Predictive Fairness in Healthcare Patient Readmissions: Equity in Hospital Decision-Making

Master Thesis

Faculty of Economics and Business

MSc. Technology and Operations Management

Author: Biniaris, Nikolaos – S5367417 – n.biniaris@student.rug.nl

First Supervisor: dr. C. Emmanouilidis

Second Supervisor: prof. dr. I.F.A. (Iris F.A.) Vis

Submission date: 28/01/2024

PREFACE

This project poses the last step of my candidacy for the MSc in Technology and Operations Management at the University of Groningen. I would like to sincerely thank Dr. C. Emmanouilidis, as my supervisor, for allowing me to work on a project that highly interests me, for his guidance in shaping the project, and for his feedback and support. Additionally, I would like to thank Prof. Dr. I.F.A. Vis for providing additional feedback and advice, enabling me to define the bounds of the research project, and providing food for thought on my research goals. I am deeply grateful for the support and encouragement of my family, who have always believed in my potential and have provided me with the necessary resources to pursue my academic goals. I am also thankful for the friendships I have made throughout my time at university, which have enriched my experience, and supported me through the process.
Finally, a special thank you to Irene, the person who provided me with the most encouragement and filled me with joy and company during the whole journey.

Thank you all for your support.

ABSTRACT

The expansion of Machine Learning (ML) algorithms into diverse decision-making processes has brought a growing impact of algorithmic decisions on people's lives, and with it, the effect of unintended biases in algorithms has increased. These biases can lead to unfair discrimination, or exacerbate existing inequalities among various groups, raising concerns in social, ethical, and legal realms. Therefore, data bias must be carefully considered, particularly in fields that heavily utilize human data.

This paper aims to present a comprehensive framework for detecting, handling, and mitigating bias within the context of ML model development in healthcare clinical predictions, where biological data are used to detect patient hospital readmissions. Conducting an extensive background analysis, the paper explores the sources of bias and their implications in the healthcare ML development cycle and introduces methods to handle and quantify diverse aspects of unfairness utilizing programming tools. The study also provides empirical findings from healthcare predicting modeling [1], spotlighting biases in ML algorithm performance, particularly in the presence of imbalanced data. Through experimentation with various bias mitigation approaches, further assisted by xAI tools, the research emphasized the role of tailored data engineering during the pre-processing phase in effectively mitigating bias, especially in health datasets. The objective is to achieve equitable predictions and outcomes for diverse patient groups.

The study underscores the critical need for gaining domain knowledge before diving into the data engineering stage, and advocates for the availability of more diverse and open-source healthcare datasets, encompassing a broad range of patient attributes.

Use of GenAI assistance: In this paper, ChatGPT was implemented to assist in grammar, vocabulary, and Python syntax checks. Any output has been critically evaluated by the author, claiming responsibility of the final content.

Key Words: Hospital Decision Making, Fair Machine Learning (ML), Bias in Data and Bias Mitigation, Sensitive Attribute Handling, Diabetes Patients Readmissions, UCI Diabetes Dataset.

[1] The code to the project is available at: https://github.com/NickolasB98/TOM_Thesis

Table of Contents

Register of Tables and Figures 5
List of Abbreviations 6
1. INTRODUCTION 7
1.1 Pursuit of Healthcare Equity 7
1.2 Research Aim 8
2. THEORETICAL BACKGROUND 9
2.1 Understanding and Addressing Bias in ML 10
2.2 Understanding and Defining Fairness in ML 14
2.3 Utilization of Biological Attributes in ML 16
2.4 Utilization of Representative Datasets in ML 17
2.5 ML Models Appropriate for Imbalanced Datasets 20
2.6 The Role of xAI Tools in Healthcare 20
2.7 Diversity Considerations of Health Datasets Utilized in ML 22
2.8 UCI Diabetes Dataset – Utilization in Prior Research 22
2.9 Conclusion and Problem Statement 24
3. METHODOLOGY 24
3.1 Dataset Information and Pre-Processing 24
3.2 Experimental Design: Methods and Tools 28
4. IMPLEMENTATION AND RESULTS 31
4.1 Sensitive Attribute Selection and Dataset's Race Distribution 31
4.2 Implementation of Different Approaches 32
4.3 Model Evaluation and Representativity Across the Approaches 33
4.4 Prediction Accuracy for Race Groups Across the Approaches 37
4.5 Overall Comparison of Unawareness Method 39
4.6 Fairness Exploration Between the New Approaches 42
5. DISCUSSION 52
5.1 Model Performance and Parameters 53
5.2 Addressing the Research Questions 53
5.3 Limitations and Further Extensions 54
5.4 Broader Implications 55
6. CONCLUSION 56
References 57

Register of Tables and Figures

Table 1 Healthcare Types of Bias 11
Table 2 Techniques for Mitigating Bias at Different Stages of ML Model Development 13
Table 3 Advantages and Challenges of Different Bias Mitigation Stages. 14
Table 4 Different Types of Fairness. 15
Table 5 Fair ML Libraries 16
Table 6 Tools/Techniques for Addressing Imbalanced Datasets in ML 19
Table 7 ML Models Suitable for Imbalanced Datasets 20
Table 8 XAI Tools and Libraries 21
Table 9 Findings of Prior Research Paper Focusing on Fairness 23
Table 10 Prior Studies Using the UCI Diabetes Dataset 23
Table 11 Types of Bias Found in Research and Linkage to Literature 32
Table 12 Grouped Codes Pre-Processing Approach 33
Table 13 Precision-Recall Curves Conclusion 37
Table 14 Comparison of Unawareness Method and New Approaches 41

Figure 1 Datasets with Diversity Characteristics. 22
Figure 2 Dataset Overview 25
Figure 3 Flowchart: from Dataset Import until Model Training 26
Figure 4 Distribution of Race Groups and their Readmissions. 32
Figure 5 AUC Scores Across the First Set of Classifiers 33
Figure 6 AUC Scores Across the Second Set of Classifiers 34
Figure 7 Race Distribution Training Set 35
Figure 8 Race Distribution Logarithmic Scale Training Set 35
Figure 9 Precision-Recall Curve – Unaware Grouped Codes Scenario 36
Figure 10 Precision-Recall Curve - Top Codes Scenario 36
Figure 11 Precision-Recall Curve - One Code Scenario 37
Figure 12 Accuracy Differences of Race Groups 38
Figure 13 Total Comparison of Unawareness Method 40
Figure 14 TNR Comparison of Race Groups 41
Figure 15 TPR Comparison of Race Groups 41
Figure 16 Top Codes Fairness Metrics Differences across Scenarios 42

Figure 17 One Code Fairness Metrics Differences across Scenarios 42
Figure 18 Fairness Check for First Approach 43
Figure 19 Fairness Radar for First Approach 44
Figure 20 Stacked Cumulated Parity Loss for First Approach 45
Figure 21 SHAP Implementation in RF for First Approach 45
Figure 22 SHAP Implementation in XGBoost for First Approach 46
Figure 23 ELI5 Implementation for First Approach 47
Figure 24 Fairness Check for Second Approach 48
Figure 25 Fairness Radar for Second Approach 49
Figure 26 Stacked Cumulated Parity Loss for Second Approach 49
Figure 27 SHAP Implementation in RF in Second Approach 50
Figure 28 SHAP Implementation in XGB in Second Approach 51
Figure 29 ELI5 Implementation for Second Approach 52

List of Abbreviations

AI	Artificial Intelligence
AIF360	AI Fairness 360
AP	Average Precision
CatBoost	Category Boosting
Dalex	Descriptive mAchine Learning EXplantations
ELI5	Explain Like I'm 5
FNR	False Negative Rate
FPR	False Positive Rate
ICD-9	International Classification of Diseases, 9th Revision
ML	Machine Learning
PER	Predictive Equality Ratio
RF	Random Forest
SHAP	Shapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
STP-SPR	Statistical Parity Rate
TNR	True Negative Rate
TPR	True Positive Rate
xAI	Explainable Artificial Intelligence
XGBoost	eXtreme Gradient Boosting

1. INTRODUCTION

Machine learning (ML) has steadily become an integral tool in various aspects of healthcare, offering new pathways for enhancing diagnostics, patient care, and administrative efficiency. The imperative for this principal domain emanates from the escalating need for more accurate diagnostics, efficient clinical predictions, and streamlined administrative processes, particularly in the medical field, and is driven by the hope to enhance clinical processes and overall healthcare provision. The growth of medical data, coupled with technological advancement in big data analytics, has facilitated AI's expansion in disease management, leading to more informed, efficient, and comprehensive care. (Saxena & Chandra, 2021)

However, the growing reliance on ML applications in medicine raises substantial concerns regarding their ethical implications and the propensity to perpetuate biases and harm, leading to the exacerbation of pre-existing health disparities and adversely affecting marginalized populations. For instance, certain ML-based software applied in hospitals has been found to reinforce social disparities, favoring one group of patients over others who might be more in need of care. There have been instances where state-of-the-art clinical prediction models have manifested underperformance on ethnic minorities, women, and those with public insurance, potentially leading to biased and unjust healthcare outcomes. (R. J. Chen et al., 2021) Such disparities and biases embedded in algorithms can reflect and perpetuate systemic inequities, adversely affecting marginalized populations and contributing to the persistence of racial inequities in healthcare use and disease outcomes. (Gervasi et al., 2022)

1.1 Pursuit of Healthcare Equity

The pursuit of healthcare equity is paramount in the current healthcare and public health landscape. With ML's expanding applications in healthcare, it is crucial to precisely inspect its implications on equity. Studies have revealed disparate predictive accuracies of ML algorithms across different racial and social groups. (Barton et al., 2023) Analyzing disparities between socially advantaged and disadvantaged populations, it becomes fundamental to identify the root causes, such as biased data or algorithms, and to strategize solutions addressing these challenges. (Rojas et al., 2022)

1.1.1 The Dilemma of Utilizing Biological Attributes in Machine Learning

Mitigating healthcare disparities and addressing ML biases demand a clear understanding between fairly incorporating or completely omitting biological attributes, which fall into the category of protected attributes, during the training of predictive models. One methodology proposed for bias mitigation is namely the "Fairness through Unawareness". It proposes that by making models unaware of protected attributes like race and gender, fairer outcomes can be achieved. (J. Chen et al., 2019) However, this method has limitations, as shown in recent literature, in comparison with more appropriate bias mitigation techniques. (R. J. Chen et al., 2021) The significance of including protected labels properly is highlighted, emphasizing that while such inclusion can refine model accuracy in recognizing diverse genetic data and biological factors, unsuited handling can also lead to biased output. In the European Union, these attributes need to be hidden, as per the General Data Protection Regulation (GDPR) passed in 2018. (GDPR, 2020) However, the omission or inappropriate handling of such data might unintentionally exacerbate unfairness in healthcare ML predictions, considering the substantial influence of biological attributes on clinical outcomes and model accuracies.

1.1.2 Diabetes Patient Readmissions: A Focus Area

To demonstrate the practical application of the theoretical insights, this paper focuses on the specific context of hospital readmissions for diabetes patients. This serves as a metric of quality and a vital parameter reflecting the economic implications and the level of provided healthcare. The examination of readmission rates is central for discussions on health equity and fairness, in case potential unequal readmission rates are observed, leading to adverse treatment of socially disadvantaged groups. Also, the escalating number of diabetics globally, from 108 million in 1980 to a projected 642 million by 2040, underscores the importance of efficient hospital care and post-discharge follow-ups. (Mahmoud et al., 2023)

1.2 Research Aim

This paper aims to explore the impact of imbalanced datasets and biological attribute handling, focusing on fairness, within the context of predicting diabetes patient readmissions. A quantitative approach is used, grounded in the empirical analysis of a diabetes readmissions dataset to investigate the below interconnected research questions.

1.2.1 Research Questions:

Q1: "How does efficient pre-processing handling of imbalanced data contribute to the mitigation of bias in machine learning models for hospital readmission predictions within diverse patient communities?"

Q2: "How effective is the unawareness method of omitting sensitive attributes from the model training, compared to attribute-inclusive pre-processing methods?"

The findings of this research will offer evidence regarding determining a suitable approach to handling racial attributes in healthcare ML predictive applications: either incorporating them carefully or omitting them altogether to promote fairness among diverse patient groups. The paper will begin with exploratory data analysis with different data engineering approaches, followed by the development and evaluation of several machine learning models. Performance fairness metrics, comparisons between the unawareness method and other techniques, and experiments with ethnic representativity will inform the study's conclusions.

1.2.2 Dataset

This research will delve into practical investigation to answer the research objectives, taking a focused look at an important specific case, that of hospital readmissions for diabetes patients. This will be accomplished by utilizing the "Diabetes 130-Hospitals" dataset sourced from the UCI Machine Learning repository. Since the paper discusses the management of hyperglycemia in hospitalized patients, the types of diabetes studied are both 1 and 2. (UCI, 2014). This dataset, encompassing ten years of clinical care data from 130 U.S. hospitals, provides a suitable foundation to explore the impacts of biological attribute handling and data pre-processing, with a focus on fairness.

The remainder of the paper is organized as follows: Section 2 reviews the relevant information and literature, sufficiently covering aspects relevant to the research's topics, and concludes with the problem statement. Section 3 then introduces the proposed quantitative research method, this section elaborates on the data used, the engineering approaches implemented, and how performance and fairness will be assessed. Finally, Section 4 presents the results of the case studies, followed by a discussion on limitations and further research opportunities in Section 5 and a brief conclusion in Section 6.

2. THEORETICAL BACKGROUND

This section begins by examining biases in the pipeline of applying machine learning to healthcare predictions and their implications in decision-making. A comprehensive examination of the approaches to address these biases, as well as a comprehensive understanding of fairness, forms a big part of this exploration. The ethical considerations of omitting biological attributes are highlighted, followed by the role of representative datasets in model accuracy and fairness. Appropriate ML models for imbalanced datasets are analyzed,

and the role of xAI tools in healthcare ML. Then, the diversity of healthcare datasets with fairness implications, including a look at our utilized dataset is presented. Various papers using the dataset are studied, along with their model performances, and a prior case study that explores the dataset from a fairness perspective.

2.1 Understanding and Addressing Bias in ML

2.1.1 Manifestation of Biases

Bias can emerge in every step of the ML process: from data collection and pre-processing to the development of ML algorithms, and even during interpreting the results. (Giovanola & Tiribelli, 2023) for example, point out that flawed data labeling can lead to qualitatively skewed datasets, which can in turn propagate bias into the resulting ML technologies. Biases in datasets can also arise inherently from their source, from historically skewed human decisions, biased device measurements, or erroneous reports. (Pessach & Shmueli, 2020) An exploration into biases in healthcare ML, illuminated by the insights offered by (Giovanola & Tiribelli, 2023), unveils the many-sided way of how biases can permeate various layers and interactions within ML models. They can be categorized into four major domains, as per Table 1:

Bias in Healthcare	Types
Emerging from Model Design	
	Label Biases: In the labeling phase, biases arise when labels utilized as proxies are imprecise, not meaning the same across all patient contexts and thus, being subject to healthcare disparities instead of reflecting a universal truth.
	Cohort Biases: Emerges when the focus is predominantly on conventional or easily measured groups, often overlooking other potentially protected groups or granular levels.
Emerging from Training Data	
	Minority Bias: Ensues when there is insufficient representation of patients in a protected group, inhibiting the model from learning accurate statistical patterns.
	Missing Data Bias: Occurs due to a lack of patient data from protected groups, making accurate prediction challenging. For instance, clinical deterioration may be under-detected in isolated patients due to fewer vital sign recordings.
	Informativeness Bias: Present when available features are less informative for making accurate predictions in a protected group, such as identifying melanoma in patients with darker skin.
	Training–Serving Skew: Takes place when ML is applied to patients whose data diverge from the training data, potentially due to non-standardized data collection methods or disparate data recording schemas. Notably, Watson for Oncology, although widely utilized in China for diagnostic predictions via image recognition, showcased suboptimal results for Chinese patients compared to their Western counterparts, having been primarily trained on Western datasets. (Liu et al., 2018)
Resulting from ML-Clinician Interactions	
	Automation Bias: Comes into play when clinicians, perhaps unknowingly of inaccuracies in the ML model for a protected group, excessively rely on the ML model, thereby acting on potentially flawed predictions.
	Feedback Loop Biases: Develop when clinicians accept model recommendations, even incorrect ones, leading the model to erroneously learn from these affirmations.
	Dismissal Biases: Evolve from clinicians becoming insensitively accustomed to alerts that are systematically inaccurate for a protected group.
	Allocation Discrepancy Biases: Occur when resources, predicated by predictions (like additional clinical attention or social services), are disproportionately withheld from a particular group.
Resulting from ML-Patient Interactions	
	Privilege Bias: Certain models might be inaccessible in environments where protected groups receive care or necessitate technologies disproportionately available to non-protected classes, thereby amplifying existing disparities.
	Informed Mistrust Bias: Stemming from patient skepticism, this bias is especially pertinent among protected groups who may believe that a model is biased against them, possibly leading them to avoid seeking care or deliberately withholding information.
	Agency Bias: Often interlinked with privilege bias, it manifests when protected groups lack input into model development, usage, and evaluation, thereby depriving them of resources or political influence to discern biases and facilitate correction.

