KNN	0	0.6999	0.6974	0.7074	0.7023	0.7639	0.1978	0.5076	0.3075	0.2926	0.6925
	1	0.7021	0.6942	0.7058	0.6999	0.7569	0.2006	0.5005	0.3015	0.2942	0.6985
DT	0	0.7157	0.7289	0.6876	0.7076	0.7410	0.1983	0.4720	0.2561	0.3124	0.7439
	1	0.7088	0.7021	0.7094	0.7058	0.7364	0.2017	0.4974	0.2919	0.2906	0.7081
RF	0	0.7100	0.7285	0.6702	0.6981	0.7715	0.1943	0.4603	0.2501	0.3298	0.7499
	1	0.7078	0.7102	0.6865	0.6981	0.7578	0.2006	0.4759	0.2716	0.3135	0.7284
MLP	0	0.7172	0.7303	0.6895	0.7093	0.7755	0.1922	0.4724	0.2550	0.3105	0.7450
	1	0.7139	0.7094	0.7094	0.7094	0.7682	0.1957	0.4923	0.2817	0.2906	0.7183

(a) Kaggle CVD

Model	Gender	Performance Metrics						Fairness / Error Rate Metrics			
		Acc.	Prec.	Recall (TPR)	F1	ROC–AUC	Brier	Sel. Rate	FPR	FNR	TNR
KNN	0	0.8913	0.9200	0.8846	0.9020	0.8923	0.1087	0.5435	0.1000	0.1154	0.9000
	1	0.9481	**0.9767**	0.9333	0.9545	0.9510	0.0519	0.5584	**0.0312**	0.0667	**0.9688**
DT	0	**0.9565**	0.9286	**1.0000**	**0.9630**	0.9712	**0.0421**	0.6087	0.1000	**0.0000**	0.9000
	1	0.9351	0.9348	0.9556	0.9451	0.9230	0.0673	0.5974	0.0938	0.0444	0.9062
RF	0	**0.9565**	0.9286	**1.0000**	**0.9630**	**0.9856**	0.0552	0.6087	0.1000	**0.0000**	0.9000
	1	0.9351	0.9545	0.9333	0.9438	0.9852	0.0502	0.5714	0.0625	0.0667	0.9375
MLP	0	0.8696	0.9545	0.8077	0.8750	0.9731	0.0944	0.4783	0.0500	0.1923	0.9500
	1	0.9416	0.9551	0.9444	0.9497	0.9806	0.0579	0.5779	0.0625	0.0556	0.9375

(b) Mendeley

Model	Gender	Performance Metrics						Fairness / Error Rate Metrics			
		Acc.	Prec.	Recall (TPR)	F1	ROC–AUC	Brier	Sel. Rate	FPR	FNR	TNR
KNN	0	0.8947	0.6667	0.6667	0.6667	0.9531	0.0706	0.1579	0.0625	0.3333	0.9375
	1	0.8699	0.9425	0.8542	0.8962	0.9181	0.1133	0.5959	0.1000	0.1458	0.9000
DT	0	0.8421	0.5000	0.6667	0.5714	0.8151	0.1431	0.2105	0.1250	0.3333	0.8750
	1	0.8151	0.9367	0.7708	0.8457	0.8786	0.1433	0.5411	0.1000	0.2292	0.9000
RF	0	0.8421	0.5000	0.8333	0.6250	0.9661	0.0849	0.2632	0.1562	0.1667	0.8438
	1	0.8288	0.8989	0.8333	0.8649	0.8999	0.1247	0.6096	0.1800	0.1667	0.8200
MLP	0	0.8684	0.5714	0.6667	0.6154	0.9167	0.0868	0.1842	0.0938	0.3333	0.9062
	1	0.8425	0.9294	0.8229	0.8729	0.8992	0.1263	0.5822	0.1200	0.1771	0.8800

(c) Composite HF

Table 23: Gender-stratified performance in the gender-balanced scenario using Fairlearn

and F1-score. These are also mirrored in the lower TPR and higher FNR for female patients, implying the algorithms perform worse in diagnosing CVD for females.

Considering the outcomes of the gender-balanced scenario, female patients achieved the best overall results in most of the metrics across all models, as reflected in the Mendeley dataset. The DT and RF models, in particular, achieved outstanding results for females, demonstrating the highest accuracy, recall, F1-score, and ROC-AUC values. However, male patients demonstrated slightly higher precision, slightly lower FPR and higher TNR.

In fact, the performance outcomes per gender reflect the results of the fairness metrics. Due to similar accuracy, F1, and selection rates for each gender in the Kaggle CVD dataset, both DPD and EOD are nearly zero. In contrast, modest differences are seen in the Mendeley dataset, particularly for the MLP model, indicated through a DPD value of *0.0997* and an EOD of *0.1368*. The latter reflects the observed male-favoring FNR and recall gaps. Having high DPD ranging from *0.3306* to *0.4380* and elevated EOD values, the Composite HF dataset shows the greatest discrepancies among all models. This is consistent with significant gender differences in recall, F1, and selection rates.

Model	AUC Diff	BA Diff	EOD	PPV Diff	SPD
KNN	0.0070	**-0.0022**	**0.0060**	0.0031	**0.0071**
DT	0.0047	0.0069	-0.0357	0.0268	-0.0254
RF	0.0137	0.0026	-0.0215	0.0183	-0.0155
MLP	0.0073	0.0034	-0.0267	0.0209	-0.0199

(a) Kaggle CVD

Model	AUC Diff	BA Diff	EOD	PPV Diff	SPD
KNN	-0.0587	-0.0587	0.0688	-0.0567	-0.0150
DT	0.0482	0.0191	0.0444	-0.0062	0.0113
RF	**0.0003**	0.0146	0.0667	-0.0260	0.0373
MLP	-0.0075	-0.0621	-0.1368	**-0.0005**	-0.0997

(b) Mendeley

Model	AUC Diff	BA Diff	EOD	PPV Diff	SPD
KNN	0.0350	-0.0750	-0.1875	-0.2759	-0.4380
DT	-0.0635	-0.0646	-0.1042	-0.4367	-0.3306
RF	0.0663	0.0119	-0.0237	-0.3989	-0.3464
MLP	0.0175	-0.0650	-0.1562	-0.3580	-0.3980

(c) Composite HF

Table 24: FairMLHealth group fairness metrics differences across three CVD datasets in the gender-balanced scenario

FairMLHealth: The results of the FairMLHealth fairness metrics in the gender balanced scenario are delineated in Table 24 across all three datasets. Once more, the Kaggle CVD dataset demonstrates a very fair and balanced performance among both genders, throughout all fairness metrics. All fairness metrics yield values very close to zero across each model, signifying equal treatment of female and male patients. Regarding the Mendeley dataset, generally, it can be observed that the fairness metrics are further away from zero, demonstrating slight to moderate inequalities in the majority of models used. This implies that females are at a disadvantage in prediction, which manifests as a lower TPR or a higher FPR. For the latter, it's an indicator of moderate underselection of females. Consistent discrimination against females is illustrated in the evaluation of the Composite HF dataset. A substantial under-selection of women is evident in all models, with SPD

values ranging from *-0.3306* to *-0.4380*, and the precision of detecting true cases is worse, as indicated by PPV Difference between *-0.2759* and *-0.4367*. Moreover, the EOD yielded negative results, with values of *-0.1875* for KNN and *-0.1562* for MLP, suggesting that women exhibit diminished detection performance.

A direct comparison of the EOD and DPD fairness metrics from FairMLHealth and Fairlearn reveals equivalent results across female-dominated and gender-balanced scenarios. However, within the male-dominant setting, inconsistencies emerged in both the Mendeley dataset and the Composite HF dataset. For the Mendeley data, disparities are negligible in most cases, with a notable difference in the performance of the KNN model for the EOD metric: FairMLHealth reports *0.0431*, whereas Fairlearn reports *0.1752*. A close examination of the Composite HF dataset reveals that discrepancies manifest exclusively within the DT model. A slight variation is observed in the EOD metric, and a significant difference is identified in the DPD metric, with FairMLHealth reporting *0.3727* and Fairlearn *0.2549*. This comparison indicates that the model selection and the distribution of the gender attribute exert a more significant influence on the fairness analysis than the choice of fairness tool. Therefore, assessing bias across several fairness tools strengthens the reliability of the outcomes.

In light of the findings from the bias evaluation presented in Table 25, it is evident that the Kaggle CVD dataset is not subject to gender bias, showing only negligible differences in performance between male and female patients. These equitable results are consistent across all three gender distribution scenarios, including male-skewed, female-skewed, and balanced samples. Conversely, the Mendeley dataset exhibits slight to moderate bias in scenarios dominated by either males or females. While the utilization of a balanced sample does enhance the fairness of the process, disparities nevertheless persist. The Composite HF dataset reveals a significant and consistent level of bias across all scenarios and models. The findings indicate that gender bias becomes visible through the comparison of fairness metrics across different gender distributions and models, particularly in the Mendeley and Composite HF datasets.

	Kaggle CVD				Mendeley				Composite HF			
	FairMLHealth		Fairlearn		FairMLHealth		Fairlearn		FairMLHealth		Fairlearn	
Model	**DPD**	**EOD**	**DPD**	**EOD**	**DPD**	**EOD**	**DPD**	**EOD**	**DPD**	**EOD**	**DPD**	**EOD**
KNN	0.0358	0.0425	0.0358	0.0425	0.0860	0.0431	0.0802	0.1752	0.3796	0.0521	0.3796	0.0521
DT	0.0114	0.0155	0.0114	0.0155	0.0387	0.0287	0.0387	0.0406	0.3727	0.0950	0.2549	**0.0812**
RF	0.0101	0.0267	0.0101	0.0267	0.0438	0.0556	0.0438	0.0688	0.4081	0.0833	0.4081	0.0833
MLP	0.0021	0.0119	0.0021	0.0119	0.0604	0.0709	0.0604	0.0709	0.3396	0.1562	0.3396	0.1562

(a) Scenario I (75M/25F)

	Kaggle CVD				Mendeley				Composite HF			
	FairMLHealth		Fairlearn		FairMLHealth		Fairlearn		FairMLHealth		Fairlearn	
Model	**DPD**	**EOD**	**DPD**	**EOD**	**DPD**	**EOD**	**DPD**	**EOD**	**DPD**	**EOD**	**DPD**	**EOD**
KNN	0.0458	0.0564	0.0458	0.0564	0.0220	0.1062	0.0220	0.1062	0.4301	0.1146	0.4301	0.1146
DT	0.0038	0.0078	0.0038	0.0078	0.0017	0.0594	0.0017	0.0594	0.3439	0.0938	0.3439	0.0938
RF	0.0096	0.0228	0.0096	0.0228	0.0285	0.1000	0.0285	0.1000	0.4243	0.1288	0.4243	0.1288
MLP	0.0289	0.0421	0.0289	0.0421	0.0003	0.0231	0.0003	0.0231	0.4712	0.1888	0.4712	0.1888

(b) Scenario II (75F/25M)

	Kaggle CVD				Mendeley				Composite HF			
	FairMLHealth		Fairlearn		FairMLHealth		Fairlearn		FairMLHealth		Fairlearn	
Model	**DPD**	**EOD**	**DPD**	**EOD**	**DPD**	**EOD**	**DPD**	**EOD**	**DPD**	**EOD**	**DPD**	**EOD**
KNN	0.0071	0.0060	0.0071	0.0060	0.0150	0.0688	0.0150	0.0688	0.4380	0.1875	0.4380	0.1875
DT	0.0254	0.0357	0.0254	0.0357	0.0113	0.0444	0.0113	0.0444	0.3306	0.1042	0.3306	0.1042
RF	0.0155	0.0215	0.0155	0.0215	0.0373	0.0667	0.0373	0.0667	0.3464	0.0237	0.3464	0.0237
MLP	0.0199	0.0267	0.0199	0.0267	0.0997	0.1368	0.0997	0.1368	0.3980	0.1562	0.3980	0.1562

(c) Scenario III (50F/50M): Gender balanced CVD datasets

Table 25: Fairness metric comparison across Fairlearn and FairMLHealth

6.2 Bias Mitigation

Following the implementation of bias detection, the subsequent mitigation process is necessary to address the identified disparities. As previously stated, FairMLHealth does not provide any form of mitigation techniques. Consequently, bias mitigation approaches from Fairlearn and AIF360 are employed to compare the effectiveness of the tools. The following sections present the results of the bias methods employed for each fairness tool. Those results provide insight into the impact of specific bias mitigation strategies and the potential trade-offs between accuracy and fairness.

6.2.1 Bias Mitigation | Scenario I (75M/25F):

The Mendeley and Composite HF datasets both demonstrate moderate to significant bias across all scenarios. For this reason, bias mitigation measures will be implemented on these two datasets. Consequently, the outcomes of distinct bias mitigation techniques applied to both datasets in Scenario I are presented in the subsequent paragraphs.

AIF360: The gender bias mitigation approaches applied to male-skewed CVD datasets leveraging AIF360 yielded the following results, as depicted in Table 26. In general, the baseline models show good predictive accuracy for the Mendeley dataset, with the RF baseline yielding the highest accuracy value of *95.5%*. Nonetheless, considerable group inequalities are associated with this superior performance. Using reweighting as a pre-processing technique is not always effective in increasing fairness. In fact, this strategy worsens equity in many cases without generating accuracy improvements. Overall, post-processing aimed at reducing bias yields the same metrics as the initial baseline. In the case of DT, using post-processing in Equalized Odds significantly lowers accuracy and worsens both fairness indicators. Meanwhile, the in-processing method, ADV, achieves the best balance, especially when tuned. Consequently, ADV produces an accuracy of *91.5%*, almost completely removing differences across genders, with a DPD of *0.0003* and an EOD of *0.0376*. In the context of the male-skewed data composition from the Mendeley dataset, the implementation of the in-processing method appears to be a more effective approach in enhancing fairness without significantly compromising accuracy.

Although the baselines in the Composite HF dataset are very accurate, significant disparities in fairness persist, particularly with regard to DPD. Reweighting does not substantially reduce bias without compromising accuracy or increasing EOD. The most equitable performance is achieved through post-processing with Equalized Odds applied to the DT model. Furthermore, its accuracy is maintained in comparison to the established baseline, with an accuracy of *80.43%*. Additionally, it fosters fair performance, as evidenced by the attainment of a DPD of *0.1896* and an EOD of *0.0312*. Furthermore, the ADV method also demonstrates its ability to enhance fairness, as it reduces EOD and increases accuracy, but at the cost of higher DPD. Therefore, the DT model with post-processing using Equalized Odds still achieves the fairest result.

	Mendeley			Composite HF		
Model / Variant	**Accuracy**	**DPD**	**EOD**	**Accuracy**	**DPD**	**EOD**
KNN – Baseline	0.9300	0.0082	0.1752	0.8859	0.3796	0.0521
KNN – Pre: Reweigh	0.9100	0.0542	0.1308	0.8641	0.3659	0.1562
KNN – Post: EqOdds	0.9300	0.0082	0.1752	0.8859	0.3796	0.0521
DT – Baseline	0.9050	0.0387	**0.0103**	0.8098	0.2549	0.0104
DT – Pre: Reweigh	0.9300	0.0497	0.0983	0.8098	0.2138	0.1979
DT – Post: EqOdds	0.8400	0.0740	0.1342	0.8043	0.1896	0.0312
RF – Baseline	0.9550	0.0438	0.0556	0.8804	0.4081	0.0833
RF – Pre: Reweigh	**0.9550**	0.0438	0.0556	0.8804	0.4081	0.0833
RF – Post: EqOdds	0.9550	0.0438	0.0556	0.8750	0.3818	0.0833
MLP – Baseline	0.9200	0.0604	0.0709	0.8587	0.3396	0.1562
MLP – Pre: Reweigh	0.9000	0.1062	0.1641	0.8587	0.3612	0.1146
MLP – Post: EqOdds	0.9200	0.0604	0.0709	0.8098	0.3327	0.2083
ADV in-proc	0.8950	0.0392	0.0179	0.8533	0.2859	0.1771
Tuned ADV in-proc	0.9150	**0.0003**	0.0376	0.8696	0.3396	0.0104

Table 26: Bias mitigation results using AIF360 on two male-dominated CVD datasets

Fairlearn: Table 27 presents the application of in- and post-processing mitigation approaches offered by Fairlearn. In this particular context, Exponentiated Gradient (EG) Reduction and Grid Search (GS) Reduction were employed for in-processing, while the Threshold Optimizer (TO) was utilized for post-processing, respectively. Given that in-processing techniques require gradient-optimized models, while KNN is instance-based and thus cannot be used, post-processing is the only option. To enhance the clarity of the results, abbreviations are employed exclusively in the table, while the complete method wording is detailed in the text.

Consequently, post-processing was not an effective mitigation strategy for the KNN models in either dataset. The other models exhibited slight inequality alongside superior accuracy. Neither in-processing nor post-processing led to fairer performance in the DT without reducing overall accuracy and increasing inequalities. Conversely, all bias mitigation approaches applied to RF have been shown to yield fairer performance. The Grid Search reduction under Equalized Odds (EO)

	Mendeley			Composite HF		
Model / Variant	**Accuracy**	**DPD**	**EOD**	**Accuracy**	**DPD**	**EOD**
KNN – Baseline	0.9300	0.0802	0.1752	0.8859	0.3796	0.0521
KNN – TO (DP/EO)	0.9300	0.0802	0.1752	0.8859	0.3796	0.0521
DT – Baseline	0.9050	0.0387	0.0406	0.8098	0.2549	0.0812
DT – EG (EO)	0.9200	0.0418	0.0750	0.8098	0.3212	0.1875
DT – EG (DP)	0.9000	0.0322	0.1562	0.8098	0.2880	0.0300
DT – GS (EO)	0.9050	0.0387	0.0406	0.8098	0.2549	0.0812
DT – GS (DP)	0.9150	0.0692	0.0983	0.7826	0.3727	0.3021
DT – TO (EO)	0.8950	0.0025	0.0250	0.8098	0.3075	0.1771
DT – TO (DP)	0.8900	0.0265	0.1094	0.8152	0.2422	0.1125
RF – Baseline	0.9550	0.0438	0.0688	0.8804	0.4081	0.0833
RF – EG (EO)	0.9550	0.0308	0.0531	0.8804	0.4081	0.0833
RF – EG (DP)	0.9550	0.0438	0.0688	0.8804	0.4081	0.0833
RF – GS (EO)	**0.9650**	0.0127	0.0188	0.8859	0.4676	0.1063
RF – GS (DP)	0.9450	0.0409	0.0821	0.8804	0.3886	0.0729
RF – TO (EO/DP)	0.9600	0.0220	0.0556	0.8750	0.4012	0.0938
MLP – Baseline	0.9200	0.0604	0.0709	0.8587	0.3396	0.1562
MLP – EG (EO/DP)	0.9200	0.0604	0.0709	0.8478	0.3396	0.1667
MLP – GS (EO)	0.9250	**0.0003**	0.1031	0.8478	0.3727	0.1667
MLP – GS (DP)	0.9200	0.0519	0.1145	0.8424	0.3327	0.1771
MLP – TO (EO)	0.9200	0.0604	0.0709	0.8641	0.4366	0.1437
MLP – TO (DP)	0.9200	0.0604	0.0709	0.8533	0.3659	**0.0150**

Table 27: Bias mitigation results using Fairlearn on two male-dominated CVD datasets

was able to increase accuracy while also minimizing the disparities in EOD and DPD. Furthermore, enhancements in regard to the MLP were only possible in DPD, while EOD worsened. In addition, Exponentiated Gradient Reduction and post-processing strategies did not yield any enhancements.