Table 1 Healthcare Types of Bias

Formed by Insights from (Giovanola & Tiribelli, 2023)

2.1.2 Effects of Biases in Healthcare

The effects of biases are significant, hindering the development of generalized healthcare solutions and yielding treatments that may be unequal for underrepresented groups. (Pot et al., 2021; Rajkomar et al., 2018) Biases influence the precision of algorithms, particularly affecting individuals with lower socioeconomic status and racial minorities, leading to disparities in

healthcare access and later disease detection. (Pot et al., 2021) They do not just affect the performance of ML models, but can also increase healthcare inequalities, or negatively impact certain demographic groups, by interfering with equitable resource allocation or cultivating grounds for unintentional discriminatory practices, as evidenced by these studies.

2.1.3 The Challenge of High Data Quality in Healthcare

The use of ML in healthcare faces challenges due to the data complexity, which goes beyond volume and is highly diverse (e.g. electronic health records, patient observations, imaging data). (Feldman et al., 2017) This complication is further intensified by their multidimensional and dynamic healthcare nature (new treatments, procedures, medical technologies), as well as practical challenges such as missing or corrupted data, and the need to incorporate heterogeneous and multimodal data. (Di Camillo et al., 2020; Johnson et al., 2016) In addition, the unique challenges of learning in a clinical setting, such as poorly labeled data and subgroup underrepresentation, further complicate the proper use of ML methodologies. (Ghassemi et al., 2020) These challenges highlight the need for higher data quality in the health domain, otherwise unintentional bias can arise.

2.1.4 Intentional Biases and their Role.

Ensuring ML fairness sometimes requires the intentional incorporation of biases. (Pot et al., 2021) The utilization of intentional biases in ML is complex, with potential benefits and risks. The authors (Gurupur & Wan, 2020; Pot et al., 2021) highlight the intentional biases to resolve historical inaccuracies and enhance representation, particularly in healthcare. For example, oversampling data from economically deprived populations or darker-skinned individuals in image learning algorithms can enhance their representation, ensuring more equitable algorithm training. (Pot et al., 2021) However, intentional biases can also lead to ethical concerns (Gaonkar et al., 2020), and therefore must be carefully managed to avoid negative consequences.

2.1.5 Mitigation of Unintentional Bias

Addressing unconscious bias presents a significant challenge due to its inherent nature: biases that remain unrecognized cannot be directly countered. ML developers play a crucial role in navigating this challenge, necessitating a deep understanding of protected features interdependencies, and sensitivities with the rest data, while ensuring that utilized datasets are accurately representative of the population under study. Adopting strict standards for data annotations, cautiously identifying features associated with target attributes, and exercising

caution to prevent bias emerging from ML algorithm training, through meticulous data analysis and validation, are fundamental practices in bias mitigation. (Srinivasan & Chander, 2021)

ML algorithms also offer a promising future for unintended bias mitigation, with their capability to identify patterns, particularly in datasets of a scale beyond human processing capabilities. (Zhang et al., 2018) present a framework for mitigating unwanted biases in ML through adversarial learning, which has the potential to counteract built-in biases. However, (Lum, 2017) cautions that ML algorithms may not be effective in eliminating judicial bias, as they can reproduce existing biases. Therefore, while ML can help uncover and confront unaware biases better than humans, it is not a foolproof solution and should be used along with human-in-the-loop methods.

Table 2 shows techniques used to reduce biases during different ML phases. The methods are divided based on the time they are used: pre-processing, in-processing, or post-processing. Each method is briefly explained with examples for clear understanding, The libraries and tools highlighted here present a foundational framework for ML practitioners.

Technique Category	***Sub-category***	***Brief Description***	***Example Methods/Algorithms***
Pre-processing	Relabelling and Perturbation	Modify truth labels or dataset features to balance it.	Disparate Impact Remover, Massaging (Kamiran and Calders)
	Sampling	Change the distribution of samples to ensure fairness.	SMOTE, Reweighing Algorithm
	Representation	Learn new data representations to reduce bias.	Learning Fair Representation (LFR), Prejudice Free Representations (PFR)
In-processing	Regularization and Constraints	Adjust loss functions or impose constraints to enhance fairness during model training.	Prejudice Remover, Exponentiated Gradient Reduction, Meta Fair Classifier
	Adversarial Learning	Train competing models to enhance performance and fairness.	Adversarial Debiasing
	Adjusted Learning	Develop or adjust algorithms to mitigate bias during training.	Multi-Party Computation (MPC) methods for Logistic or Linear Regression
Post-processing	Input Correction	Modify testing data after model training to affect model inputs.	Gradient Feature Auditing (GFA)
	Classifier Correction	Adjust trained classification models to enhance fairness.	Linear Programming (LP), Calibrated Equalized Odds
	Output Correction	Modify model outcomes directly to yield fairer predictions.	Reject Option based Classification (ROC), Randomized Threshold Optimizer

Table 2 Techniques for Mitigating Bias at Different Stages of ML Model Development

Below is an exploration of these divisions' advantages and challenges, as there is no universally applicable method to confront bias. The table is formed by insights from (Pessach & Shmueli, 2020)

Pre-Processing Mechanisms
• Advantages: Pre-process mechanisms boast versatility, being applicable across various classification algorithms.
• Challenges: Their non-specificity to particular algorithms may cause unpredictability regarding the final accuracy level. Furthermore, by altering data before it reaches the algorithm, they might diminish the explainability of the outcomes, making it challenging to decipher how input variations affect predictions.
In-Processing Mechanisms
• Advantages: Offering the capacity to directly influence the model during training, in-processing mechanisms facilitate an explicit balancing act between fairness and accuracy within the algorithm's objective function.
• Challenges: These mechanisms are intimately tied to machine learning algorithms, limiting their adaptability across diverse algorithmic structures and potentially complicating their integration and optimization.
Post-Processing Mechanisms
• Advantages: Applicable to any classification algorithm, post-process mechanisms offer flexibility and enable bias mitigation without tweaking the original model.
• Challenges: Applied after model training, their intervention is belated and may yield suboptimal results compared to mechanisms integrated earlier in the learning process. They also grapple with ethical dilemmas, potentially initiating a skewed compensation by deliberately sacrificing accuracy for certain individuals to redress biases, which might be perceived as a form of reverse discrimination.

Table 3 Advantages and Challenges of Different Bias Mitigation Stages.

Formed by insights from (Pessach & Shmueli, 2020)

Although a range of fairness mechanisms have been proposed in the literature, no single method consistently outperforms others across different datasets and cases, as shown in the findings of (Pagano et al., 2023).

2.2 Understanding and Defining Fairness in ML

2.2.1 Fairness is a Multifaceted Concept

Fairness in machine learning is a complicated and multidimensional concept, influenced by various challenges such as bias in the data, model inaccuracies, and the potential for unintended unfairness coming from algorithm output. (Pessach & Shmueli, 2022) The need for a standard definition of fairness in the ML community is emphasized, with a call for broader community

involvement in defining what is fair. (Skirpan & Gorelick, 2017) (Makhlouf et al., 2021) further explore the applicability of different fairness notions in real-world scenarios, providing a decision diagram to guide practitioners and policymakers in selecting the most suitable fairness notion. (Gajane & Pechenizkiy, 2017) illuminate fairness in ML as the pursuit of impartiality and unbiasedness towards individuals or groups, categorized into group and individual fairness, which is further supported by (Mehrabi et al., 2019), who present fairness as the absence of any prejudice or favoritism toward an individual or groups based on their inherent or acquired characteristics.

The Table represents different types of fairness formed by insights from (Mehrabi et al., 2019).

Name	Fairness Type	Brief Description
Equalized Odds	Group	Ensures equal true and false positive rates across both protected and unprotected groups.
Equal Opportunity	Group	Mandates identical probabilities of positive class assignment for all demographic sectors.
Demographic Parity	Group	Requires equal likelihoods of favorable outcomes across every demographic subset.
Fairness Through Awareness	Individual	Allocates similar predictions to similar individuals based on a specified similarity metric.
Fairness Through Unawareness	Individual	Advocates for the omission of protected attributes in decision-making to uphold fairness.
Treatment Equality	Group	Maintains a consistent ratio of false negatives to positives across all protected group categories.
Test Fairness	Group	Ascertains that score S represents a consistent likelihood of a specific outcome regardless of group membership.
Counterfactual Fairness	Individual	Ensures consistent decisions across both real and hypothetical scenarios, even when demographic associations change.
Fairness in Relational Domains	Individual	Integrates the relational structures between individuals when assessing fairness, surpassing individual attributes.
Conditional Statistical Parity	Group	Advocates for equal probabilities of positive outcome assignment, conditional upon a set of legitimate factors L.

Table 4 Different Types of Fairness.

Formed by insights from (Mehrabi et al., 2019)

This section suggests that the coexistence of all notions of fairness simultaneously is challenging. (Mehrabi et al., 2019) points out that ethical aspects of fairness in ML are closely tied to societal contexts and the involved stakeholders.

2.2.2 Tools and Techniques for Ensuring Fairness in ML

In the field of healthcare ML, there is a growing need for fair AI tools and libraries. These resources aim to ensure that models are fair and ethical in their applications. Table 5 presents libraries used in Fair ML, the most frequently used including AIF360, Fairlearn, and Themis-ML.

Library	*Description*	*Source/Repository*
FairSight	Assists in achieving fair decision-making through visual analytics tools.	FairSight Source
AIF360	Provides algorithms to detect and mitigate bias in ML models across the AI application lifecycle.	AIF360 GitHub
Fairlearn	Facilitates fair ML and supports various fairness metrics.	Fairlearn Website
Themis-ML	Provides methods to enforce and measure group fairness in ML models.	Themis-ML GitHub
Debiaswe	Designed to debias word embeddings in natural language processing.	Debiaswe GitHub
ML-fairness-gym	A platform to study the long-run impact of ML systems using reinforcement learning environments.	ML-fairness-gym GitHub
fairness-in-ml	Focuses on fair and transparent machine learning.	fairness-in-ml GitHub
bias-correction	Provides tools and methods for removing bias from ML models.	bias-correction GitHub
Fairness-comparison	Offers functionality for the benchmarking of fairness-aware ML algorithms.	Fairness-comparison GitHub
BlackBoxAuditing	Features a sample implementation of Gradient Feature Auditing (GFA).	BlackBoxAuditing GitHub
Aequitas	An open-source bias audit toolkit for ML models, developers, and policymakers.	Aequitas GitHub
fairness-indicators	A TensorFlow Model Analysis tool that evaluates model performance using fairness indicators and visualizations.	fairness-indicators GitHub
Responsible-AI-Toolbox	Provides various group-fairness metrics across sensitive features.	Responsible AI Toolbox
LinkedIn Fairness Toolkit (LiFT)	Provides fairness metrics and tools for ML models.	LiFT Blog Post
Responsibly	Aims at auditing and mitigating bias and ensuring fairness in ML applications.	Responsibly GitHub
smclarify	Amazon's tool for bias detection and mitigation in datasets and models.	smclarify GitHub
inFairness	A PyTorch package designed to train and audit ML models for individual fairness.	inFairness GitHub
Dalex	Provides explanations of model predictions and addresses both explainability and fairness in ML.	Dalex GitHub

Table 5 Fair ML Libraries

2.3 Utilization of Biological Attributes in ML

This section addresses the ethical and practical challenges of incorporating biological attributes in healthcare analytics and ML models. It explores the balance between accurate data representation and the risk of bias perpetuation. The focus is on the impact of including or omitting such attributes on the fairness and accuracy of models within existing legal and ethical data policies.

2.3.1 Ethnicity Data Quality and Regulations in Healthcare Analytics

(Gervasi et al., 2022) emphasize the need for comprehensive regulatory and oversight policies to ensure the ethical use of ethnicity data in healthcare analytics. (Aspinall & Jacobson, 2007). argue that despite concerns over the quality of ethnicity data, it should still be used to identify disparities in healthcare. They emphasize the need for healthcare organizations to engage actively with existing data, acknowledging its limitations while recognizing its potential to highlight ethnic inequalities, helping also to drive improvements in data quality and healthcare equity.

2.3.2 Sensitive Attributes Implications on the Fairness-Accuracy Trade-off

There is a debate between fairness and accuracy in ML algorithms, (Caton & Haas, 2020) but proper utilization of protected attributes has made significant strides in addressing this issue. (Han et al., 2023) suggest that explicitly considering sensitive attributes can indeed foster fairer outcomes in ML outputs. They introduce the Fairness Feedback Benchmark (FFB) as a testament to the feasibility of harmonizing accuracy with fairness. (Deho et al., 2023) found that the impact of maintaining or not the protected attributes depends on their correlation with the target label. Lastly, (Yan et al., 2020) propose Fair Class Balancing, achieving fair outputs by balancing the group distribution of training data across multiple sensitive attributes before training.

2.3.3 Bias Elimination, Contextual Integrity, and GDPR Compliance

The use of sensitive biological data in machine learning is necessary to ensure non-discrimination. (Žliobaitė & Custers, 2016) However, this creates a dilemma with privacy concerns, particularly under the GDPR. (Baumann & Rumberger, 2018) More specifically, Article 9 of GDPR protects 'special categories of personal data', such as racial, ethnic details, and biometric data. This regulation acts as a shield for personal data, also mandating user consent before data collection and allowing individuals to know the utilization of their data in applications. (GDPR, 2020). As per (Bogen et al., 2020), omitting or including sensitive data can substantially impact the fairness and accuracy of ML algorithms in real-world applications. The GDPR's impact on AI algorithms and automated decision-making is significant and valuable, but there are still some deficiencies, particularly related to specific consent and the scope of the subject's right to information. (Allen & Masters, 2020)

2.4 Utilization of Representative Datasets in ML

This section explores the role of representative datasets in the development of ML models, particularly in healthcare. It examines how the composition and representativeness of datasets

influence the fairness and accuracy of model predictions, highlighting the necessity of balanced and inclusive data in improved model outputs, ensuring equitable outcomes.