In the case of the DT model of the Composite HF dataset, the efforts to mitigate bias were unsuccessful in the majority of attempts. However, the application of post-processing with DPD yielded a mitigated discrepancy in DPD, accompanied by minor gains in accuracy. These gains were associated with a compromise in fairness,

	Mendeley			Composite HF		
Model / Variant	**Accuracy**	**DPD**	**EOD**	**Accuracy**	**DPD**	**EOD**
KNN – Baseline	0.8900	0.0220	0.1063	0.8315	0.4301	0.1146
KNN – Pre: Reweigh	0.8900	0.0068	0.0043	0.8261	0.3590	0.0208
KNN – Post: EqOdds	0.8700	0.0215	0.1038	0.8315	0.4301	0.1146
DT – Baseline	0.9000	**0.0017**	0.0594	0.8261	0.3439	0.0938
DT – Pre: Reweigh.	0.8950	0.0330	0.0393	0.8533	0.4618	0.2604
DT – Post: EqOdds	0.8850	0.0635	0.0421	0.8152	0.2913	0.0938
RF – Baseline	**0.9250**	0.0285	0.1000	0.8424	0.4106	0.1088
RF – Pre: Reweigh	0.9250	0.0285	0.1000	0.8424	0.4243	0.0312
RF – Post: EqOdds	0.9250	0.0285	0.1000	0.8424	0.4243	0.0312
MLP – Baseline	0.8850	0.0234	0.0219	0.8207	0.4712	0.1888
MLP – Pre: Reweigh	0.8650	0.0669	0.0650	0.7989	0.4322	0.1354
MLP – Post: EqOdds	0.8850	0.0234	**0.0009**	0.7989	0.3133	0.0312
ADV in–proc	0.8800	0.0020	0.0316	0.8207	0.3991	0.0208
Tuned ADV in–proc	0.9050	0.0025	**0.0009**	0.8696	0.3590	0.0208

Table 28: Bias mitigation results using AIF360 on two female-dominated CVD datasets

in terms of EOD, leading to a deterioration in predictive performance. The RF did not respond to in-processing through EG Reduction, overall generating results that were identical to the baseline. No accuracy gain could be achieved by applying bias mitigation to the MLP. Additionally, the fairness gaps increased for at least one metric while improving the other. In summary, the most accurate and fair results for MLP are achieved with the postprocessing method, which yields an accuracy of *85.55%* and an EOD value of *0.0150*, while still showing unequal selection rates with a DPD of *0.3659*.

6.2.2 Bias Mitigation | Scenario II (75F/25M)

The results of gender bias mitigation approaches incorporated from AIF360 and Fairlearn, applied to female-skewed CVD datasets, are illustrated in Tables 28 and 29.

AIF360: The baseline RF model demonstrates the greatest accuracy, exhibiting a moderate bias despite the implementation of an elevated EOD of *0.10*. Consequently, RF is not the optimal model for

deployment in CVD diagnosis. The ADV in-processing approach, which was further improved through tuning, yields the best balance of predictive accuracy and fairness for the Mendeley dataset. With an accuracy rate of *90.5%*, this technique demonstrates fair performance for both genders, with the lowest differences in DPD and EOD at *0.0025* and *0.0009*, respectively.

The tuned ADV approach is the most equitable choice for the Composite HF dataset, as it exhibits high accuracy and balance in error rates through low EOD. Despite this, there remains a significant disparity in the selection rates, as indicated by a DPD of *0.3590*. Applying the Equalized Odds post-processing to the DT is another well-balanced choice for the Composite HF dataset. With a DPD of *0.2913*, this method substantially enhances fairness while nearly fully maintaining accuracy. Other mitigation techniques also show low discrimination in diagnosing true CVD cases as well. However, they still show significant bias in favoring one gender for positive predictions, signified through high DPD.

Fairlearn: Except for the EOD value in the DT model, the fairness indicators exhibited only a minimal degree of bias for the models applied to the Mendeley dataset. The KNN model improves fairness by integrating a TO with DP, which halves the fairness gaps in DPD and EOD while maintaining nearly the same level of accuracy. Although the DT baseline shows only slight inequality in predictive performance, implementing Exponentiated Gradient reduction with Equalized Odds improves the EOD value to *0.0393*. While the application of Exponentiated Gradient reduction and the Threshold Optimizer were ineffective in eliminating any fairness gaps for the RF model, employing Grid Search reduction under DP resulted in moderate accuracy gains and much fairer performance. Furthermore, the MLP model didn't respond to any of the bias mitigation approaches. Consequently, fairer performance could not be facilitated. Of all the algorithms and their various bias-mitigation variants, RF with grid search reduction using DPD yields the most favorable results. It produces outstanding results with *96%* accuracy and respective fairness metrics of *0.0090* and *0.0444*.

The Composite HF dataset revealed substantial disparities in selection rates and moderate inequalities in predictive performance. Therefore, the mitigation approaches applied to KNN were unsuccessful, yielding

the same results as the baseline. The Exponentiated Gradient Reduction with DP applied to the DT model substantially reduces the DPD. Moreover, EOD could also be further minimized without significantly depreciating accuracy. Other mitigation approaches applied to DT either led to a significant decrease in accuracy or an increase in inequality. Grid Search Reduction with EO was the only successful mitigation strategy for RF, while other techniques diminished performance and fairness. Nevertheless, in that particular instance, the improvements in fairness, while maintaining the accuracy level, are minimal. In regard to the MLP, the majority of the mitigation strategies stagnate at the baseline in terms of accuracy, as well as fairness. Implementing Exponentiated Gradient reduction under DP yielded slight accuracy improvements while maintaining the disparities observed in DPD and EOD. Even when integrating the Threshold Optimizer, higher accuracy comes at the cost of substantial increases in both fairness metrics.

	Mendeley			Composite HF		
Model / Variant	**Accuracy**	**DPD**	**EOD**	**Accuracy**	**DPD**	**EOD**
KNN – Baseline	0.8900	0.0220	0.1063	0.8315	0.4301	0.1146
KNN – TO (DP)	0.8850	0.0104	0.0594	0.8315	0.4301	0.1146
KNN – TO (EO)	0.8900	0.0548	0.1094	0.8315	0.4301	0.1146
DT – Baseline	0.9000	0.0017	0.0594	0.8261	0.3439	0.0938
DT – EG (EO)	0.8900	**0.0017**	0.0393	0.8261	0.3497	0.0729
DT – EG (DP)	0.8950	0.0308	0.0615	0.8098	0.2823	0.0413
DT – GS (EO)	0.9000	0.0017	0.0594	0.7337	0.2390	0.0521
DT – GS (DP)	0.8950	0.0308	0.0615	0.8043	0.4239	0.1813
DT – TO (EO)	0.8950	0.0483	0.1750	0.7174	0.3079	0.2917
DT – TO (DP)	0.8950	0.0567	0.1219	0.8207	0.3659	0.1562
RF – Baseline	0.9250	0.0285	0.1000	0.8424	0.4106	0.1088
RF – EG (EO/DP)	0.9250	0.0285	0.1000	0.8424	0.4106	0.1088
RF – GS (EO)	0.9450	0.0178	0.0444	0.8424	0.3911	0.0775
RF – GS (DP)	**0.9600**	0.0090	0.0444	0.8315	0.3911	0.0975
RF – TO (EO/DP)	0.9200	0.0533	0.1000	0.8424	0.4632	0.1400
MLP – Baseline	0.8850	0.0234	**0.0219**	0.8207	0.4712	0.1888
MLP – EG (EO)	0.8850	0.0234	0.0219	0.8207	0.4712	0.1888
MLP – EG (DP)	0.8850	0.0234	0.0219	0.8315	0.4712	0.1688
MLP – GS (EO)	0.8850	0.0234	0.0219	0.8207	0.4712	0.1888
MLP – GS (DP)	0.8850	0.0285	0.1063	0.8207	0.4712	0.1888
MLP – TO (EO)	0.8850	0.0234	0.0219	0.8261	0.4722	0.3438
MLP – TO (DP)	0.8850	0.0234	0.0219	0.8424	0.5512	0.3438

Table 29: Bias mitigation results using Fairlearn on two female-dominated CVD datasets

6.2.3 Bias Mitigation | Scenario III (50F/50M)

The following paragraph presents the results of the bias mitigation applied to the balanced CVD datasets from Scenario III.

AIF360: Table 30 presents the results of bias mitigation achieved through the implementation of AIF360 in CVD datasets where the gender distribution was balanced. Based on the gender-balanced composition of the Mendeley dataset, the initial state of fairness status appears well-balanced in terms of performance for both genders, except for the MLP, which exhibits moderate bias in its fairness metrics. In

	Mendeley			Composite HF		
Model / Variant	**Accuracy**	**DPD**	**EOD**	**Accuracy**	**DPD**	**EOD**
KNN – Baseline	0.9350	0.0150	0.0688	0.8750	0.4380	0.1875
KNN – Pre: Reweigh	0.9200	**0.0068**	0.0120	0.8315	0.3580	0.0417
KNN – Post: EqOdds	0.9350	0.0150	0.0487	0.8750	0.4380	0.1875
DT – Baseline	0.9400	0.0113	0.0444	0.8207	0.3306	0.1042
DT – Pre: Reweigh	0.8650	0.0308	0.1444	0.7391	0.1968	0.1771
DT – Post: EqOdds	0.9250	0.0234	0.0060	0.8261	0.3064	0.1667
RF – Baseline	**0.9400**	0.0373	0.0667	0.8315	0.3464	0.0237
RF – Pre: Reweigh	0.9400	0.0373	0.0667	0.8315	0.3464	**0.0000**
RF – Post: EqOdds	0.9400	0.0373	0.0667	0.8261	0.3533	0.0000
MLP – Baseline	0.9250	0.0997	0.1368	0.8533	0.4243	0.1562
MLP – Pre: Reweigh	0.9000	0.0497	0.0650	0.8315	0.3911	0.1354
MLP – Post: EqOdds	0.9250	0.0997	0.1368	0.8370	0.3453	0.1562
ADV in–proc	0.9250	0.0127	0.0598	0.8587	0.3864	0.0312
Tuned ADV in–proc	0.9250	0.0155	0.0103	0.8641	0.3854	0.0104

Table 30: Bias mitigation results using AIF360 on two gender balanced CVD datasets

this regard, Reweighting as a pre-processing technique yields fairer outcomes at the expense of slightly reduced accuracy. Nevertheless, for the best performance, KNN with Reweighting implemented reaches a good compromise between fairness and predictive performance. With the smallest group disparities indicated by a DPD of *0.0068* and an EOD of *0.0120*, it performs slightly less accurately than the baseline. The applied mitigation methods either decrease fairness for higher performance or fail to enhance both fairness indicators consistently. However, the RF baseline offers balanced fairness and high accuracy without the need for pre- or post-processing.