2.4.1 Evaluation of Representativeness in Health Databases

Representative datasets are important for training ML models in healthcare, ensuring fair and equitable predictions (Petersen et al., 2023; Yuan et al., 2021), as diverse representation in training data enhances subgroup performances and contributes to achieving population-level objectives. (Rolf et al., 2021) To achieve this, (Wojtusiak, 2021) emphasizes the need for detailed reporting of results, proposing ten criteria for model evaluation. (Kleinberg et al., 2023) further explores the concept of representativeness, presenting a statistical framework for creating balanced datasets. This approach, which considers group-specific performance metrics and learning rates, outperforms traditional sampling strategies. Their findings underscore the importance of addressing racial underrepresentation in dermatological datasets, as biased models can exacerbate healthcare disparities.

A meta-analysis by (Kamikubo et al., 2022) underlines the need for more representative and inclusive healthcare datasets for improving model training, pointing out the underrepresentation of gender and race in accessibility datasets. This is further emphasized by (Clemmensen & Kjærsgaard, 2022), who stresses the potential of balanced datasets in promoting equitable predictions.

2.4.2 Challenges in Compilation and Documentation

The compilation of representative datasets is a difficult task, as highlighted by (Clemmensen & Kjærsgaard, 2022; Kamikubo et al., 2022). Challenges include inconsistencies and a lack of standardized documentation in reporting demographic-related metadata, as well as the collection of sensitive attribute data, which raises concerns about participant consent and reidentification risks. (Dixon et al., 2015) discuss the biases and representativeness issues in electronic health records, underscoring the need for rigorous assessment and mitigation strategies.

2.4.3 Tools and Techniques for Dealing with Imbalanced Datasets in ML

Datasets that are highly non-representative and fail to accurately depict the real-world distribution are commonly referred to as imbalanced datasets. To address imbalances, various population sampling techniques have been proposed, with Synthetic Minority Oversampling Technique (SMOTE) being a popular choice. (Elreedy & Atiya, 2019; Santoso et al., 2019) found that integrating SMOTE into classification algorithms can significantly improve accuracy. (Chawla et al., 2002) further demonstrated that a combination of oversampling the

minority class and undersampling the majority class, namely hybrid sampling tools, can achieve even better performance.

However, in most cases, oversampling alone is favored in highly imbalanced data (Oskouei & Bigham, 2017), as undersampling can result in the loss of critical information by eliminating instances from the dataset, thus from the model training. Lastly, some techniques handle imbalances without the need for data sampling. The table below provides an overview of these tools and techniques.

Tool / Technique	*Description*	+	-
SMOTE (Oversampling)	An oversampling technique generating synthetic samples for the minority class to overcome the overfitting problem posed by random oversampling.	Mitigates bias and captures important features of the minority class, contributing to more accurate predictions and better model performance.	Can create noisy data points in feature space and complicate the decision surface for some classifiers
Tomek (Undersampling)	Tomek links are pairs of very close instances but of opposite classes. Removing the instances of the majority class of each pair increases the space between the two classes, facilitating the classification process.	Increases the space between classes, potentially improving classifier performance.	May remove useful data points, losing potentially important information.
ENN (Undersampling)	The nearest neighbors of each of the majority class is estimated. If the nearest neighbors misclassify that particular instance of the majority class, then that instance gets deleted.	Can improve the quality of the dataset by removing noisy and borderline examples that may cause overfitting, leading to a cleaner decision boundary.	Risks eliminating not only noise but also potentially informative minority class instances, which can result in the loss of valuable data and possibly lead to underfitting.
ADASYN (Oversampling)	Adapts the amount of synthetic data generated for each minority instance based on its level of difficulty in classification. It focuses on the minority instances that are more challenging to classify, creating synthetic samples accordingly	Addresses class imbalance by emphasizing the generation of synthetic data for the minority class instances that are harder to classify.	May increase computational complexity and may still introduce some noise into the dataset.
SMOTE & Tomek (Hybrid)	Enhances the quality of synthetic sample generation by combining over-sampling of the minority class with cleaning overlaps between classes, potentially leading to better model generalization.	Aims to clean overlapping data points between classes, potentially reducing overfitting and enhancing model performance.	May still generate noisy data; effectiveness depends on the extent of class overlap and specific dataset characteristics.
SMOTE & ENN (Hybrid)	Integrates SMOTE oversampling with ENN undersampling to perform extensive data cleaning for clearer class separation.	Balances classes by generating synthetic samples and then refining them by removing any misclassified instances, leading to a more distinct class separation.	May result in overcleaning the dataset, where too many valuable instances are discarded, potentially causing loss of important information and underfitting.
Change the Performance Metric	Changing the performance metric involves using more suitable metrics for imbalanced datasets, such as F1 score or AUROC, instead of accuracy.	Provides a more accurate assessment of model performance on imbalanced datasets.	May lead to more complex evaluation processes that are not as straightforward as accuracy.
Penalize Algorithms (Cost-Sensitive Training)	Involves using penalized learning algorithms that increase the cost of misclassification of the minority class.	Focuses the learning process on the minority class, which can improve performance on this often-underrepresented class.	Can introduce a bias towards the minority class, potentially leading to overfitting.
Change the Algorithm	Changing the algorithm refers to using machine learning algorithms better suited for imbalanced datasets, like tree ensembles.	Algorithms like decision trees and tree ensembles may inherently perform better on imbalanced datasets.	Not all algorithms will perform well on all types of imbalanced datasets; selection is crucial.

Table 6 Tools/Techniques for Addressing Imbalanced Datasets in ML

2.5 ML Models Appropriate for Imbalanced Datasets

Machine-learning practitioners have identified models better suited for handling imbalanced datasets, with various studies exploring the use of ensemble models like Random Forest. (Zhu et al., 2018) introduces the concept of class weights in RF, which assigns individual weights to each class to enhance overall performance. XGBoost and CatBoost are both effective for imbalanced classifications, with the former's weighted loss function and the latter's robustness to overfitting being their key strengths. (Li & Mao, 2014) However, XGBoost's computational intensity and parameter tuning can be challenging, along with CatBoost's potential speed issues on large datasets. Both models are also complex for interpretation, called black-box models, but are known to outperform in multi-class imbalanced datasets. (Tanha et al., 2020)

Model	***Description***	+	-
Random Forest	Ensemble of decision trees for classification, using bootstrap samples and random feature subsets.	High performance on a variety of problems due to capturing complex data structures.	May not perform well on highly imbalanced datasets, with bias towards the majority class.
Random Forest with Class Weighting	A modified version of Random Forest that adjusts the class weights to emphasize the minority class.	Improves performance for the minority class by addressing the imbalance issue.	May increase the false positive rate for the majority class due to adjusted class weights.
XGBoost	An efficient and scalable implementation of gradient boosting that uses a weighted loss function to address imbalances.	Superior performance and handling of weighted classes.	Computationally intensive with complex parameter tuning.
CatBoost	Gradient boosting algorithm that handles categorical features and is robust to imbalanced data distributions.	Directly processes categorical features and is robust to overfitting.	Potentially slower on very large datasets and complex to interpret.
Bagging with Random Undersampling	A modification of bagging that applies random undersampling to the majority class within each bootstrap sample.	Addresses class imbalance by creating balanced samples during the ensemble training.	May not consider class imbalance in its initial sampling.
Easy Ensemble	An ensemble technique using AdaBoost with resampling, creating balanced samples for training.	Effectively addresses class imbalance by using resampling and boosting.	May require more computational resources due to boosting iterations.

Table 7 ML Models Suitable for Imbalanced Datasets

The use of black-box algorithms in healthcare, while effective, presents challenges in terms of output interpretability (Rasheed et al., 2021), being particularly vital in the health domain. This has led to the development of Explainable Artificial Intelligence (xAI) tools, which aim to make these models more transparent and understandable.

2.6 The Role of xAI Tools in Healthcare

(Mathews, 2019; Pawar et al., 2020) both emphasize the importance of xAI in healthcare, with Mathews specifically highlighting the use of LIME in NLP and cancer detection, and Pawar proposing a solution for integrating xAI into standard ML workflows. (Linardatos et al., 2021)

provides a comprehensive view of xAI methods, while (Adadi & Berrada, 2020) discuss the impact of xAI on healthcare, particularly in terms of trust and transparency. These studies collectively underscore the crucial role of xAI in enhancing the interpretability of ML models in healthcare. However, it is vital to consider the strengths and weaknesses of these methods, as they can change the interpretation of model output. (Salih et al., 2023)

Tool / Library	***Description***	***Key Features***
InterpretML	An open-source library providing a suite of interpretability tools for ML models, with model-agnostic and model-specific techniques.	Global/local interpretability, feature importance, summary plots, individual explanations, compatibility with ML frameworks.
SHAP (SHapley Additive exPlanations)	A library for computing Shapley values to understand feature contributions, supporting a range of models for feature importance analysis.	Model-agnostic SHAP values, unified feature importance measure, support for various model types.
LIME (Local Interpretable Model-agnostic Explanations)	Generates locally faithful explanations for model predictions by approximating the decision boundary around specific instances.	Model-agnostic, local interpretations, support for text/image/tabular data.
ANCHORS	Offers high-precision, model-agnostic explanations with clear 'if-then' rules and defined coverage, improving upon LIME's limitations.	Offers clear 'if-then' rules that delineate the conditions under which the explanations are valid, ensuring transparency.
DALEX	Package for exploration, explanation, and validation of ML models, offering visualizations and a consistent API for interpretation methods.	Consistent API, visualizations for model understanding, model validation techniques.
ELI5 (Explain Like I'm 5)	Library that explains ML model predictions, providing feature importances and textual explanations for various models.	Model-agnostic feature importances, textual explanations, compatibility with ML libraries.
AI Explainability 360	An IBM toolkit offering a comprehensive set of algorithms and tools for interpretable and explainable AI, with bias detection capabilities.	Model-agnostic techniques, fairness and bias detection, compatibility with scikit-learn and TensorFlow.
SHAPash	Built on top of SHAP, it offers a user-friendly interface for visualizing and understanding model predictions with simplified SHAP value interpretation.	Simplified interface for SHAP, visualizations for feature contributions, supports classification and regression models.
Alibi-Explain	Part of the Alibi library focusing on model explanation, offering methods like counterfactual and anchor explanations for model decisions.	Model-agnostic explanation methods, counterfactual/anchor explanations, support for tabular/image data.

Table 8 XAI Tools and Libraries

Table 8 illustrates a selection of xAI tools and libraries. They range from model-agnostic techniques that offer global and local insights into model predictions, to specialized libraries designed for feature importance analysis and explanation frameworks. The diversity of these tools underlines the need for careful selection of the most suitable tool for every healthcare case.

2.7 Diversity Considerations of Health Datasets Utilized in ML

The health domain is a diverse field, comprising various subdomains such as medicine, pharmacology, and psychology, each demonstrating its distinct concerns or challenges for deploying AI applications. The essence of addressing fairness within this domain involves designing models that are uniformly effective across varying patient demographics, particularly for race sub-populations and vulnerable groups. (Fletcher et al., 2021)

Most datasets come from patient data collected at medical centers. They provide information on a variety of health conditions. Datasets like Framingham and the Infant Health and Development Program, track data over long periods, helping researchers study health trends over time. (Fabris et al., 2022) In our case, this is useful for understanding fairness in different clinical predictions and building unbiased algorithms. However, the quality of these datasets is central, as biases from bad-quality data can lead to discriminatory classifications and perpetuate structural marginalization (Quaresmini & Primiero, 2023), or to an unfair allocation of resources and excessive health risks for certain groups. (Feng et al., 2022)

The figure below provides a list of Health Datasets with Diverse Characteristics, suitable to test potential issues of fairness.

Health	
Sleep medicine	Apnea
Critical care medicine	MIMIC-III
Public health	Kidney exchange program, willingness-to-pay for vaccine, kidney matching
Cardiology	Arrhythmia, heart disease, framingham
Neurology	Epileptic seizures
Pediatrics	Infant Health and Development Program (IHDP)
Dermatology	HAM10000, SIIM-ISIC melanoma classification
Medicine	Stanford medicine research data repository
Pharmacology	Warfarin
Endocrinology	Diabetes 130-US Hospitals, Pima Indians diabetes dataset (PIDD)
Nephrology	Renal failure
Radiology	CheXpert, MIMIC-CXR-JPG, national lung screening trial (NLST)
Health policy	Heritage Health, MEPS-HC
Applied psychology	Drug consumption
Experimental psychology	FACES

Figure 1 Datasets with Diversity Characteristics.

Source: (Fabris et al., 2022)

2.8 UCI Diabetes Dataset – Utilization in Prior Research

2.8.1 Overview of the UCI Diabetes Dataset

The UCI Diabetes dataset, sourced from the UCI machine learning repository (UCI Diabetes Dataset et al., 2014), comprises approximately 100,000 instances collected from 130 hospitals across the US over a decade (1999-2008). This dataset provides a detailed representation of

diabetic encounters, characterized by 55 features. These features range from demographic details and diagnoses to medications. The comprehensiveness of this dataset offers a significant opportunity for analysis, making it a prevalent choice among ML researchers.

2.8.2 Building upon Prior Research and Initial Framework

After researching the same dataset for fairness issues, using the AIF360 Fairness Library to find and mitigate bias, and exploring the importance of the sensitive attribute of Race with Lime (xAI Tool), the findings of this paper (Aldrees et al., 2022) are illustrated below:

Libraries/Tools	*Use*	*Evaluations*
AIF360	Revealed high bias against Asian Race Group	Found Minority Bias, Training-Serving Skew Bias
LIME	Generated a perturbed dataset and trained a model, providing localized prediction explanations	Found Race feature having high importance by explaining the XGBoost model's decisions. Especially for Asian Race, the feature was negatively affecting the final output of readmission.

Table 9 Findings of Prior Research Paper (Aldrees et al., 2022) focusing on Fairness

Several studies have also leveraged the UCI Diabetes Dataset for predictive analyses aiming to achieve the highest model performance. However, the outcomes of these studies have varied significantly, often due to disparities in data pre-processing approaches and ML algorithms employed. A brief overview of some notable studies include:

Study	*Best Performances*	*Comments*
(Hempstalk et al., 2016)	Logistic Regression with AUC: 0.670 (10-fold training result)	Developed classic machine learning models for 30-day readmission risk predictions without any explainability. Only provided training performance, with no test AUC data.
(Alloghani et al., 2019)	Naïve Bayes with AUC: 0.640	Focused on recognizing patterns and combinations of risk factors.
(Shang et al., 2021)	Random Forest with AUC: 0.64 (over-sampling) and AUC: 0.661 (down-sampling)	Employed varied machine learning classifiers for the 30-day readmission prediction, with the random forest model demonstrating the optimal performance. Lacked deeper analysis into feature importance.
(Ahmed & Kabir, 2019)	Multilayer Perceptron with AUC: 0.6548	Worked on reducing the number of features and identifying the critical ones that had higher impact in predictions
(Lu & Uddin, 2022)	Stacking model with AUC: 0.6736	Aimed to evaluate the 30-day readmissions risk for diabetic patients using explainable AI techniques

Table 10: Prior Studies Using the UCI Diabetes Dataset

The AUC Score represents the ability of a classifier to distinguish between classes, in our case Not Readmitted patients against Readmitted. An AUC Score of 1 indicates a perfect classifier, while an AUC Score of 0.5 suggests that the classifier is no better than random guessing. As per Table 10, it is evident that the model output scores of these past papers are not very high. This underscores the importance of suitable data engineering and pre-processing, and also

implies that potentially the dataset has some limitations, like missing labels or important information.

2.9 Conclusion and Problem Statement

As presented in this chapter, the presence of biases, data imbalances, and underrepresentation in healthcare predictions poses a significant challenge to developing fair and accurate ML applications. While various tools and techniques exist to address the presented challenges, a focus on the essential role of proper data engineering, distinct for every different case, is missing in the literature for achieving ML fairness. Noteworthily, the important diversion of past papers utilizing this dataset was on the algorithms chosen, and not on data engineering during the pre-processing phase, which impacts overall model performance. This identified gap serves as a foundational ground in our research, prompting an exploration in the subsequent methodology chapter to experiment with careful data engineering approaches tailored to the specific case. This exploration aims to enhance both model accuracy and fairness, dealing with challenges posed by the dataset's imbalances while striving for more representative inclusion of minorities.