Fairlearn: The outcomes of Fairlearn's bias mitigation techniques on the CVD datasets with equal gender representation are displayed in Table 31. Moderate bias is still present in the gender-balanced dataset composition, as evidenced by the EOD metric in the Mendeley dataset. Furthermore, the employed MLP model demonstrates moderate bias in DPD, suggesting imbalanced selection rates. Nevertheless, the application of bias mitigation via post-processing to KNN was

	Mendeley			Composite HF		
Model / Variant	**Accuracy**	**DPD**	**EOD**	**Accuracy**	**DPD**	**EOD**
KNN – Baseline	0.9350	0.0150	0.0688	0.8750	0.4380	0.1875
KNN – TO (DP/EO)	0.9350	0.0150	0.0688	0.8750	0.4380	0.1875
DT – Baseline	0.9400	0.0113	0.0444	0.8207	0.3306	0.1042
DT – EG (EO)	0.9300	**0.0017**	0.0444	0.8261	0.4448	0.3229
DT – EG (DP)	0.9300	0.0017	0.0444	0.7826	0.2812	0.0208
DT – GS (EO)	0.9400	0.0113	0.0444	0.8207	0.3306	0.1042
DT – GS (DP)	0.9400	0.0113	0.0444	0.7880	0.2743	0.1979
DT – TO (EO)	0.9350	0.0104	**0.0063**	0.8315	0.3327	0.1667
DT – TO (DP)	0.9300	0.0017	0.0444	0.8424	0.4380	0.3333
RF – Baseline	0.9400	0.0373	0.0667	0.8315	0.3464	0.0237
RF – EG (EO/DP)	0.9400	0.0373	0.0667	0.8315	0.3464	0.0237
RF – GS (EO)	**0.9600**	0.0344	0.0436	0.8587	0.4254	0.0975
RF – GS (DP)	0.9450	0.0025	0.0375	0.8750	0.3991	0.0312
RF – TO (EO/DP)	0.9300	0.0026	0.0375	0.8533	0.4849	0.2000
MLP – Baseline	0.9250	0.0997	0.1368	0.8533	0.4243	0.1562
MLP – EG (EO)	0.9250	0.0997	0.1368	0.8098	0.4106	0.1175
MLP – EG (DP)	0.9250	0.0997	0.1368	0.8261	0.4174	0.1250
MLP – GS (EO)	0.9250	0.0997	0.1368	0.8533	0.4243	0.1562
MLP – GS (DP)	0.9100	0.1344	0.1752	0.8424	0.4791	0.1875
MLP – TO (EO)	0.9150	0.1126	0.1368	0.8478	0.4506	0.3229
MLP – TO (DP)	0.9150	0.1126	0.1368	0.8533	0.4654	0.1875

Table 31: Bias mitigation results using Fairlearn on two gender balanced CVD datasets

unsuccessful, neither under the constraint of DP nor EO. Despite the DT's initially equitable performance, further enhancement of fairness could be achieved through the integration of the Threshold Optimizer. Each post-processing restriction improves the respective fairness metric with only minor accuracy loss. Grid Search under EO achieves the highest overall accuracy when applied to RF, while helping to minimize disparities in error and selection rates. However, the Grid Search constrained to DP yielded the most equitable performance overall while enhancing accuracy in comparison to the baseline. Regarding the MLP applied to the Mendeley dataset, no bias mitigation approaches were capable of reducing the moderate bias present.

The distribution of gender balance did not yield substantial improvement in terms of bias presence in the Composite HF dataset. Notwithstanding the integration of the Threshold Optimizer into the KNN algorithm as a mitigation method, both accuracy and the fairness indicators stagnated at the baseline. With respect to the DT model, enhancements in fairness were only achieved through a considerable decrease in accuracy. Therefore, the employment of the Exponentiated Gradient with a DP constraint results in the most equitable outcomes within the mitigated DT models. With an accuracy rate of *78.26%*, a DPD of *0.2812*, and an EOD of *0.0208*, a fairness gap in terms of selection rate remains. In the case of the Composite HF data, RF did not respond to any of the implemented approaches to bias mitigation. An enhancement in the accuracy of the model was accompanied by an increase in its bias. Bias mitigation for MLP was only achieved through the implementation of Exponentiated Gradient Reduction. Nonetheless, the enhancements in the fairness metrics are only minor, and they simultaneously reduce the overall accuracy of the MLP.

To provide a more comprehensive overview, Figures 32 and 33 condense the mitigation results for each dataset, including the fairness metrics and the accuracy across all scenarios in terms of gender composition. These combined findings reveal several recurring trends regarding the dynamics between the accuracy and fairness of the various mitigation techniques employed. In general, bias reduction increased fairness in all models, however, the extent of the corresponding performance trade-off depended on the model and its mitigation strategy.

Scenario	Model	Tool	Method	Acc	DPD	EOD
75M/25F	KNN	Baseline		0.9300	0.0802	0.1752
		AIF360	Pre: Reweight.	0.9100	0.0542	0.1308
		Fairlearn	Post: ThreshOpt	0.9300	0.0802	0.1752
	DT	Baseline		0.9050	0.0387	0.0103
		AIF360	Pre: Reweight.	0.9300	0.0497	0.0983
		Fairlearn	Post: ThreshOpt	0.8950	0.0025	0.0250
	RF	Baseline		0.9550	0.0438	0.0556
		AIF360	Pre+Post	0.9550	0.0438	0.0556
		Fairlearn	In: GridSearch	**0.9650**	0.0127	0.0188
	MLP	Baseline		0.9200	0.0604	0.0709
		AIF360	Post: EOdds	0.9200	0.0604	0.0709
		Fairlearn	In: GridSearch	0.9250	**0.0003**	0.1031
	ADV	AIF360	FairClassifier	0.9150	**0.0003**	0.0376
75F/25M	KNN	Baseline		0.8900	0.0220	0.1063
		AIF360	Pre: Reweight.	0.8900	0.0068	0.0043
		Fairlearn	Post: ThreshOpt	0.8850	0.0104	0.0594
	DT	Baseline		0.9000	0.0017	0.0594
		AIF360	Pre: Reweight.	0.8950	0.0330	0.0393
		Fairlearn	In: ExpGrad	0.8900	0.0017	0.0393
	RF	Baseline		0.9250	0.0285	0.1000
		AIF360	Pre+Post	0.9250	0.0285	0.1000
		Fairlearn	In: GridSearch	0.9600	0.0090	0.0444
	MLP	Baseline		0.8850	0.0234	0.0219
		AIF360	Post: EOdds	0.8850	0.0234	**0.0009**
		Fairlearn	In+Post	0.8850	0.0234	0.0219
	ADV	AIF360	FairClassifier	0.9050	0.0025	**0.0009**
50F/50M	KNN	Baseline		0.9350	0.0150	0.0688
		AIF360	Pre: Reweight.	0.9200	0.0068	0.0120
		Fairlearn	Post: ThreshOpt	0.9350	0.0150	0.0688
	DT	Baseline		0.9400	0.0113	0.0444
		AIF360	Post: EOdds	0.9250	0.0234	0.0060
		Fairlearn	Post: ThreshOpt	0.9350	0.0150	0.0688
	RF	Baseline		0.9400	0.0373	0.0667
		AIF360	Pre+Post	0.9400	0.0373	0.0667
		Fairlearn	In: GridSearch	0.9450	0.0025	0.0375
	MLP	Baseline		0.9250	0.0997	0.1368
		AIF360	Pre: Reweight.	0.9000	0.0497	0.0650
		Fairlearn	Post: ThreshOpt	0.9150	0.1126	0.1368
	ADV	AIF360	FairClassifier	0.9250	0.0155	0.0103

Table 32: Mitigation results on the Mendeley dataset under different gender distributions.

With relatively minor impacts on accuracy, pre-processing reweighting from AIF360 generated the most reliable decrease in disparity measures for KNN. This implies that KNN benefits greatly from data-level mitigation when applied to the Mendeley data. In the context of DT, no particular

technique emerged as dominant. The implementation of the Equalized Odds from AIF360 and the Threshold Optimizer from Fairlearn resulted in a reduction of bias following post-processing, accompanied by a decline in predictive performance. Considering this, the DT model applied to the Mendeley dataset tends to show a clearer relationship between fairness and performance. The incorporation of the in-processing Grid Search from Fairlearn consistently yielded the most fair outcomes for Random Forest, while simultaneously enhancing fairness and augmenting accuracy. Therefore, in the case of the Mendeley data, the ensemble approach achieved a positive fairness-performance compromise. The MLP model produced less consistent outcomes, although post-processing Equalized Odds yielded the greatest reduction in disparities for the Mendeley data. However, no single solution outperformed the others consistently across the three gender composition scenarios. Finally, the AIF360 fair classifier demonstrated considerable accuracy and equality in both fairness measures. In circumstances where both fairness metrics are pertinent, the ADV method by AIF360 emerges as a highly effective solution, as evidenced by its application to the Mendeley data.

When compared to the results obtained from the Mendeley dataset, the mitigation outcomes for the Composite HF dataset display a more varied pattern. The degree and direction of the fairness-accuracy relationship differed considerably based on gender distribution and the mitigation approach used, despite the achievement of bias reduction across models.

AIF360's Reweighting approach reduced gender disparities for KNN; however, in the case of Scenario I, the male-dominant scenario, this approach was found to be ineffective. AIF360's Reweighting approach reduced gender disparities for KNN applied to the Composite HF data. However, in the case of Scenario I, the male-dominant scenario, this approach was found to be ineffective. Moreover, these gains in fairness were accompanied by modest declines in accuracy, suggesting sensitivity of the model to data redistribution. The Exponentiated Gradient from Fairlearn and Equalized Odds by AIF360 displayed their capacity to mitigate disparities for the DT model across all gender distribution scenarios of the Composite HF dataset. This led to modest declines in performance once more. Subsequently, none of the mitigation categories are universally beneficial for the DT model. Nevertheless, the results demonstrates a preference for fairness constraints applied at the threshold level or during optimization. The most balanced outcomes for RF were accomplished through Fairlearn's Grid Search enhancing fairness while preserving reasonable accuracy.

Scenario	Model	Tool	Method	Acc	DPD	EOD
75M/25F	KNN	Baseline		**0.8859**	0.3796	0.0521
		AIF360	Post: EOdds	0.8859	0.3796	0.0521
		Fairlearn	Post: ThreshOpt	0.8859	0.3796	0.0521
	DT	Baseline		0.8098	0.2549	0.0104
		AIF360	Post: EOdds	0.8043	**0.1896**	0.0312
		Fairlearn	In: ExpGrad	0.8098	0.2549	0.0104
	RF	Baseline		0.8804	0.4081	0.0833
		AIF360	Post: EOdds	0.8750	0.3818	0.0833
		Fairlearn	In: GridSearch	0.8804	0.3886	0.0729
	MLP	Baseline		0.8587	0.3396	0.1562
		AIF360	Pre: Reweight.	0.8587	0.3612	0.1146
		Fairlearn	In: ThreshOpt	0.8533	0.3659	0.0150
	ADV	AIF360	FairClassifier	0.8696	0.3396	0.0312
75F/25M	KNN	Baseline		0.8315	0.4301	0.1146
		AIF360	Pre: Reweight.	0.8261	0.3590	0.0208
		Fairlearn	Post: ThreshOpt	0.8315	0.4301	0.1146
	DT	Baseline		0.8261	0.3439	0.0938
		AIF360	Post: EOdds	0.8152	0.2913	0.0938
		Fairlearn	In: ExpGrad	0.8098	0.2823	0.0413
	RF	Baseline		0.8424	0.4243	0.1288
		AIF360	Pre+Post	0.8424	0.4243	0.0312
		Fairlearn	In: GridSearch	0.8424	0.3911	0.0775
	MLP	Baseline		0.8207	0.4712	0.1888
		AIF360	Post: EOdds	0.7989	0.3133	0.0312
		Fairlearn	In: ExpGrad	0.8315	0.4712	0.1688
	ADV	AIF360	FairClassifier	0.8696	0.3590	0.0208
50F/50M	KNN	Baseline		0.8750	0.4380	0.1875
		AIF360	Pre: Reweight.	0.8315	0.3580	0.0417
		Fairlearn	Post: ThreshOpt	0.8750	0.4380	0.1875
	DT	Baseline		0.8207	0.3306	0.1042
		AIF360	Pre: Reweight.	0.7391	0.1968	0.1771
		Fairlearn	In: ExpGrad	0.7826	0.2812	0.0208
	RF	Baseline	–	0.8315	0.3464	0.0237
		AIF360	Pre: Reweight.	0.8315	0.3464	**0.0000**
		Fairlearn	In+Post	0.8315	0.3464	0.0237
	MLP	Baseline		0.8533	0.4243	0.1562
		AIF360	Post: EOdds	0.8370	0.3453	0.1562
		Fairlearn	In: ExpGrad	0.8098	0.4106	0.1175
	ADV	AIF360	FairClassifier	0.8641	0.3854	0.0104

Table 33: Mitigation results on the Composite HF dataset under different gender distributions.

In contrast, the AIF360 mitigation strategies demonstrated inconsistent outcomes, indicating that ensemble models might respond more favorably to in-processing optimization than to pre- or post-processing modifications. Furthermore, the mitigation methods implemented in MLP were found to

be less consistent when applied to the Composite HF dataset. Regardless of the mitigation strategy applied, mitigation could only improve one fairness metric at a time across all scenarios, with some loss of accuracy. Overall, the mitigation techniques displayed no consistent pattern across the scenarios for the MLP model. Consequently, the extent and direction of the impact on fairness are determined by the gender distribution composition provided to the MLP. Finally, ADV produced consistent results across the scenarios when used on the Composite HF data. Inequalities in selection rates persist, while fairness is enhanced by generating low disparities in error rates and achieving reasonable performance.

Taken together, this comparison indicates that gender bias can be reduced across all models and datasets. However, depending on the model and the underlying gender composition of the respective dataset, the trade-off between accuracy and fairness varies substantially.

7 Discussion

This discussion chapter synthesizes the findings of the legal analysis relevant to the specified use case. Furthermore, it discusses the outcomes of the technical analysis, comparing and contrasting distinct fairness tools and their effectiveness in detecting and mitigating gender bias in AI-based CVD diagnosis. In doing so, the overarching research question, *"How can gender bias in AI-based diagnosis of cardiovascular disease be effectively detected and mitigated in the healthcare sector?"* is explored.

7.1 Discussion of the Research Questions

This subsection addresses the sub-research questions intended to facilitate the investigation of the results concerning the main research question. These sub-research questions are addressed individually, using insights from the preceding chapters.

RQ 1: What are the legal implications of the EU AIA for the employment of bias detection and mitigation methods to detect gender bias in medical data-driven AI systems for predicting CVD?

We based this legal analysis on a concrete use case in which a hospital plans to use an AI-supported tool for diagnosing CVD. The legal basis for our investigation is the EU AIA, which came into effect on 1 August 2024.

Using the IRAC method to analyze this new regulatory framework, we derived practical requirements for detecting and mitigating gender bias in this setting.

Given that the AIA introduces regulation to a scenario that was previously unregulated, the applicable rules for this use case are not yet fully clear. The central concern lies in the identification of the relevant articles for an AI system within the context of medical care, particularly with regard to the management of sensitive patient data and the safeguarding of patient rights, such as the right to equal treatment. In accordance with *Article 6(2)* and Annex III, we categorized the use case as a high-risk scenario, requiring adherence to the provisions outlined in Chapter III. These provisions pertain to data quality and management (*Article 10*), technical documentation (*Article 11*), record keeping (*Article 12*), transparency (*Article 13*), and human oversight (*Article 14*). Consequently, Table 34 maps each applicable article to its concrete requirements and the corresponding compliance measures, thereby showing how the AIA obligations can be translated into this healthcare context. The compliance measures reflect central expectations regarding medical AI systems, namely transparency, trust, and fairness. Nevertheless, several provisions remain ambiguous regarding the identification and mitigation of bias in practice. While certain articles explicitly reference data quality and representativeness, the AIA does not specify operational methods for bias detection or mitigation. Consequently, formal compliance may be met through ineffective measures, resulting in a regulatory gap with potential consequences for patient safety in the diagnosis of CVD.

AIA Article	Requirement	Actions to Ensure Compliance
Article 10	- Data utilized for the training, validation, and testing needs to be representative, relevant, and free of errors - Data employed needs to align with the context of its deployment - Implementation of bias detection and mitigation techniques	- Ensure training, validation, and test data cover both gender groups - Ensure alignment with the deployment setting - Successful bias detection and mitigation techniques
Article 11	- Establishment of technical documentation according to Annex IV	- Explain system development process - Document training procedures - Report performance across gender groups - Specify data origin and describe data characteristics - Document and describe bias detection and mitigation methods
Article 12	- Development of an automatic record-keeping system	- Maintain records for verifiability and error tracking - Monitor performance across demographic groups
Article 13	- Ensure transparency - Description of the system's technical features	- Elaborate on the system's behavior
Article 14	- Establishment of human control and intervention mechanisms	- Enable review and override outputs

Table 34: Concrete Compliance Measures under the AIA for AI-Supported CVD Diagnostic Systems

RQ 2: How can gender bias in cardiovascular disease be identified and addressed?

Given that biases can stem from both the behavior of models and the inherent properties of the data, bias detection demands a comparative evaluation across datasets and model types. For the purpose of gender bias detection in CVD predictions, Fairlearn and FairMLHealth were employed. Both toolkits were used to analyze gender-based differences in model performance through established fairness metrics in three distinct gender distribution scenarios. While both tools offer a comprehensive array of fairness metrics, the primary focus of the comparison lies in the employment of two common metrics: the EOD and the DPD metric.

The findings indicate that, while some datasets show negligible gender differences, others reveal significant disparities in imbalanced scenarios. Consequently, it can be argued that fairness does not uniformly generalize across clinical datasets. Rather, it is impacted by factors such as gender composition and how medical attributes reflect gender-specific patterns of CVD risk. The manifestation of bias is more prominent in scenarios where there is an imbalance in the representation of gender, whereas the balanced scenario may mask that bias. Consequently, rather than depending solely on a single training-test split, a more robust approach involves evaluating bias across multiple distribution scenarios.

Subsequently, the identified bias can be addressed at three stages in the model pipeline: at the data level, on the algorithmic level and at the outcome level. Pre-processing methods, like reweighting, modify the initial distribution of data prior to training. The incorporation of fairness objectives into the learning process of algorithms is achieved through in-processing techniques, such as ADV. Once training is complete, post-processing techniques adjust the decision thresholds to balance error rates in order to reduce bias.

RQ 3: How does using gender bias mitigation strategies in AI-based cardiovascular disease diagnosis affect model performance and fairness?

The employment of bias mitigation techniques from Fairlearn and AIF360 enables the mitigation of gender bias in AI-supported CVD diagnosis, thereby enhancing fairness across genders. However, these gains in fairness are accompanied by a decline in predictive accuracy. The implemented mitigation techniques resulted in a reduction in inequalities in selection and error rates across both datasets. Nonetheless, the impact on predictive performance varies depending on the model type and the underlying gender ratio of the dataset under consideration. In the context of models with less complexity, such as KNN, the application of pre-processing techniques, specifically reweighting, yielded notable enhancements in terms of fairness, while preserving a reasonable level of accuracy. Ensemble models, such as RF, exhibited the most substantial gains from in-processing optimization. With regard to RF, the bias reduction not only enhanced fairness but concurrently preserved or even raised accuracy. Post-processing techniques frequently mitigated discrepancies, but they largely compromised predictive accuracy when implemented. Overall, the findings suggest fairness can be improved without significantly compromising diagnostic accuracy if mitigation measures align with the chosen model and dataset.

7.2 Limitations

This thesis provides valuable insights into the real-world application of the AIA in a specific use case within the healthcare domain. Furthermore, it offers a thorough comparison of the fairness tools and techniques that have been implemented. That being said, there are some shortcomings that need to be taken into account.

The first limitations concern the data utilized. When generating the gender compositions to establish distinct male-dominant, female-dominant, and gender balanced datasets, downsampling the overrepresented group and upsampling the underrepresented group were necessary. Unfortunately, the CVD Kaggle dataset lacks details about its origin and collection. Due to its immense size of 70000 observations, it's unclear whether the data was collected from real patients or created artificially. This would probably explain the subsequent limitation. Regardless of the gender distribution scenario, the CVD Kaggle dataset showed only very minor and negligible indices of bias. However, this outcome is not realistic in practice. Additionally, the specific compositions of the datasets according to the scenarios present another limitation concerning the data. When generating the gender compositions to establish distinct male-dominant, female-dominant, and gender balanced datasets, downsampling the overrepresented group and upsampling the underrepresented group were necessary. This process resulted in data loss and duplication of cases, respectively. Despite the low upsampling count, this should be considered a limitation.