3. METHODOLOGY

The pursuit of proper data engineering approaches aligns with the core research objective of this thesis. By exploring different approaches during data pre-processing, the phase before model training, this study aims to address the identified gaps in prior research. These approaches are further examined with the use of xAI tools, to illustrate their importance in deciding for the fairest and most balanced approach.

Secondly, the examination and comparison of the unawareness method as an alternative to different data-preprocessing approaches will answer how biological attributes are more appropriately handled within the healthcare domain. All of these techniques are implemented before reaching the model training phase.

3.1 Dataset Information and Pre-Processing

3.1.1 Reason for Utilization

The selection of the UCI Diabetes dataset was driven by a consideration of critical criteria aligning with this paper's objectives. Primarily, this dataset stood out due to its substantial size and the inherent challenge of class imbalance, essential factors mirroring real-world challenges within the health domain. The dataset's focus on clinical predictions, encompassing a wide array of patient features, aligns closely with the research aim to analyze the proper handling of

biological attributes. Its long-term information gathering (1999-2008) provides a rich foundation to apply different approaches to its engineering methods and see the impact on model outcomes.

3.1.2 Dataset Overview

Feature Name	Type	Feature Name	Type
Encounter ID	Numeric	Number of medications	Numeric
Patient number	Numeric	Number of outpatient visits	Numeric
Race	Nominal	Number of emergency visits	Numeric
Gender	Nominal	Number of inpatient visits	Numeric
Age	Nominal	Diagnosis 1	Nominal
Weight	Numeric	Diagnosis 2	Nominal
Admission type	Numeric	Diagnosis 3	Nominal
Discharge disposition	Numeric	Number of diagnosis	Numeric
Admission source	Numeric	Glucose serum test result	Nominal
Time in hospital	Numeric	A1c test result	Nominal
Payer code	Nominal	Change of medications	Nominal
Medical specialty	Nominal	Diabetes medications	Nominal
Number of lab procedures	Numeric	23 features for medications	Nominal
Number of procedures	Numeric	Readmitted	Nominal

Figure 2 Dataset Overview

Furthermore, it can be confirmed that 27 of 50 attributes are associated with medications and their changes, the patient's characteristics are the remaining attributes (5 for personal features, 10 for patient's encounter, 7 for their historical data). A distinctive 'Readmitted' column serves as our target feature. This column classifiers patients based on their readmission pattern:

- Within 30 days post-discharge
- After 30 days post-discharge
- No readmission

3.1.3 New Data Pre-Processing Approaches

The flowchart illustrates our roadmap, from importing the dataset, until we reach the model training phase. It clearly shows the two pre-processing approaches we conduct, along with the frequent approach as implemented in most papers. The rationale behind the new approaches lies behind tailored ways to experiment with the diagnosis codes utilization, which are directly linked to patient readmissions, a decision rooted in a thorough analysis of the domain.

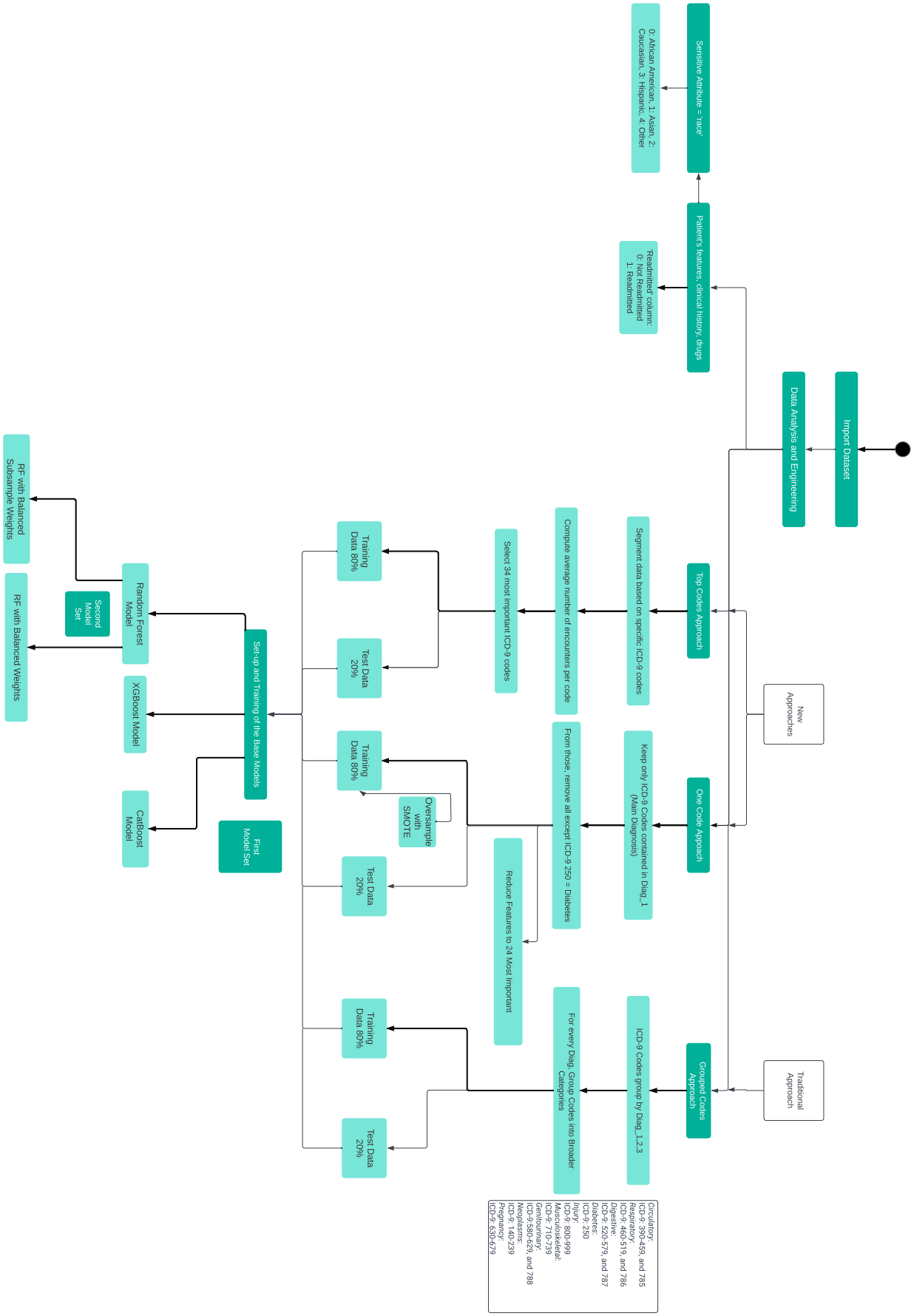

Figure 3 Flowchart: from Dataset Import until Model Training

3.1.3.1 Top Codes Approach

A primary observation is that Readmission is intrinsically linked to the diagnosis, according to the dataset's paper. (UCI Diabetes Dataset et al., 2014) It is essential to highlight that each ICD-9 Code corresponds to a distinct diagnosis. From a methodological standpoint, segmenting the data based on broad diagnosis categories, as done in past papers, may not be as insightful as examining it based on specific types of diagnosis codes.

To address the limitations identified in previous studies and potentially improve predictive performance, our first approach is as follows:

- Instead of grouping the ICD-9 Codes, we will segment the data based on specific codes. Given the vast number of unique ICD-9 codes in the dataset, we compute the average number of encounters per code. Our analysis will focus on a selected number of ICD-9 codes that exhibit the highest numbers of encounters, meaning they are the most common diagnoses.
- Data entries with smaller numbers of encounters per ICD-9 codes will be excluded, leaving us with 28446 data points. This filtering ensures our models train on the most significant and representative data.

This data engineering strategy prioritizes codes that offer the most valuable insights to classifier training, filtering out codes with a lesser impact on the final prediction. By narrowing our focus to data with the highest encounters of ICD-9, we anticipate enhanced model performance and a decrease in imbalances between different race groups.

3.1.3.2 One Code Approach

In the original paper of the dataset (UCI Diabetes Dataset et al., 2014) it has been described that the probability of hospital readmission depends on the primary diagnosis of the patients (feature name: diag_1). The ICD-9 Code 250 denotes that a patient is diagnosed with diabetes. The data points that have a diag_1 value in this range of ICD-9:250 consider the patients who have diabetes assigned as the primary diagnosis. Only these data points will remain for this data pre-processing approach. SMOTE oversampling tool is also implemented in the training set. 8772 data points fall under this criterion. The dataset's dimensionality was reduced to the 24 most significant features. This approach aims to narrow down the information to the minimal point of keeping only the patients directly linked with Diabetes as their main diagnosis, eliminating the rest of ICD-9, the secondary diagnoses, and keeping low feature dimensionality.

3.2 Experimental Design: Methods and Tools

The experiment was conducted in Python (version 3.9), within a Jupyter environment, utilizing several key libraries: Skicit-Learn (version 1.1.3), serving as the foundation for implementing ML models, FairLearn (version 0.8.0), providing tools for assessing fairness metrics, AIF360 (version 0.5.0), offering bias mitigation techniques and fairness evaluation criteria, and Dalex (version 1.6.0) for fairness metrics visualization through diverse plots and frameworks. The experimental setup is divided into five steps; each step is detailed below:

3.2.1 Exploratory analysis of the data:

This entails the representation of distributions, counts, and correlations of groups and attributes. This step aims to understand the representativeness of different features in the data, the relationship between the sensitive attribute and other features, and determine the sample size of different groups. This is crucial to comprehend what types of fairness measured can and should be used, and what are the best possible engineering approaches or bias mitigation techniques. This step also determines if there is a label and class imbalance. For this step, the features are encoded accordingly. The sensitive attribute of Race was encoded as such: 0: African American, 1: Asian, 2: Caucasian, 3: Hispanic, 4: Other

The target label by 0: Not Readmitted, 1: Readmitted.

3.2.2 Set-up and training of the base models:

The task of the models is to detect the decision of hospital readmission through classification. After the data cleaning, transformation, and engineering, all features in the dataset have been labeled accordingly to allow ML classifiers to learn to detect the health status of a record; supervised learning is, therefore, possible.

To allow comparability between classifiers and their performance in handling the data imbalances, as shown in Table 7, an XGBoost, a CatBoost, and a Random Forest have been set up for all scenarios, representing the first model set. Additionally, one RF with Balanced Subsample Weights and one with Balanced Weights are set up, comprising the second model set.

For all classifiers, the data have been split into training and testing sets, employing an 80% - 20% ratio. All the performances that are reported in this section are test performances.

3.2.3 Model Evaluation on Performance and Fairness:

Our case focuses on Recall, AP, and AUC Scores as the most suited performance metrics. This is primarily due to the class imbalances found on the sets and our focus on fairness issues.

While Precision measures the authenticity of predicted positives, it serves as a metric to improve resource allocation costs, so that healthy patients are not falsely identified as in need of Readmission. Recall assesses the accuracy of correctly predicting the patients that are in actual need of Readmission. Our considerations for focusing on Recall to achieve more fair results are:

- Patient Safety: If you miss a necessary readmission (a False Negative), a patient might suffer adverse outcomes. High Recall ensures more patients who need readmission are identified.
- Public Health: From a broader health system perspective, catching more potential readmissions (even at the risk of higher false positives, namely precision) could prevent complications and more severe health issues that could be more costly and difficult to treat later.

We are also utilizing the Precision-Recall curves, along with their AP Scores. Average Precision score is a metric that assesses a model's overall performance by considering Precision-Recall trade-offs across various decision thresholds. It quantifies the area under the curve, illustrating the model's capability to maintain high recall without sacrificing precision, resulting in better balance across these two contradicting metrics. Regardless of our high recall priorities, we aim to achieve a balance, without sacrificing one metric for the other. Also, a higher AUC score indicates better differentiation between the Readmitted and Non-Readmitted, which is essential for our case, while given the scores from previous papers, optimizing this metric becomes a priority. Additionally, cross-validation in the form of a k-fold validation, enables a comprehensive assessment of the evaluation metrics.

When looking at fairness metrics, it is important to consider the task at hand. When considering fairness, one should aim to answer the question: Which race groups are most likely to be disproportionately negatively affected? And What are the harms of adverse treatment?

In this context, a false negative classification signifies that a patient who validly requires readmission is not receiving the necessary medical intervention. This raises concerns about equity, as unequal rates of false negatives across racial groups could perpetuate healthcare disparities. To address this issue, the models should strive for Equalized Odds, aiming to minimize false negatives while simultaneously controlling false positives, also preventing resource allocation costs. This balanced approach ensures that patients from all backgrounds receive equitable access to appropriate care without compromising resource efficiency. For a

fair model, we expect each group's FNR (also called miss rate) to be equal. This will be measured using the maximum absolute difference between the group's accuracies in the Fairlearn MetricFrame tool. The Balanced Error Rate Difference will be also considered, to consider the overall fallout rate of the model (FPR). Also, we will assess the Demographic Parity Difference to ensure that the proportion of positive predictions is roughly equal across different race groups. The objective is to reach results close to 0 for all metrics.

During the implementation of Dalex Library, the chosen metrics for our focus on assessing fairness are the Equal Opportunity Ratio (equal Recall among groups), the Statistical Parity Ratio (balance in the distribution of positive predictions, meaning how equally distributed are the Readmissions between the groups), and the Predictive Equality Ratio (equal Precision among groups). Dalex Library provides an epsilon parameter, set at 0.2, which serves as a threshold in the U.S. (EEOC, 1978), and if surpassed, represents a fairness violation of each metric. The deviations are examined for the protected race groups and the remaining groups, while the base is represented by the privileged group.

3.2.4 Bias Mitigation and Imbalance Handling Tools:

As described in the literature, bias mitigation methods can be applied in different stages, but our paper focuses on the pre-processing phase. The only technique that is categorized as in-processing, is the change of weights in our second RF model set, but this is for imbalances handling and not for directly mitigating bias.

The first step is to mitigate bias through reweighting using the AIF360 library on the first set of models. Note that pre-processing is the earliest opportunity for intervening in the data; this method offers the most opportunities for downstream metrics and is therefore preferred in practice. For the second model set, this technique is not implemented, as another way of reweighting is happening through the sample weight parameters of the RF models.

We next explore the application of "Fairness through Unawareness", implementing it while having engineered the data with the traditional Grouping of ICD-9 Codes. The rationale is to compare the unawareness technique implemented in the traditional data pre-processing approach, against the different aware pre-processing scenarios. We also applied the unawareness method in the new approaches to compare all scenarios.

Lastly, we implemented techniques addressing class imbalances. Noteworthily, sampling is implemented on balancing the target label, in our case the Readmissions label. As a result of this sampling, the race group distribution also changes. To guarantee a realistic evaluation of the models' performance, sampling tools will be applied only to the training data, leaving the

test data intact to avoid an overly optimistic outcome, by maintaining the test set's original demographic representation. Also, the second model set represents another way of handling class imbalances, without the need to change the sampling.

3.2.5 xAI Tools and Libraries Implementation

We aim to provide transparency and comprehensibility in understanding how our models make predictions and why specific features play important roles in these predictions. To accomplish this, we leverage a suite of xAI tools and libraries as found in Table 8, carefully selected for their ability to reveal model decisions. The tools will be used to dive into the model results, illustrating the importance of the Race attribute on their decisions, while Dalex will illustrate potential bias among the chosen fairness metrics, in diverse and easily understandable plots. Below is a brief overview of each xAI tool implemented in this research:

- SHAP values provide a unified measure of feature importance, helping us understand the feature contributions of each feature to local model predictions.
- ELI5 simplifies complex model predictions, offering feature importance and their weights on decisions on the global explanations level and not on single instances.
- Dalex is a library used for both Fairness Mitigation and Model Explainability. In our case, we use it to illustrate potential violations of fairness metrics. Its epsilon parameter, as analyzed during its implementation, is valuable to acknowledge whether a fairness metric shows violation comparing the protected groups, to the privileged.