A subsequent constraint manifested in the implementation of fairness tools. The integration of FairMLHealth, the specified fairness instrument for healthcare applications, was a crucial comparison. However, FairMLHealth's lack of bias mitigation approaches resulted in a less insightful comparison of mitigation techniques using AIF360 and Fairlearn.

8 Conclusion

Gender bias constitutes a serious concern in the prediction of CVD diagnosis. Data exhibiting gender bias has the capacity to directly impact diagnostic decisions and subsequently result in treatment disparities. In reality, these disparities frequently manifest as misdiagnoses and mistreatments, thus endangering patients' lives. As AI becomes more prevalent in medical

practices, the need for regulatory frameworks has become increasingly apparent. The EU AI Act is intended to regulate its deployment, especially in contexts where individual safety and health are at risk.
Therefore, the present thesis examines this real-world case through the lens of the following use case: A hospital is in the process of implementing an AI-powered diagnostic tool for cardiovascular diseases. Healthcare professionals would utilize this system to facilitate the diagnosis of cardiac conditions by leveraging diverse patient data and examination findings. In light of this use case, the thesis's investigation is twofold: firstly, it illuminates the legal perspective in accordance with the EU AIA; secondly, it elucidates the practical implications through a technical implementation. Consequently, this work contributes to ongoing legal research on the use of AI in the EU, particularly the use of high-risk AI systems, and to the technical discussion on algorithmic fairness in healthcare, through an empirical evaluation of fairness across various datasets and tools.

The legal analysis, employing the IRAC method, has confirmed that AI-based CVD diagnosis systems fall within the scope of high-risk AI systems, as defined in Article 6 in conjunction with Annex III of the EU AIA. The results of the legal examination indicate that gender bias in medical AI is not solely a technical deficiency, but also a potential infringement on fundamental rights, including the right to non-discrimination, and thus, equal treatment in medical care, regardless of the patient's gender. The AIA stipulates obligations for high-risk systems, including the need for high-quality and representative training data, transparency and human oversight. Moreover, the provisions mandate the implementation of bias detection and mitigation measures to ensure compliance. In consequence, ensuring fairness in AI applications becomes a matter of regulatory necessity for lawful usage.

Accordingly, the legal requirement established by the AIA is associated with the technical implementation of this thesis, namely the gender bias detection and mitigation techniques. Drawing upon this legal foundation, the technical implementation illuminates the practical realization of these fairness requirements concerning gender bias in CVD prediction. The technical analysis revealed that bias demonstrated a high degree of sensitivity to the gender composition of the utilized datasets. Across the three datasets under consideration, different gender representation influences both bias outcomes and model accuracy. Among the models examined, the RF model consistently demonstrated the most robust performance across gender groups, balancing predictive accuracy and fairness. Similar

to KNN and Decision Tree models, the MLP's fairness metrics exhibited increased sensitivity to shifts in gender distribution, despite achieving competitive performance levels. Furthermore, an imbalance in gender ratios has been shown to have a detrimental effect on the underrepresented gender, particularly in the context of comparison error-based metrics such as EOD and DPD, as evidenced by bias detection tools Fairlearn and FairMLHealth. Furthermore, the findings obtained through the utilization of FairMLHealth are largely consistent with those derived from Fairlearn. However, FairMLHealth provides a more comprehensive array of fairness metrics and augmented contextual information, thereby offering supplementary interpretive value.

Applying bias mitigation strategies reveals the trade-off between fairness and predictive ability. Both AIF360 and Fairlearn demonstrate the ability to mitigate gender disparities in all scenarios. However, the extent of mitigation varies depending on factors such as gender distribution, dataset features, the selected mitigation approach, and the stage at which it is implemented. Notably, no single mitigation method outperformed the others consistently across all scenarios. This aligns with the AIA's focus on risk-based assessments in conjunction with continuous human supervision of high-risk AI systems. Furthermore, it underscores the necessity of conducting context-specific fairness evaluations.

Rather than being regarded as a purely computational problem, ensuring fairness in AI-driven healthcare diagnoses must be understood as a multifaceted undertaking, encompassing technical and legal dimensions. Alongside the implementation of effective mitigation tools, ongoing oversight and regulatory awareness are essential for the effective diagnosis of gender bias in CVD. In light of the increasing integration of AI systems in healthcare decision-making processes, the alignment of algorithmic fairness with prevailing legal frameworks, particularly within the context of the AIA, is crucial for assuring the trustworthiness, fairness, and responsible development of AI innovation in the healthcare sector. In conclusion, this thesis underscores the significance of evaluating specific use cases within the context of the new legislative framework, AIA, and further emphasizes the necessity for enhanced specificity to strengthen its practical impact.

References

[1] AI Service Desk | RTR — rtr.at. Retrieved from: https://www.rtr.at/rtr/service/ki-servicestelle/KI-Servicestelle.en.html. [Accessed 11-03-2025].

[2] Annex III: High-Risk AI Systems Referred to in Article 6(2) | EU Artificial Intelligence Act — artificialintelligenceact.eu. Retrieved from: https://artificialintelligenceact.eu/annex/3/. [Accessed 13-03-2025].

[3] Article 10: Data and Data Governance | EU Artificial Intelligence Act — artificialintelligenceact.eu. Retrieved from: https://artificialintelligenceact.eu/article/10/. [Accessed 14-03-2025].

[4] Article 6: Classification Rules for High-Risk AI Systems | EU Artificial Intelligence Act. Retrieved from: https://artificialintelligenceact.eu/article/6/. [Accessed 14-01-2025].

[5] Article 8: Compliance with the Requirements | EU Artificial Intelligence Act — artificialintelligenceact.eu. Retrieved from: https://artificialintelligenceact.eu/article/8/. [Accessed 14-03-2025].

[6] Artificial Intelligence — digitalaustria.gv.at. Retrieved from: https://www.digitalaustria.gv.at/eng/topics/AI.html. [Accessed 11-03-2025].

[7] Cardiovascular Disease dataset — kaggle.com. Retrieved from: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset. [Accessed 05-06-2025].

[8] Cardiovascular_Disease_Dataset — data.mendeley.com. Retrieved from: https://data.mendeley.com/datasets/dzz48mvjht/1. [Accessed 23-07-2025].

[9] Recital 46 | EU-Gesetz — artificialintelligenceact.eu, howpublished = Retrived from: https://artificialintelligenceact.eu/de/recital/46/, year = , note = [Accessed 22-04-2025],.

[10] Understanding Blood Pressure Readings — heart.org. Retrieved from: https://www.heart.org/en/health-topics/

high-blood-pressure/understanding-blood-pressure-readings. [Accessed 14-08-2025].

[11] Article 1: Subject Matter | EU Artificial Intelligence Act. Retrieved from: https://artificialintelligenceact.eu/article/1/, 2024. [Accessed 13-01-2025].

[12] Implementation Timeline | EU Artificial Intelligence Act. Retrieved from: https://artificialintelligenceact.eu/implementation-timeline/, 2024. [Accessed 13-01-2025, 10-03-2025].

[13] Margaux Achtari, Adil Salihu, Olivier Muller, Emmanuel Abbé, Carole Clair, Joëlle Schwarz, and Stephane Fournier. Gender bias in ai's perception of cardiovascular risk. *Journal of Medical Internet Research*, 26:e54242, 2024.

[14] TA Addissouky, I El Tantawy El Sayed, MM Ali, MH Alubiady, and Y Wang. Recent developments in the diagnosis, treatment, and management of cardiovascular diseases through artificial intelligence and other innovative approaches. *J Biomed Res*, 5(1):29–40, 2024.

[15] M.A. Ahmad, C. Eckert, V. Kumar, A. Patel, C. Allen, and A. Teredesai. Fairness in machine learning for healthcare. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, pages 3529–3530. August 2020.

[16] Muhammad Aurangzeb Ahmad, Arpit Patel, Carly Eckert, Vikas Kumar, and Ankur Teredesai. Fairness in machine learning for healthcare. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3529–3530, 2020.

[17] Mugahed A Al-Antari. Artificial intelligence for medical diagnostics-existing and future ai technology, 2023.

[18] Abdullah Al Hamid, Rachel Beckett, Megan Wilson, Zahra Jalal, Ejaz Cheema, Dhiya Al-Jumeily Obe, Thomas Coombs, Komang Ralebitso-Senior, and Sulaf Assi. Gender bias in diagnosis, prevention, and treatment of cardiovascular diseases: A systematic review. *Cureus*, 16(2), 2024.

[19] Mana Saleh Al Reshan, Samina Amin, Muhammad Ali Zeb, Adel Sulaiman, Hani Alshahrani, and Asadullah Shaikh. A robust heart

disease prediction system using hybrid deep neural networks. *IEEE Access*, 11:121574–121591, 2023.

[20] Md Mamun Ali, Bikash Kumar Paul, Kawsar Ahmed, Francis M Bui, Julian MW Quinn, and Mohammad Ali Moni. Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine*, 136:104672, 2021.

[21] C. Allen, M.A. Ahmad, C. Eckert, J. Hu, V. Kumar, and A. Teredesai. fairmlhealth: Tools and tutorials for fairness evaluation in healthcare machine learning. Retrieved from: `https://github.com/KenSciResearch/fairMLHealth`, 2020.

[22] Nadiah A Baghdadi, Sally Mohammed Farghaly Abdelaliem, Amer Malki, Ibrahim Gad, Ashraf Ewis, and Elsayed Atlam. Advanced machine learning techniques for cardiovascular disease early detection and diagnosis. *Journal of Big Data*, 10(1):144, 2023.

[23] Shahab S Band, Atefeh Yarahmadi, Chung-Chian Hsu, Meghdad Biyari, Mehdi Sookhak, Rasoul Ameri, Iman Dehzangi, Anthony Theodore Chronopoulos, and Huey-Wen Liang. Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. *Informatics in Medicine Unlocked*, 40:101286, 2023.

[24] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: an extensible toolkit for detecting. *Understanding, and Mitigating Unwanted Algorithmic Bias*, 2, 2018.

[25] Chintan M Bhatt, Parth Patel, Tarang Ghetia, and Pier Luigi Mazzeo. Effective heart disease prediction using machine learning techniques. *Algorithms*, 16(2):88, 2023.