4. IMPLEMENTATION AND RESULTS

This chapter elaborates on the implementation of the abovementioned methodology and the results of the study.

4.1 Sensitive Attribute Selection and Dataset's Race Distribution

As stated in the background analysis, previous research has shown that imbalances in the race attribute can impact the performance of classifiers. This study will primarily consider race as a sensitive attribute, focusing on fair outcomes among race groups. According to Table 11, certain types of Bias have been found for this sensitive attribute. Connecting with the Bias Types in healthcare presented in Figure 1, the types found in (Aldrees et al., 2022)
are:

Type of Bias	Description	Example in our case
Minority bias	Refers to the uneven prediction accuracy across different groups, potentially leading to healthcare disparities. Occurs when models perform better for the majority group than for minority groups.	A model may be more accurate in predicting hospital readmissions for the majority racial group (Caucasian), but less accurate for the minorities (Hispanic, Asian).
Training-serving skew bias	Related to the difference between the model training phase conditions and data distributions versus the deployment phase. This can lead to skewed predictions and inconsistencies in healthcare decision-making for the under-represented groups.	A model can be trained on data from a hospital that serves a predominantly Caucasian population. When the model is deployed in a hospital that serves a more diverse population, it may not be able to accurately predict readmission rates for minority patients, leading to disparities in care.

Table 11 Types of Bias Found in Research (Aldrees et al., 2022) & Linkage to Literature

During our data analysis, these race group distributions were visualized to show the highly imbalanced patient distribution, as shown in the first figure. However, as shown in their readmission rates on the right figure, there is a fairly normal readmission rate for the analogy of each group.

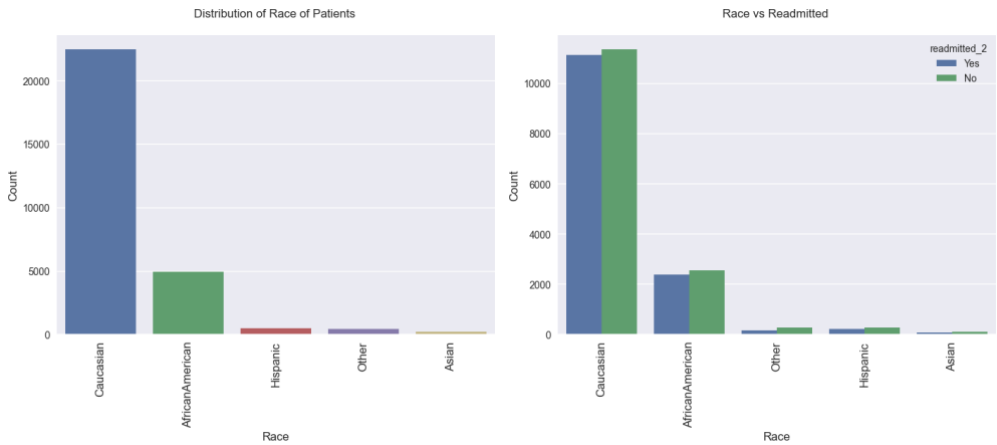

Figure 4 Distribution of Race Groups and their Readmissions.

It is worth investigating whether this high imbalance significantly influences model performance and fairness concerning the under-represented groups.

4.2 Implementation of Different Approaches

Aiming to compare the model performance across the different pre-processing scenarios, the implementation started with the traditional approach widely used in past papers, namely the Grouped Codes approach.

Column	Description
Diag_1	Primary diagnosis in the dataset, denoting the main reason for a patient's hospital admission. The diagnosis codes follow the ICD-9 coding system and represent distinct medical conditions.
Diag_2	Secondary diagnosis column, containing codes for additional medical conditions that may coexist with the primary diagnosis or develop subsequently. These secondary diagnoses can impact treatment and length of stay.
Diag_3	Additional secondary diagnosis column, capturing further medical conditions relevant to the patient's healthcare record.
Categorization of ICD-9 Diagnoses Codes	Due to the extensive number of unique ICD codes and their complexity, they have been categorized into three types: primary, secondary, and additional. These categories are further condensed into 9 disease groups. The resulting groups, mirroring the original dataset publication, include Circulatory, Respiratory, Digestive, Diabetes, Injury, Musculoskeletal, Genitourinary, Neoplasms, and Others.

Table 12 Grouped Codes Pre-Processing Approach

In different Jupyter Notebooks, the new approaches (Top Codes, One Code) were implemented, as described and visualized in the methodology, to be compared to the Grouped approach. For experimentation reasons, different sampling tools were applied in the first model set (oversampling, hybrid). Reweighting and Unawareness bias mitigation techniques were then applied, showcasing different performances for each scenario.

4.3 Model Evaluation and Representativity Across the Approaches

The figures below compare the average AUC Scores across the different scenarios.

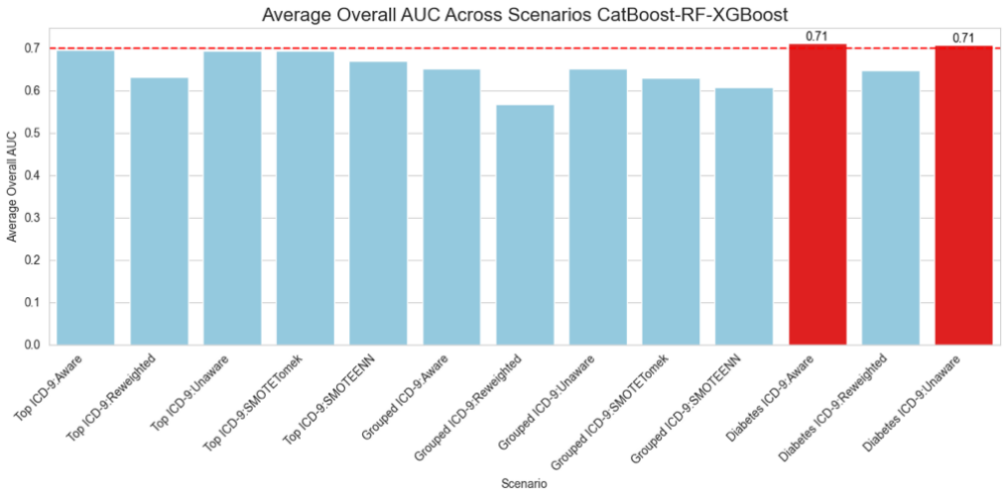

Figure 5 AUC Scores Across the First Set of Classifiers

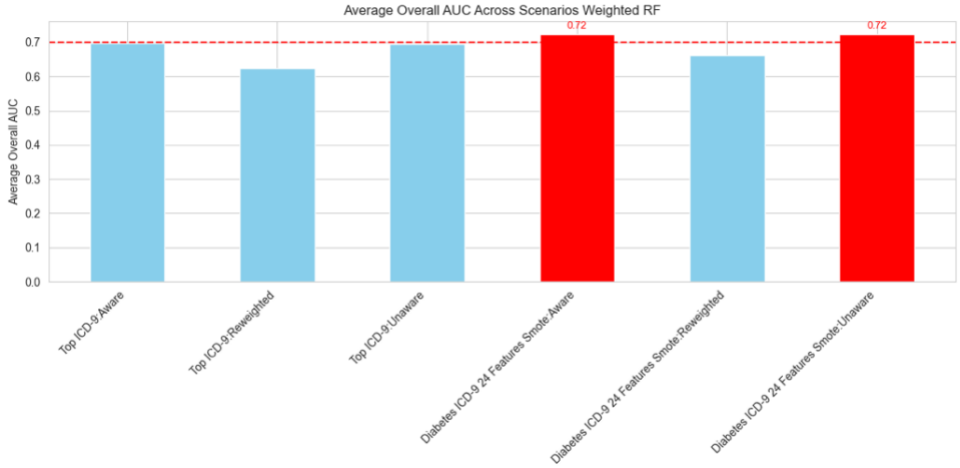

Figure 6 AUC Scores Across the Second Set of Classifiers

Upon initial examination, it is evident that the attained AUC scores surpass the previously shown benchmarks, reaching 0.72, notably exceeding the registered highest AUC score of 0.6736 in Table 10. The improvement can be attributed to different data engineering. However, it is clear in Figure 5 that the implementation of Reweighting, and the Hybrid Sampling tools (loss of information due to undersampling) reduced the AUC scores. Generally, oversampling only was preferred to hybrid tools, and as a result, it was selected to utilize SMOTE in the cases where sampling is needed and to drop further examination of the scenarios including the Hybrid sampling tools, and the Reweighting technique. This is the reason SMOTE was implemented in the One Code Approach, to oversample the few data points remaining. Lastly, in the new pre-processing approaches, implementing the Unawareness Method did not drop the AUC scores significantly, but our focus is to compare the Unaware Grouped Codes with the Aware Top Codes and One Codes approaches. At this point, it is vital to note that a higher model accuracy does not mean fairer results, as more accuracy may translate into even more overshadowing of the under-represented, as the majority groups may represent the most instances correctly predicted.

In comparing the 3 pre-processing approaches, a noteworthy aspect of interest lies in the different Race Group Distributions. There is an observable trend where the majority race group is significantly reduced in the training set of each new approach, compared to the Grouped Codes one. Consequently, the training set undergoes a visible shift towards greater balance among the race groups. This reflects a rise in the visibility of the minority groups within the training set, mitigating their former overshadowing.

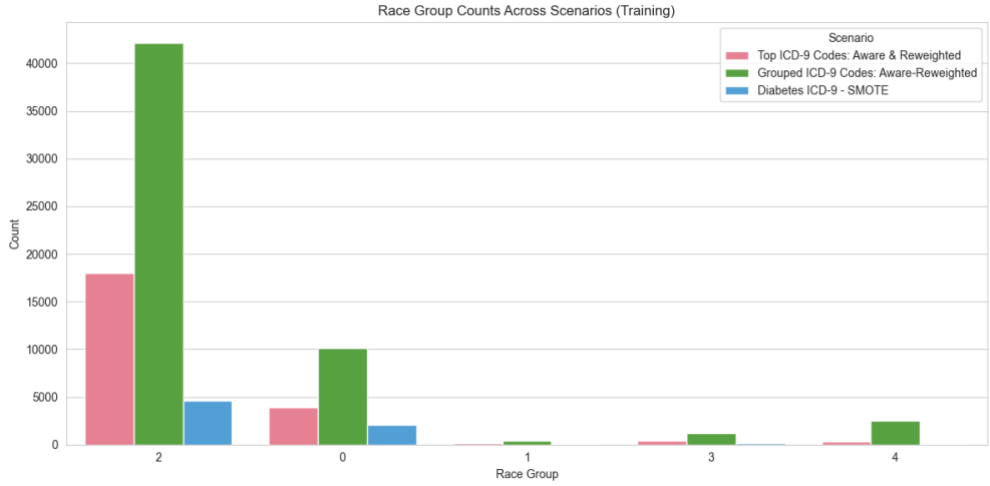

Figure 7 Race Distribution Training Set

To facilitate a clearer visual comparison, logarithmic scales can effectively address the issue when some groups have much larger counts than others, making them hard to distinguish. By employing plt.yscale('log), the y-axis can be transformed into a logarithmic scale (Fig. 8). This enhances the ability to gain a better understanding of the distribution analogy across all groups.

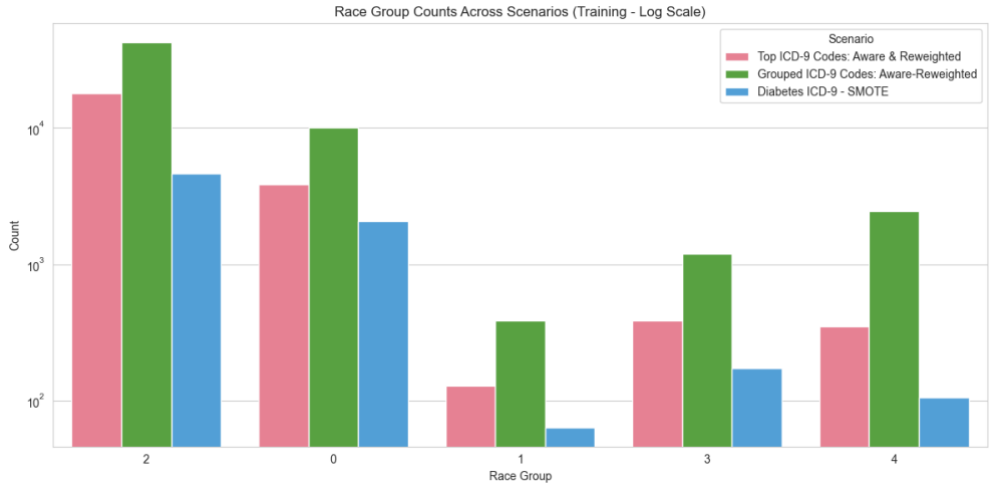

Figure 8 Race Distribution Logarithmic Scale Training Set

Precision and recall are fundamental metrics for evaluating the performance of classification models. As already described in the methodology, precision measures the proportion of correct positive predictions (minimizes the False Positives), while recall measures the proportion of actual positives that are correctly identified by the model (how many patients in need of readmission are not missed by the model, minimizing the False Negatives) The AP Scores are also presented, quantifying the area under the curve, providing an overall perspective on model performance and balance between Precision and Recall.

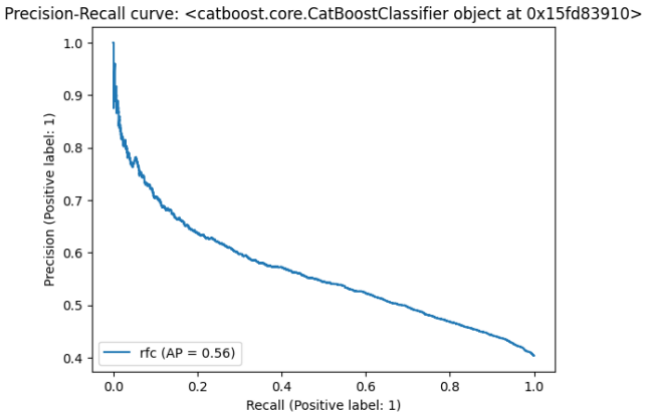

Figure 9 Precision-Recall Curve – Unaware Grouped Codes Scenario

Grouped Codes approach demonstrates suboptimal performance in maintaining precision and recall, illustrating CatBoost's curve. The curve exhibits a steep downward slope, suggesting the model cannot maintain high recall as precision increases. Also, such a low AP Score represents a moderate performance level, and this approach's capability to correctly predict the Readmitted class.

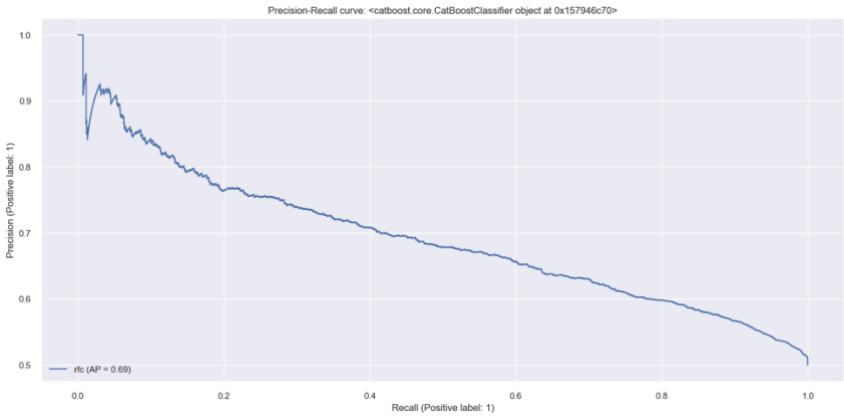

Figure 10 Precision-Recall Curve - Top Codes Scenario

Top Codes approach exhibits a significant improvement in recall compared to the Grouped Codes approach, again utilizing CatBoost. The curve's slope is more gradual, indicating a better trade-off between precision and recall, suggesting this approach is more effective in maintaining high recall without compromising precision. The AP score increased by 0.13.

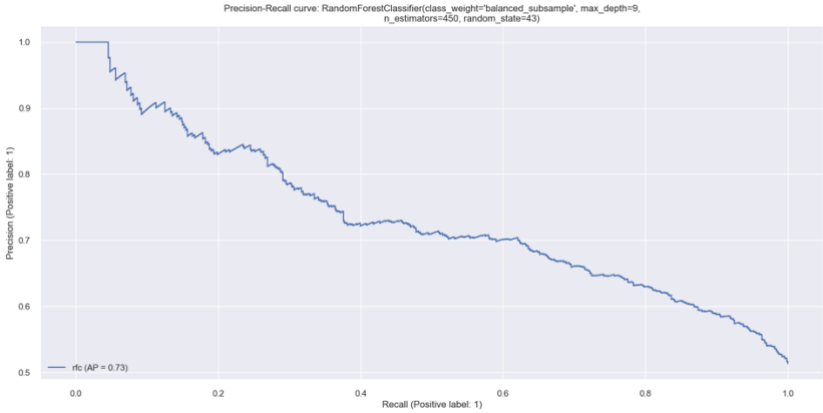

Figure 11 Precision-Recall Curve - One Code Scenario

One Code Approach established the benchmark curve in our comparison. The curve's slope is even more gradual, demonstrating a harmonious precision and recall trade-off. RF was the most optimal model in this approach, maintaining high recall and achieving high precision simultaneously. The AP score is the highest among the scenarios.

Pre-Processing Approach	**AP Score**	**Trade-Off Between Precision Recall**
Grouped Codes	0.56	Suboptimal – Curve declines steeply
Top Codes	0.69	Better – Curve declines gradually
One Code	0.73	Harmonious – Most Gradually declining Curve

Table 13 Precision-Recall Curves Conclusion

The comparative analysis, as per Table 13, highlights the better Precision-Recall trade-off of the new approaches. By maintaining a high recall, without sacrificing precision, and exhibiting a gradual slope in their curves, they identify true positives consistently while avoiding the high increase of false positives.

4.4 Prediction Accuracy for Race Groups Across the Approaches

In comparing model accuracy in correctly predicting the different race groups, for the 3 approaches, it is necessary to assess False Positive, False Negative, True Positive, and True Negative Rates for each race group. The goal is to minimize the False Rates and maximize the True Rates while maintaining consistent scores across the groups. Our focus is on minimizing the FNR metric, focusing on High Recall maintenance. Fig. 12 compares all 3 approaches in all scenarios of Awareness and Unawareness.

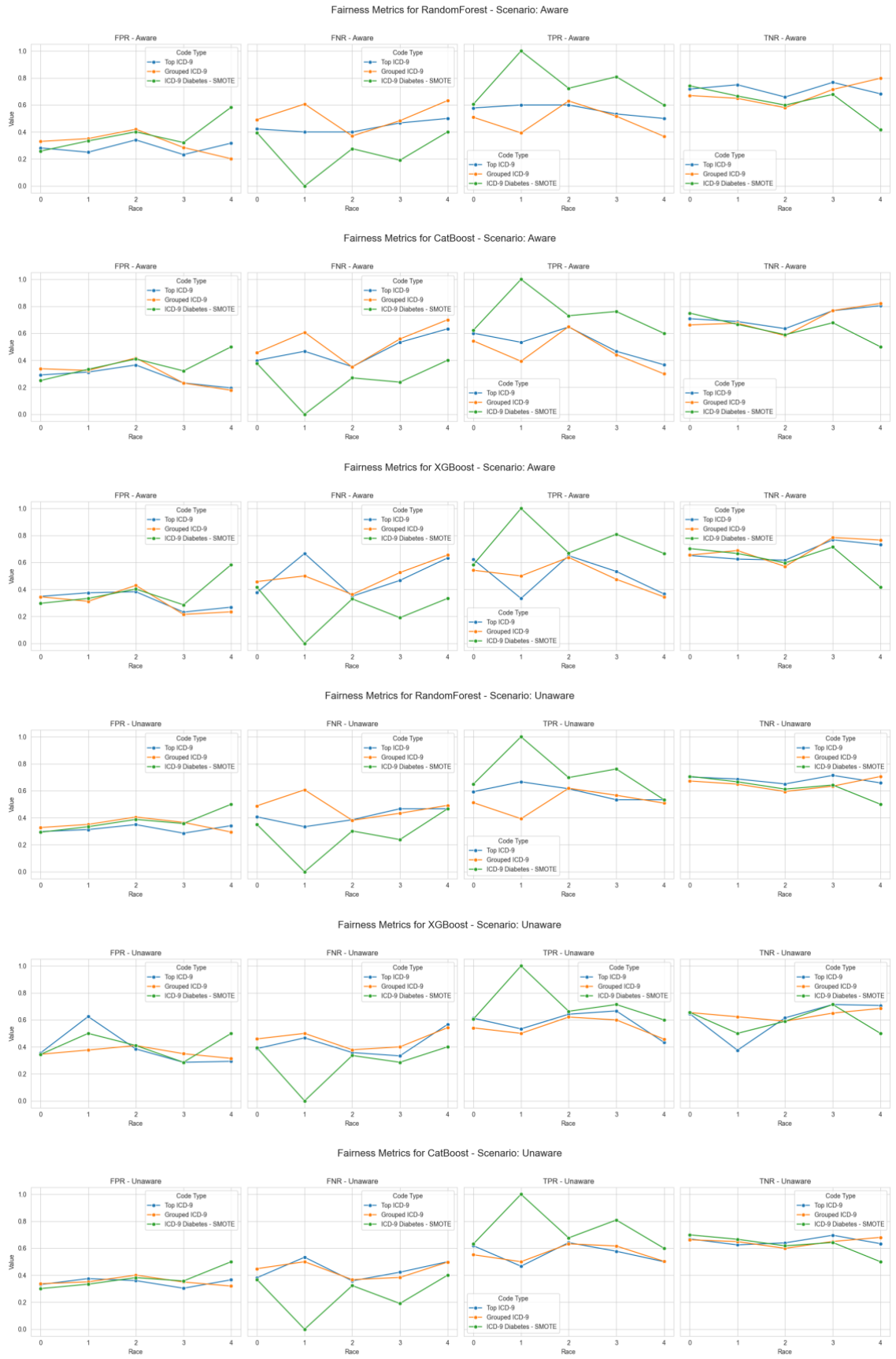

Figure 12 Accuracy Differences of Race Groups

Grouped Codes approach yielded inferior results across the different scenarios and classifiers for the False Negative rates, (except for XGB) indicating a substantial impact of data imbalances on lower model accuracy to correctly predict the minority groups (Races 1,3), while accuracy is much higher for Race 2. The False Positive rates are overall equal to other approaches.

One Code approach achieved the highest accuracies, especially for the True Positive rates (equals 1-False Negative rates), meaning high recall for every group. However, this approach shows an imbalanced accuracy distribution, with a preference for the previously protected groups, leading to new race group imbalances, as now Races 1 and 3 have way higher chances to get accurately predicted for Readmission than any other races. There is a need to keep the accuracies balanced among the subgroups, not only the previously under-represented.

Top Codes approach does not achieve such high accuracies compared to One Code, but it showcases the most balanced distribution of accurately predicting every race group. An exception to that fair accuracy distribution is illustrated in the aware XGBoost, which works better for the 2 other approaches. In most cases, every group has approximately the same chances of being correctly or falsely predicted, demonstrating an equitable manner of minimizing prediction disparities among different ethnic groups.

After illustrating the different scenarios of awareness and unawareness between all the approaches, this paper will now head to answer the second research question, showcasing an overall comparison of the Unaware Grouped Codes approach, against the new Approaches aware of the protected attribute.

4.5 Overall Comparison of Unawareness Method

To answer the second research question, we further visualize overall performance metrics to compare the Unaware Grouped, to the Aware Approaches. Our goal is to keep high AUC and AP scores, while our focus on healthcare fairness prioritizes accuracy in Readmitted predictions. Figure 13 illustrates all metrics for a comprehensive comparison.

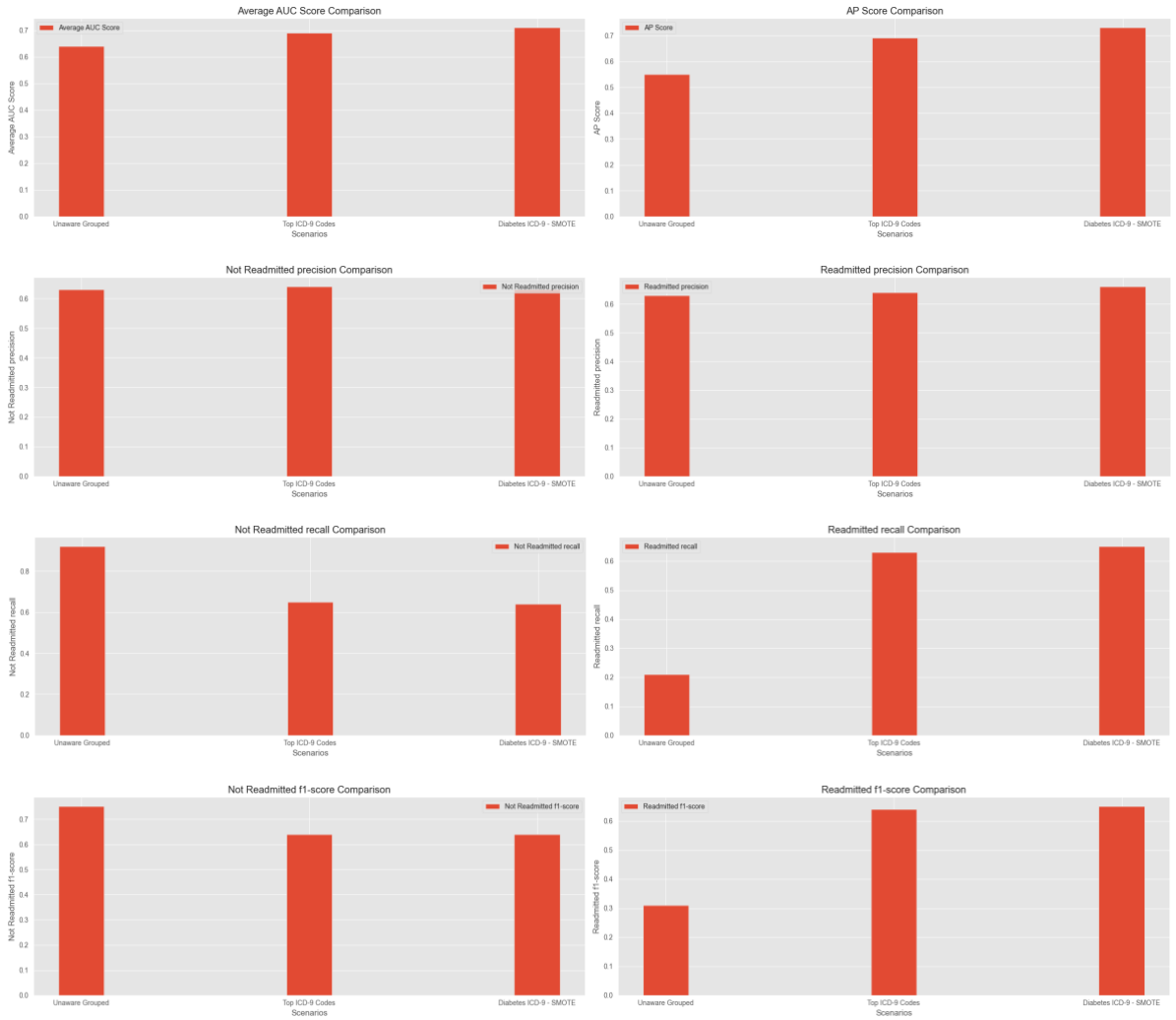

Figure 13 Total Comparison of Unawareness Method

Our comparative analysis showed that the new approaches, aware of the protected attribute, significantly outperformed the Unaware Grouped Codes. They demonstrated notably higher average AUC and AP scores, indicating more precise readmission predictions and precision-recall trade-offs. While the Unaware Grouped Codes performs better in Non-readmitted Recall (meaning it predicts more patients that do not require Readmission), it fell short in all metrics for the Readmitted class. The new approaches excelled in both precision and recall for Readmissions, indicating their ability to correctly predict the patients who are in real need of care. This finding, along with the much higher performance scores (AP, AUC), makes the new approaches more suitable for our focus on achieving both accuracy and fairness.

Next, we investigate the TN and TP Rates for the Unaware Grouped and the Aware approaches.

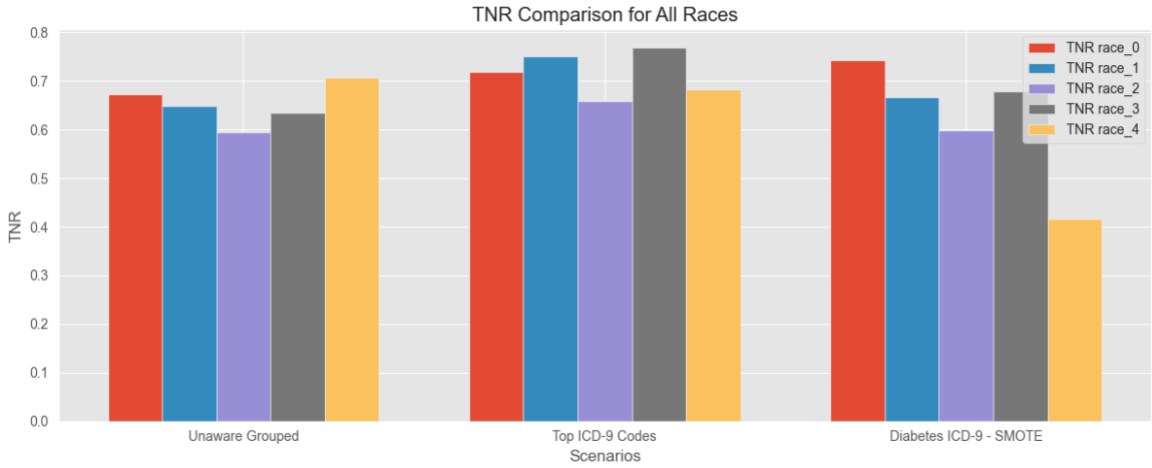

Figure 14 TNR Comparison of Race Groups

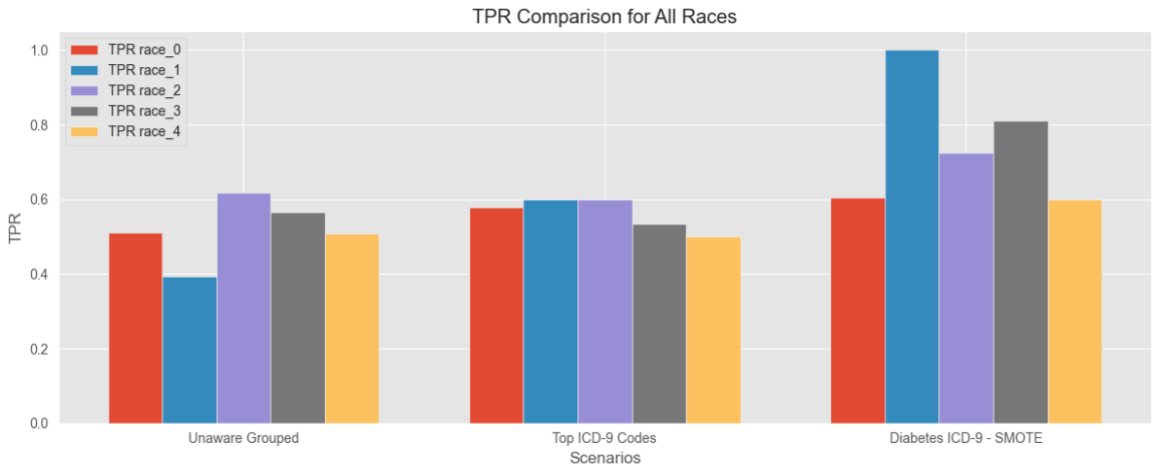

Figure 15 TPR Comparison of Race Groups

The True Negative Rates are higher and almost equal for the Unaware Grouped and the Top Codes. One Codes approach illustrates to fall a bit shorter in this metric, having some imbalances. For the True Positive Rates, the One Code approach achieves the highest metrics, but as already seen, the scores are imbalanced across groups. Top Codes successfully managed to correct the imbalances of the Unaware Grouped approach, indicating the most equitable detection of predicting readmissions across all groups.