[26] Proshanta Kumar Bhowmik, Mohammed Nazmul Islam Miah, Md Kafil Uddin, Mir Mohtasam Hossain Sizan, Laxmi Pant, Md Rafiqul Islam, and Nisha Gurung. Advancing heart disease prediction through machine learning: Techniques and insights for improved cardiovascular health. *British Journal of Nursing Studies*, 4(2):35–50, 2024.

[27] Giovanni Briganti and Olivier Le Moine. Artificial intelligence in medicine: today and tomorrow. *Frontiers in medicine*, 7:509744, 2020.

[28] Kelley Burton. Teaching and assessing problem solving: An example of an incremental approach to using irac in legal education. *Journal of University Teaching & Learning Practice*, 13(5):20, 2016.

[29] Kelley Burton. " think like a lawyer" using a legal reasoning grid and criterion-referenced assessment rubric on irac (issue, rule, application, conclusion). *Journal of Learning Design*, 10(2):57–68, 2017.

[30] D Cenitta, N Arul, T Praveen Pai, J Andrew, et al. An explainable transfer learning based residual attention bilstm model for fair and accurate prognosis of ischemic heart disease. *F1000Research*, 14(651):651, 2025.

[31] Nadikatla Chandrasekhar and Samineni Peddakrishna. Enhancing heart disease prediction accuracy through machine learning techniques and optimization. *Processes*, 11(4):1210, 2023.

[32] Feng Chen, Liqin Wang, Julie Hong, Jiaqi Jiang, and Li Zhou. Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models. *Journal of the American Medical Informatics Association*, 31(5):1172–1183, 2024.

[33] Richard J Chen, Judy J Wang, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Ming Y Lu, Sharifa Sahai, and Faisal Mahmood. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature biomedical engineering*, 7(6):719–742, 2023.

[34] Zhenpeng Chen, Jie M Zhang, Federica Sarro, and Mark Harman. A comprehensive empirical study of bias mitigation methods for machine learning classifiers. *ACM transactions on software engineering and methodology*, 32(4):1–30, 2023.

[35] Sribala Vidyadhari Chinta, Zichong Wang, Xingyu Zhang, Thang Doan Viet, Ayesha Kashif, Monique Antoinette Smith, and Wenbin Zhang. Ai-driven healthcare: A survey on ensuring fairness and mitigating bias. *arXiv preprint arXiv:2407.19655*, 2024.

[36] Davide Cirillo, Silvina Catuara-Solarz, Czuee Morey, Emre Guney, Laia Subirats, Simona Mellino, Annalisa Gigante, Alfonso Valencia, María José Rementeria, Antonella Santuccione Chadha, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ digital medicine*, 3(1):1–11, 2020.

[37] Pedro Conceição and Pedro Ferreira. The young person's guide to the theil index: Suggesting intuitive interpretations and exploring analytical applications. 2000.

[38] William DeGroat, Habiba Abdelhalim, Kush Patel, Dinesh Mendhe, Saman Zeeshan, and Zeeshan Ahmed. Discovering biomarkers associated and predicting cardiovascular disease with high accuracy using a novel nexus of machine learning techniques for precision medicine. *Scientific reports*, 14(1):1, 2024.

[39] Shailesh Desai, Atul Munshi, and Devangi Munshi. Gender bias in cardiovascular disease prevention, detection, and management, with specific reference to coronary artery disease. *Journal of mid-life health*, 12(1):8–15, 2021.

[40] Bhanu Prakash Doppala, Debnath Bhattacharyya, Midhunchakkaravarthy Janarthanan, and Namkyun Baik. A reliable machine intelligence model for accurate identification of cardiovascular diseases using ensemble techniques. *Journal of Healthcare Engineering*, 2022(1):2585235, 2022.

[41] P Ignacio Dorado-Díaz, Jesús Sampedro-Gómez, Víctor Vicente-Palacios, and Pedro L Sánchez. Applications of artificial intelligence in cardiology. the future is already here. *Revista Española de Cardiología (English Edition)*, 72(12):1065–1075, 2019.

[42] Barbara Draghi, Zhenchen Wang, Puja Myles, and Allan Tucker. Bayesboost: Identifying and handling bias using synthetic data generators. In *Third International Workshop on Learning with Imbalanced Domains: Theory and Applications*, pages 49–62. PMLR, 2021.

[43] Elias Dritsas and Maria Trigka. Efficient data-driven machine learning models for cardiovascular diseases risk prediction. *Sensors*, 23(3):1161, 2023.

[44] Martin Ebers, Veronica RS Hoch, Frank Rosenkranz, Hannah Ruschemeier, and Björn Steinrötter. The european commissions proposal for an artificial intelligence act - a critical assessment by members of the robotics and ai law society. *J*, 4(4):589–603, 2021.

[45] Hosam El-Sofany, Belgacem Bouallegue, and Yasser M Abd El-Latif. A proposed technique for predicting heart disease using machine

learning algorithms and an explainable ai method. *Scientific Reports*, 14(1):23277, 2024.

[46] Luciano Floridi. The european legislation on ai: a brief analysis of its philosophical approach. *Philosophy & Technology*, 34(2):215–222, 2021.

[47] Joelle H Fong. Disability incidence and functional decline among older adults with major chronic diseases. *BMC geriatrics*, 19(1):323, 2019.

[48] Austrian Regulatory Authority for Broadcasting and Telecommunications. Risk leveles of AI systems | AI Service Desk — rtr.at. https://www.rtr.at/rtr/service/ki-servicestelle/ai-act/risikostufen_ki-systeme.en.html, 2024. [Accessed 10-03-2025].

[49] Xiao-Yan Gao, Abdelmegeid Amin Ali, Hassan Shaban Hassan, and Eman M Anwar. Improving the accuracy for analyzing heart diseases prediction based on the ensemble method. *Complexity*, 2021(1):6663455, 2021.

[50] Armin Garmany, Satsuki Yamada, and Andre Terzic. Longevity leap: mind the healthspan gap. *NPJ Regenerative Medicine*, 6(1):57, 2021.

[51] O Alonso Gelabert, M Barcena Veciana, V Brumwell Valsells, M Crunas Baqe, and Catia Nicodemo. Gender bias in the diagnosis of cardiovascular disorders in catalonia. *Health policy*, 132:104823, 2023.

[52] Pronab Ghosh, Sami Azam, Mirjam Jonkman, Asif Karim, FM Javed Mehedi Shamrat, Eva Ignatious, Shahana Shultana, Abhijith Reddy Beeravolu, and Friso De Boer. Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques. *IEEE Access*, 9:19304–19326, 2021.

[53] Philipp Hacker. A legal framework for ai training data-from first principles to the artificial intelligence act. *Law, innovation and technology*, 13(2):257–301, 2021.

[54] Fereshteh Hasanzadeh, Colin B Josephson, Gabriella Waters, Demilade Adedinsewo, Zahra Azizi, and James A White. Bias recognition and mitigation strategies in artificial intelligence healthcare applications. *NPJ Digital Medicine*, 8(1):154, 2025.

[55] Jonathan Huang, Galal Galal, Mozziyar Etemadi, and Mahesh Vaidyanathan. Evaluation and mitigation of racial bias in clinical machine learning models: scoping review. *JMIR medical informatics*, 10(5):e36388, 2022.

[56] Ali Husnain, Ayesha Saeed, A Hussain, A Ahmad, and MN Gondal. Harnessing ai for early detection of cardiovascular diseases: Insights from predictive models using patient data. *International Journal for Multidisciplinary Research*, 6(5), 2024.

[57] Anna Isaksson. Mitigation measures for addressing gender bias in artificial intelligence within healthcare settings: a critical area of sociological inquiry. *AI & SOCIETY*, pages 1–10, 2024.

[58] Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain, and Preeti Nagrath. Heart disease prediction using machine learning algorithms. In *IOP conference series: materials science and engineering*, volume 1022, page 012072. IOP Publishing, 2021.

[59] Rizwan Karim and Muhammad Asjad. A fair approach to heart disease prediction: Leveraging machine learning model. *Systems Assessment and Engineering Management*, 2, 12 2024.

[60] Rizwan Karim and Muhammad Imran Asjad. A fair approach to heart disease prediction: Leveraging machine learning model. *Systems Assessment and Engineering Management*, 2:23–32, 2024.

[61] M Kavitha, G Gnaneswar, R Dinesh, Y Rohith Sai, and R Sai Suraj. Heart disease prediction using hybrid machine learning model. In *2021 6th international conference on inventive computation technologies (ICICT)*, pages 1329–1333. IEEE, 2021.

[62] Arsalan Khan, Moiz Qureshi, Muhammad Daniyal, and Kassim Tawiah. A novel study on machine learning algorithm-based cardiovascular disease prediction. *Health & Social Care in the Community*, 2023(1):1406060, 2023.

[63] Burak Koçak, Andrea Ponsiglione, Arnaldo Stanzione, Christian Bluethgen, João Santinha, Lorenzo Ugga, Merel Huisman, Michail E Klontzas, Roberto Cannella, and Renato Cuocolo. Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. *Diagnostic and Interventional Radiology*, pages Epub–ahead, 2024.

[64] Can Li, Sirui Ding, Na Zou, Xia Hu, Xiaoqian Jiang, and Kai Zhang. Multi-task learning with dynamic re-weighting to achieve fairness in healthcare predictive modeling. *Journal of biomedical informatics*, 143:104399, 2023.

[65] Fuchen Li, Patrick Wu, Henry H Ong, Josh F Peterson, Wei-Qi Wei, and Juan Zhao. Evaluating and mitigating bias in machine learning models for cardiovascular disease prediction. *Journal of biomedical informatics*, 138:104294, 2023.

[66] Jian Ping Li, Amin Ul Haq, Salah Ud Din, Jalaluddin Khan, Asif Khan, and Abdus Saboor. Heart disease identification method using machine learning classification in e-healthcare. *IEEE access*, 8:107562–107582, 2020.

[67] Aminat Magomedova and Ghizal Fatima. Mental health and well-being in the modern era: a comprehensive review of challenges and interventions. *Cureus*, 17(1), 2025.

[68] Tanjim Mahmud, Anik Barua, Manoara Begum, Eipshita Chakma, Sudhakar Das, and Nahed Sharmen. An improved framework for reliable cardiovascular disease prediction using hybrid ensemble learning. In *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 1–6. IEEE, 2023.