Approach	Top AUC Score	Top AP Score	Avg Recall (Readmitted)	Avg Precision (Readmitted)	Avg TPR	Avg TNR
Top Codes	0.69	0.69	0.63	0.64	0.56	0.70
One Code	0.72	0.73	0.65	0.66	0.74	0.62
Unaware Grouped	0.65	0.55	0.21	0.63	0.52	0.64

Table 14 Comparison of Unawareness Method and New Approaches

In conclusion, omitting race as a sensitive attribute during model training proved less effective alone, without changing the data engineering approach, particularly for a highly imbalanced

dataset. Better data pre-processing managed to achieve both bias mitigation and increased model accuracy.

This paper moves on to a fairness exploration of the two new data engineering approaches, with the use of Fairlearn and Dalex libraries, and feature importance investigation with SHAP and ELI5 implementation.

4.6 Fairness Exploration Between the New Approaches

4.6.1 Fairness Exploration with Fairlearn

In this section, we examine the Fairness Metrics Differences, as outlined in Section 3.2.3 of the methodology. A higher bar on these figures signifies a greater disparity between race groups, translating to unfairer, inequitable outcomes. The results from applying the Reweighting technique are also illustrated, to prove that this bias mitigation technique did not achieve fairer results for both approaches.

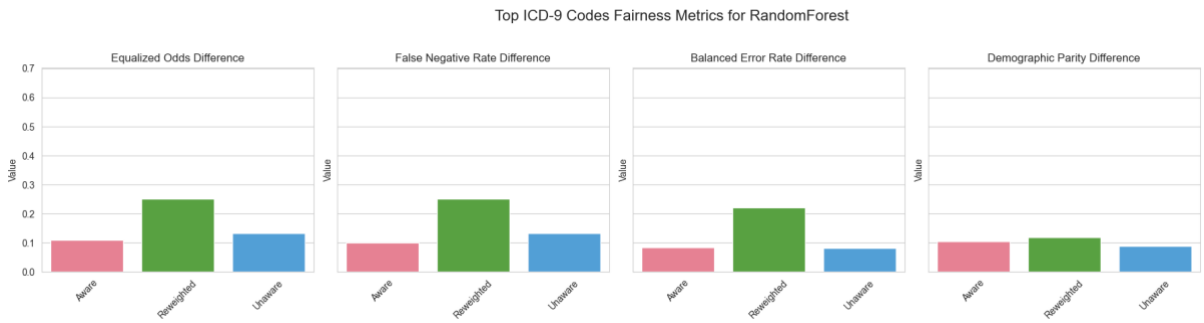

Figure 16 Top Codes Fairness Metrics Differences across Scenarios

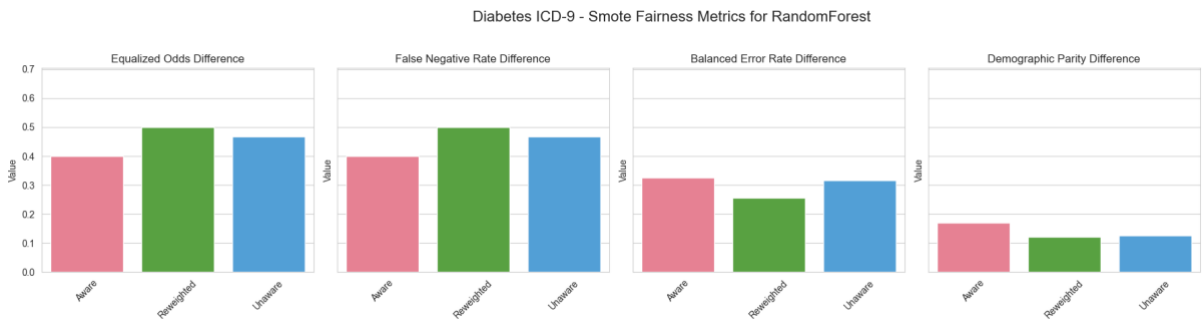

Figure 17 One Code Fairness Metrics Differences across Scenarios

As expected, Top Codes produces much lower fairness score differences across the race groups. This further supports the previous results of this approach presenting fairly balanced accuracies for all race groups, not favoring or disfavoring specific races.

One Code approach, as was before illustrated to having imbalanced accuracies, boosting specific groups, now is revealed to exhibit high fairness difference rates, indicating high disparities among groups.

The findings here highlight the importance of exploring the fairness of every approach and that more predictive accuracy does not necessarily imply more fairness. One Code approach had better AP Scores and AUC, and higher overall True Positive Rates, but proved to have more fairness issues compared to Top Codes. We aim to strike a balance between overall accuracy and fairness across all demographic patients. Our next step is to showcase the results of Dalex implementation for further fairness exploration, local explanations using SHAP, and global using ELI5.

4.6.2 Fairness Exploration with Dalex & Feature Importance with xAI

The Dalex library is employed to further explore fairness concerning the racial groups within the approaches. During the library's implementation, it was required to assign a group as the 'protected', choosing the group labeled as '1' (Asian), and a 'privileged' group, where the group labeled as '2' (Caucasian) was chosen. Then, during the creation of a fairness object, having assigned the protected and privileged groups, along with the rest as 'Others', the default epsilon value of 0.8 was tuned. This value operates within the framework of the four-fifths rule, serving as a benchmark for detecting discrimination against the protected group, as per (EEOC, 1978) The epsilon parameter provides a threshold, where ratios outside the range may indicate biases in model predictions. If the threshold is surpassed, this metric indicates a fairness violation.

4.6.2.1 Top Codes Approach

4.6.2.1.1 Implementation of Dalex Library

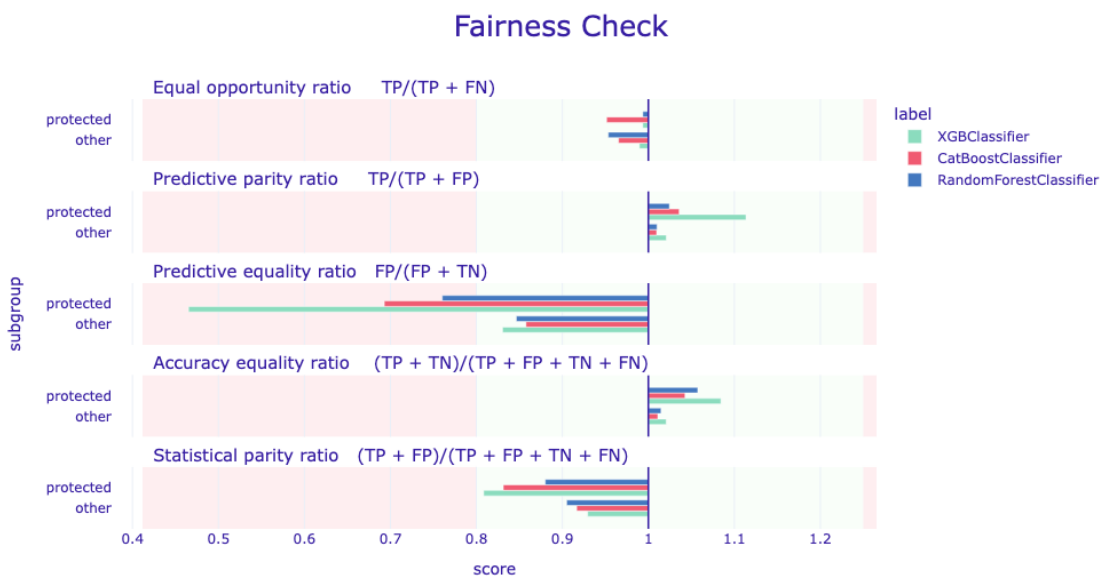

Figure 18 Fairness Check for First Approach

The choice of fairness metric plays a crucial role in each case's fairness evaluation. For the specific focus on balanced True Positives (equal Recall) among groups, the Equal Opportunity ratio serves as the most suitable choice. To achieve balanced results, the other 2 suitable metrics are the PER and SPR, translating into balance in false positives across groups (in Precision), and in predicting positive outcomes (equal Readmissions for every group).

Noteworthily, in this approach, all classifiers show only one violation of fairness for the PER metric. Our focus on keeping an Equal Opportunity Ratio for all, and having the same Readmissions Rates is highly achieved. RF emerges as the fairest model, having almost all disparity lines near the vertical line (representing the Privileged group as a baseline), for both the protected and the other groups.

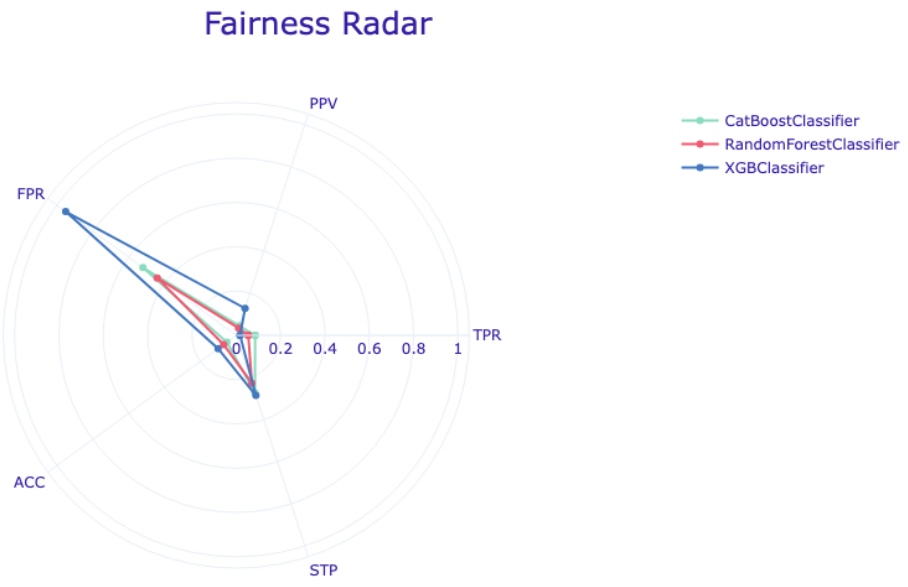

Figure 19 Fairness Radar for First Approach

The Fairness Radar plot (Fig.19) visualized the parity loss metrics across the different models. Parity loss signifies the deviation from a state of fairness. A value of zero in our case indicates no bias or disparity in model predictions across the groups. The FPR metric indicates the likelihood of false positive predictions (meaning precision). If the classifier lines extend further away from the center along this axis it suggests a higher parity loss for this metric, implying greater bias.

XGBoost's line extends the furthest along the FPR. This implies that XGBoost is more prone to generating unequal False Positive predictions, potentially leading to unfair treatment of the protected group. As expected, RF's line falls close to the center along all axes.

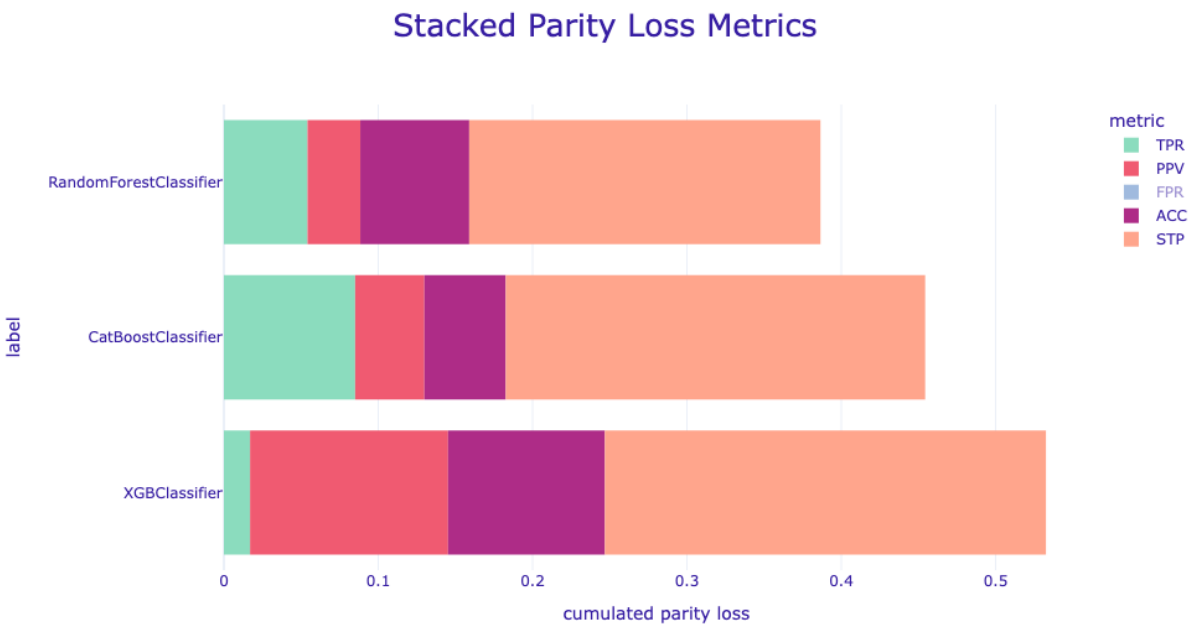

Figure 20 Stacked Cumulated Parity Loss for First Approach

The term 'stacked' in Fig.20 translates into the graphical representation of multiple fairness metrics within a single bar for each classifier. In our stacked bar chart, various segments are stacked to illustrate different metrics contributing to the total value. Each segment corresponds to the degree of parity loss for the specific metric, meaning more bias. The complete bar provides an overall view of the cumulative parity loss for each model. The smallest bar translated to the least bias, as shown in RF in our case.