[69] Ekta Maini, Bondu Venkateswarlu, Baljeet Maini, and Dheeraj Marwaha. Machine learning–based heart disease prediction system for indian population: An exploratory study done in south india. *medical journal armed forces india*, 77(3):302–311, 2021.

[70] Vitor Galioti Martini and Lilian Berton. Fairness analysis in ai algorithms in healthcare: A study on post-processing approaches. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 553–564. SBC, 2024.

[71] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[72] Ciro Mennella, Umberto Maniscalco, Giuseppe De Pietro, and Massimo Esposito. Ethical and regulatory challenges of ai technologies in healthcare: A narrative review. *Heliyon*, 10(4), 2024.

[73] Ariana Mihan, Ambarish Pandey, and Harriette GC Van Spall. Artificial intelligence bias in the prediction and detection of cardiovascular disease. *npj Cardiovascular Health*, 1(1):31, 2024.

[74] Ariana Mihan, Ambarish Pandey, and Harriette GC Van Spall. Mitigating the risk of artificial intelligence bias in cardiovascular care. *The Lancet Digital Health*, 6(10):e749–e754, 2024.

[75] Pranav Motarwar, Ankita Duraphe, G Suganya, and M Premalatha. Cognitive approach for heart disease prediction using machine learning. In *2020 international conference on emerging trends in information technology and engineering (ic-ETITE)*, pages 1–5. IEEE, 2020.

[76] G Mukazhanova, Zh Alibiyeva, A Kassenkhan, and N Mukazhanov. Using machine learning algorithms for processing medical data. *Computing & Engineering*, 1(1):13–19, 2023.

[77] THOR STALHANE MYKLEBUST, TOR VATN, and DORTHEA MATHILDE KRISTIN. *AI ACT AND THE AGILE SAFETY PLAN.* Springer, 2025.

[78] Awad Bin Naeem, Biswaranjan Senapati, Dipen Bhuva, Abdelhamid Zaidi, Abhishek Bhuva, Md Sakiul Islam Sudman, and Ayman EM Ahmed. Heart disease detection using feature extraction and artificial neural networks: A sensor-based approach. *IEEE Access*, 2024.

[79] Umarani Nagavelli, Debabrata Samanta, and Partha Chakraborty. Machine learning technology-based heart disease detection models. *Journal of Healthcare Engineering*, 2022(1):7351061, 2022.

[80] A Angel Nancy, Dakshanamoorthy Ravindran, PM Durai Raj Vincent, Kathiravan Srinivasan, and Daniel Gutierrez Reina. Iot-cloud-based smart healthcare monitoring system for heart disease prediction via deep learning. *Electronics*, 11(15):2292, 2022.

[81] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. Bias in data-driven artificial intelligence systems - an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1356, 2020.

[82] Adedayo Ogunpola, Faisal Saeed, Shadi Basurra, Abdullah M Albarrak, and Sultan Noman Qasem. Machine learning-based

predictive models for detection of cardiovascular diseases. *Diagnostics*, 14(2):144, 2024.

[83] Tiago P Pagano, Rafael B Loureiro, Fernanda VN Lisboa, Rodrigo M Peixoto, Guilherme AS Guimarães, Gustavo OR Cruz, Maira M Araujo, Lucas L Santos, Marco AS Cruz, Ewerton LS Oliveira, et al. Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big data and cognitive computing*, 7(1):15, 2023.

[84] Atharva Prakash Parate, Aditya Ajay Iyer, Kanav Gupta, Harsh Porwal, PC Kishoreraja, R Sivakumar, and Rahul Soangra. Review of data bias in healthcare applications. 2024.

[85] Ken Peffers, Tuure Tuunanen, Marcus A Rothenberger, and Samir Chatterjee. A design science research methodology for information systems research. *Journal of management information systems*, 24(3):45–77, 2007.

[86] Pooja Rani, Rajneesh Kumar, Anurag Jain, Rohit Lamba, Ravi Kumar Sachdeva, Karan Kumar, and Manoj Kumar. An extensive review of machine learning and deep learning techniques on heart disease classification and prediction. *Archives of Computational Methods in Engineering*, 31(6):3331–3349, 2024.

[87] Shaina Raza. A machine learning model for predicting, diagnosing, and mitigating health disparities in hospital readmission. *Healthcare Analytics*, 2:100100, 2022.

[88] Maria Restrepo Tique, Oscar Araque, and Luz Adriana Sanchez-Echeverri. Technological advances in the diagnosis of cardiovascular disease: a public health strategy. *International Journal of Environmental Research and Public Health*, 21(8):1083, 2024.

[89] Guoguang Rong, Arnaldo Mendez, Elie Bou Assi, Bo Zhao, and Mohamad Sawan. Artificial intelligence in healthcare: review and prediction case studies. *Engineering*, 6(3):291–301, 2020.

[90] Abdul Saboor, Muhammad Usman, Sikandar Ali, Ali Samad, Muhmmad Faisal Abrar, and Najeeb Ullah. A method for improving prediction of human heart disease using machine learning algorithms. *Mobile Information Systems*, 2022(1):1410169, 2022.

[91] Hossein Sadr, Arsalan Salari, Mohammad Taghi Ashoobi, and Mojdeh Nazari. Cardiovascular disease diagnosis: a holistic approach using the integration of machine learning and deep learning models. *European Journal of Medical Research*, 29(1):455, 2024.

[92] Devansh Shah, Samir Patel, and Santosh Kumar Bharti. Heart disease prediction using machine learning techniques. *SN Computer Science*, 1(6):345, 2020.

[93] Palak Sharma, Priya Maurya, and T Muhammad. Number of chronic conditions and associated functional limitations among older adults: cross-sectional findings from the longitudinal aging study in india. *BMC geriatrics*, 21(1):664, 2021.

[94] Sonish Sivarajkumar, Yufei Huang, and Yanshan Wang. Fair patient model: Mitigating bias in the patient representation learned from the electronic health records. *Journal of biomedical informatics*, 148:104544, 2023.

[95] Polipireddy Srinivas and Rahul Katarya. hyoptxg: Optuna hyper-parameter optimization framework for predicting cardiovascular disease using xgboost. *Biomedical Signal Processing and Control*, 73:103456, 2022.

[96] Isabel Straw, Geraint Rees, and Parashkev Nachev. Sex-based performance disparities in machine learning algorithms for cardiac disease prediction: Exploratory study. *Journal of Medical Internet Research*, 26:e46936, 2024.

[97] Sivakannan Subramani, Neeraj Varshney, M Vijay Anand, Manzoore Elahi M Soudagar, Lamya Ahmed Al-Keridis, Tarun Kumar Upadhyay, Nawaf Alshammari, Mohd Saeed, Kumaran Subramanian, Krishnan Anbarasu, et al. Cardiovascular diseases prediction by machine learning incorporation with deep learning. *Frontiers in medicine*, 10:1150933, 2023.

[98] Md Abu Sufian, Lujain Alsadder, Wahiba Hamzi, Sadia Zaman, ASM Sharifuzzaman Sagar, and Boumediene Hamzi. Mitigating algorithmic bias in ai-driven cardiovascular imaging for fairer diagnostics. *Diagnostics*, 14(23):2675, 2024.

[99] Rivansyah Suhendra, Noviana Husdayanti, Suryadi Suryadi, Ilham Juliwardi, Sanusi Sanusi, Abdurrahman Ridho, Muhammad

Ardiansyah, Murhaban Murhaban, and Ikhsan Ikhsan. Cardiovascular disease prediction using gradient boosting classifier. *Infolitika Journal of Data Science*, 1(2):56–62, 2023.

[100] Satyam Suman, Jakkula Pravalika, Pulluru Manjula, and Umar Farooq. Gender and cvd-does it really matters? *Current Problems in Cardiology*, 48(5):101604, 2023.

[101] Xiaoyu Sun, Yuzhe Yin, Qiwei Yang, and Tianqi Huo. Artificial intelligence in cardiovascular diseases: diagnostic and therapeutic perspectives. *European journal of medical research*, 28(1):242, 2023.

[102] Prasannavenkatesan Theerthagiri and Jyothiprakash Vidya. Cardiovascular disease prediction using recursive feature elimination and gradient boosting classification techniques. *Expert systems*, 39(9):e13064, 2022.

[103] Hannah Van Kolfschooten. Eu regulation of artificial intelligence: Challenges for patients' rights. *Common Market Law Review*, 59(1), 2022.

[104] Amalia Vanacore, Maria Sole Pellegrino, and Armando Ciardiello. Fair evaluation of classifier predictive performance based on binary confusion matrix. *Computational Statistics*, 39(1):363–383, 2024.

[105] Emmanouil P Vardas, Maria Marketou, and Panos E Vardas. Medicine, healthcare and the ai act: gaps, challenges and future implications. *European Heart Journal-Digital Health*, page ztaf041, 2025.

[106] Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. Fairlearn: Assessing and improving fairness of ai systems. *Journal of Machine Learning Research*, 24(257):1–8, 2023.

[107] Roel J Wieringa. *Design science methodology for information systems and software engineering*. Springer, 2014.

[108] Jie Xu, Yunyu Xiao, Wendy Hui Wang, Yue Ning, Elizabeth A Shenkman, Jiang Bian, and Fei Wang. Algorithmic fairness in computational medicine. *EBioMedicine*, 84, 2022.

[109] Zhe Yu, Joymallya Chakraborty, and Tim Menzies. Fairbalance: How to achieve equalized odds with data pre-processing. *IEEE Transactions on Software Engineering*, 2024.

[110] Shasha Zhang, Yuyu Yuan, Zhonghua Yao, Jincui Yang, Xinyan Wang, and Jianwei Tian. Coronary artery disease detection model based on class balancing methods and lightgbm algorithm. *Electronics*, 11(9):1495, 2022.

[111] Chunjie Zhou, Pengfei Dai, Aihua Hou, Zhenxing Zhang, Li Liu, Ali Li, and Fusheng Wang. A comprehensive review of deep learning-based models for heart disease prediction. *Artificial Intelligence Review*, 57(10):263, 2024.

A Appendix

This appendix provides information about the GitHub repository associated with this thesis project. The GitHub repository contains the code, the data, the visualizations and other materials related to the research presented in this thesis project. It can be accessed through the following link: Github - Detecting and Mitigating Gender Bias in AI-driven CVD Diagnosis