4.6.2.1.2 Implementation of SHAP Tool

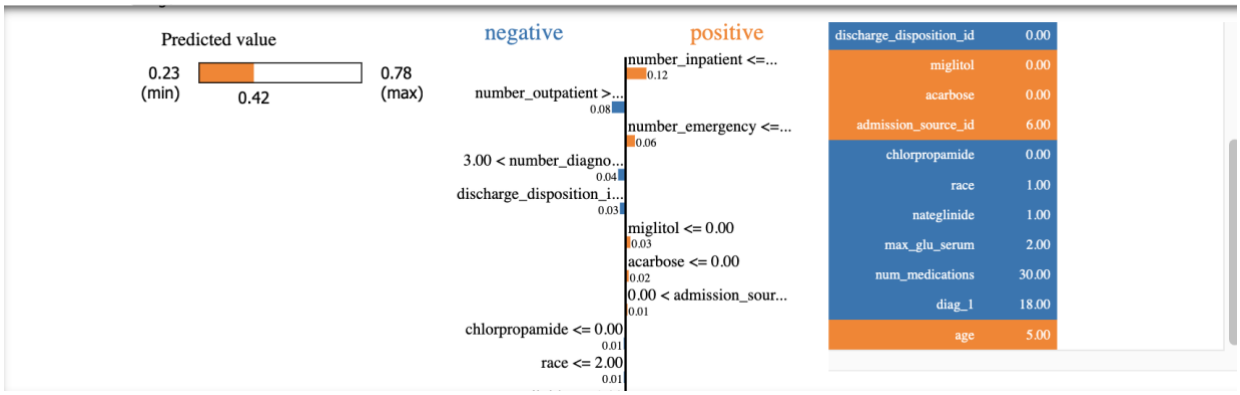

Figure 21 SHAP Implementation in RF for First Approach

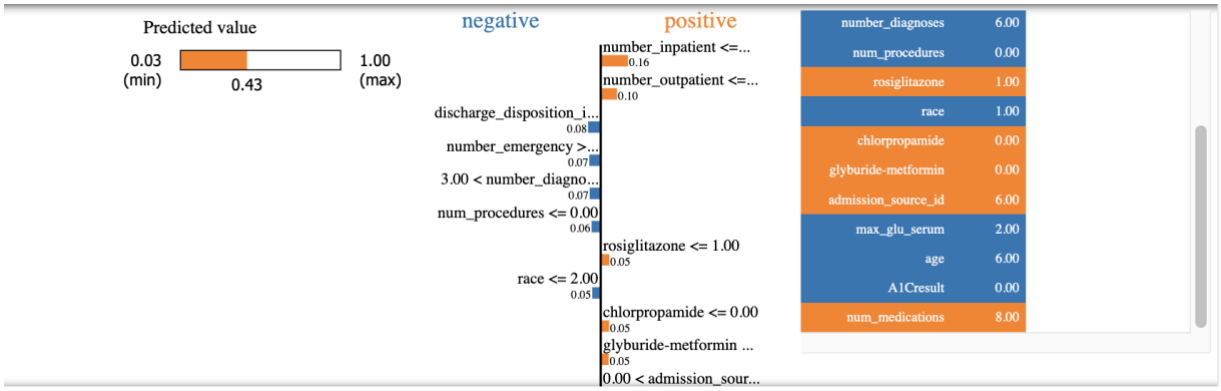

Figure 22 SHAP Implementation in XGBoost for First Approach

By analyzing the SHAP values, we can gain insights into how the model makes its decisions, meaning how important is every feature for the final decision, a patient to be readmitted. In the context of analyzing the race attribute, we are interested in understanding how much it influences this decision. To compare the two models (XGBoost and RF) in terms of their reliance on the race feature, we can consider the following factors:

- *Position and Color of Race Feature in SHAP plot*

  If the race feature is closer to the left and in blue, it suggests that it decreases the readmission chance, potentially meaning it is an important factor leading to a Not Readmission decision, as happens in XGBoost with a value of 0.05.

- *Absolute Value of SHAP Score for Race Feature*

  The higher absolute value in XGBoost indicates a stronger influence of the feature in the model's decision. (RF has a score of 0.01)

This implies that for Race 1 (the Asian group), in two local predictions, the XGBoost "punished" more the patient because of the specific race attribute, compared to RF.

However, it is important to note that SHAP values do not necessarily imply causation and high values for race do not confirm that the model is inherently biased. They simply provide insights into how the model is using this feature to make predictions.

4.6.2.1.3 Implementation of ELI5 Library

In Figure 23, we illustrate the results obtained from employing the ELI5 library to interpret the RF model, revealing the assigned weights to individual features for the decision of readmissions. It is essential to note the outcomes provide a global explanation, giving the

overall feature importance across the entire outputs, unlike the local explanations given by SHAP.

Weight	Feature
0.2137 ± 0.1465	number_inpatient
0.0917 ± 0.0759	discharge_disposition_id
0.0613 ± 0.0385	num_medications
0.0612 ± 0.0641	number_outpatient
0.0598 ± 0.0662	number_diagnoses
0.0554 ± 0.0470	diag_3
0.0515 ± 0.0563	admission_source_id
0.0475 ± 0.0344	diag_1
0.0471 ± 0.0232	num_lab_procedures
0.0460 ± 0.0319	diag_2
0.0401 ± 0.0556	number_emergency
0.0328 ± 0.0302	num_procedures
0.0319 ± 0.0196	age
0.0314 ± 0.0209	time_in_hospital
0.0202 ± 0.0168	admission_type_id
0.0147 ± 0.0125	insulin
0.0122 ± 0.0150	diabetesMed
0.0118 ± 0.0100	race
0.0111 ± 0.0099	A1Cresult
0.0076 ± 0.0088	max_glu_serum
... *14 more* ...	

Figure 23 ELI5 Implementation for First Approach

The relatively low score and its deviation together imply that Race is not a dominant factor in the model's decision, being lower than non-sensitive attributes like insulin, time in hospital, medications, and diagnoses, ranking the 18th place of importance. However, its presence in this rank warrants careful consideration, especially in the healthcare domain, attributing so much for a readmission decision. If the ML application was image-based, for example detecting skin cancer, the substantial impact of skin color on the model's decision would seem more plausible.

4.6.2.2 One Code Approach

4.6.2.2.1 Implementation of Dalex Library

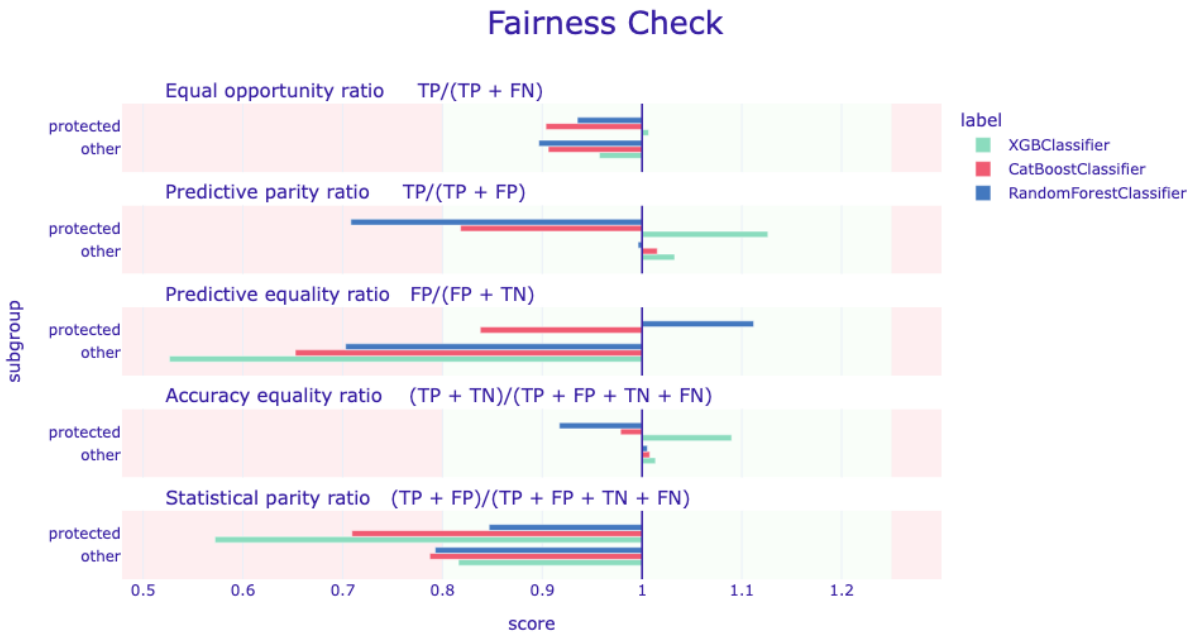

Figure 24 Fairness Check for Second Approach

In this approach, it is evident that for all models, 3 metrics are being violated in their fairness issues instead of one, as illustrated in Fig. 24. For the Equal Opportunity ratio, all models are fair, not surpassing the threshold. This implies that the classifiers are equally identifying the True Positive rates for all groups. Predictive Equality and Statistical Parity ratios, which are in our scope of investigation, show big fairness violations, especially from the XGB model.

RF surpasses the threshold the least when considering both the protected and the other groups. However, it cannot be called fairer, given its additional violation in the Predictive Parity ratio, given the compliance of the other models with this metric. Overall, the Fairness Check figure does not favor any specific classifier being significantly fairer. Further plots need to be analyzed.

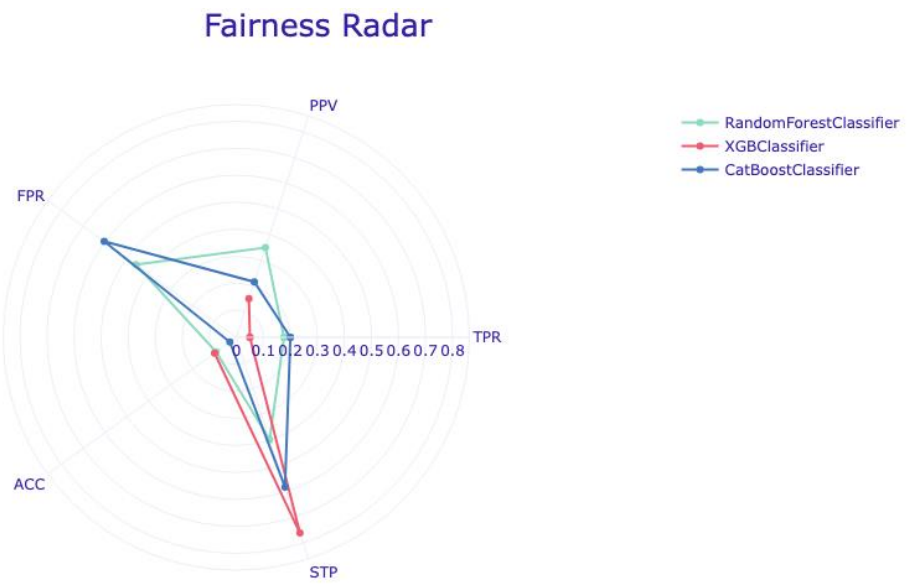

Figure 25 Fairness Radar for Second Approach

In the Radar plot, a first glance is that XGBoost implies heavy bias in only one metric, STP. The other classifiers show a fairly distributed bias between STP and FPR (False Positive Rates). RF is positioned relatively centrally in the plot, suggesting a more balanced performance across the metrics. However, the choice will be again based on the prioritization of certain notions of fairness.

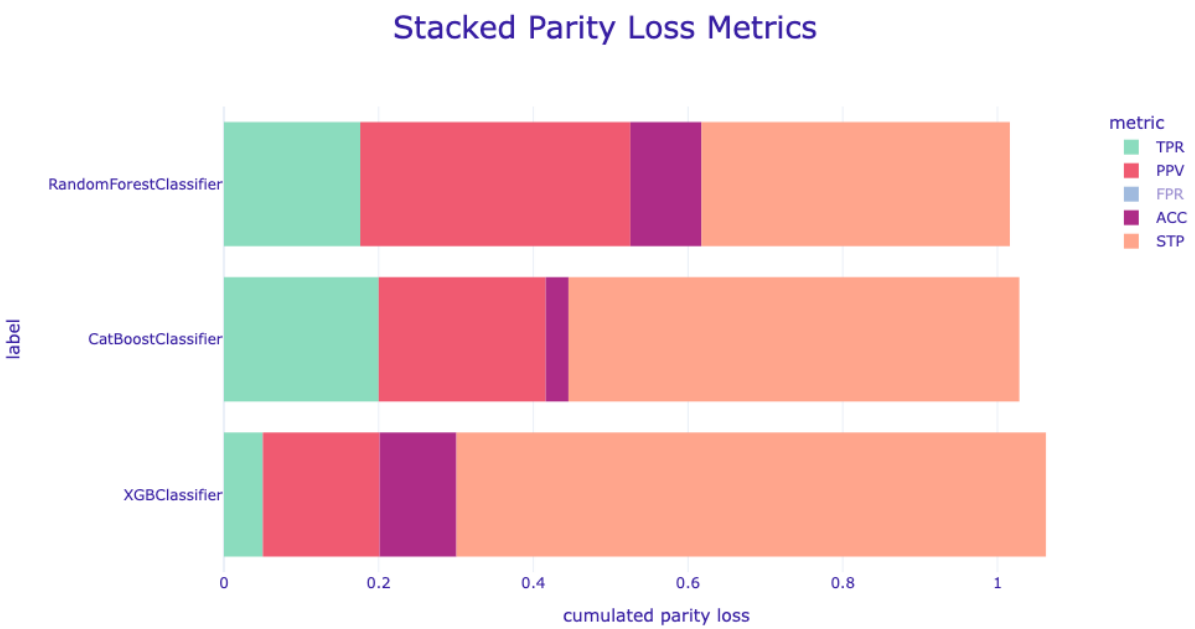

Figure 26 Stacked Cumulated Parity Loss for Second Approach

In this chart, XGBoost exhibits the least parity loss in True Positive rates. This is desirable as it reflects fewer discrepancies in correctly identifying true positives across groups. However,

the model presents a significant parity loss in STP, indicating big differences in the readmission rates among the groups. This could lead to unfair advantages or disadvantages for certain races, potentially leading to healthcare inequalities.

The assessment of the fairer classifier depends on the prioritization of the fairness metrics. If the primary objective is to minimize the disparity in true positive predictions, XGBoost would be this approach's choice. Conversely, if ensuring uniform positive outcomes is predicted across all groups, meaning each group has approximately the same readmission rates, attention should be directed towards reducing the Statistical Parity's parity loss, and RF would make a less biased choice. The decision regarding the fairer classifier lies in the fairness goals established for the model's application.

4.6.2.2.2 Implementation of SHAP Tool

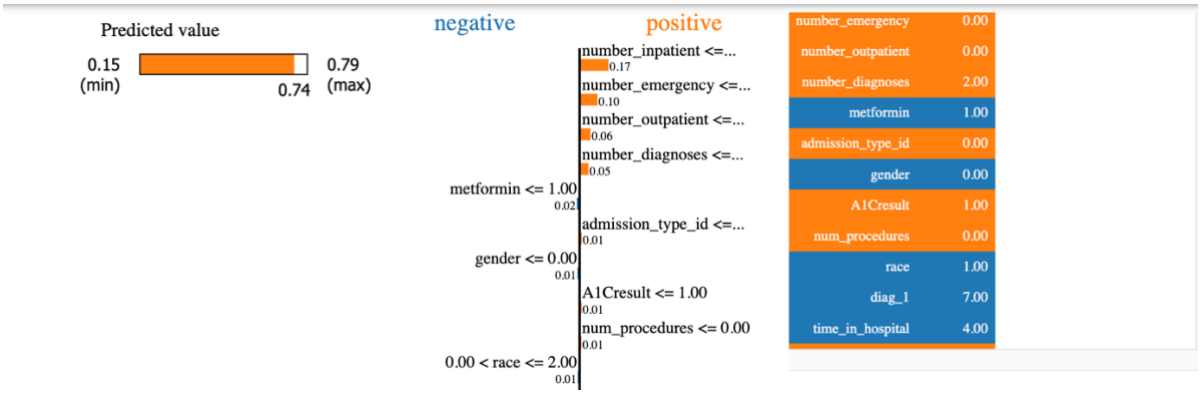

Figure 27 SHAP Implementation in RF in Second Approach

In the context of the presented local explanation, the race feature is assigned a minor importance score of 0,01 for the Asian Race, negatively impacting the readmission. This score is the same as in the Top Codes approach. We will now check the XGBoost's SHAP value for another local instance of Race=2 (Caucasian Race), as the high Statistical Parity (STP) metric implies way higher counts of readmission rates for this privileged group, compared to the others.

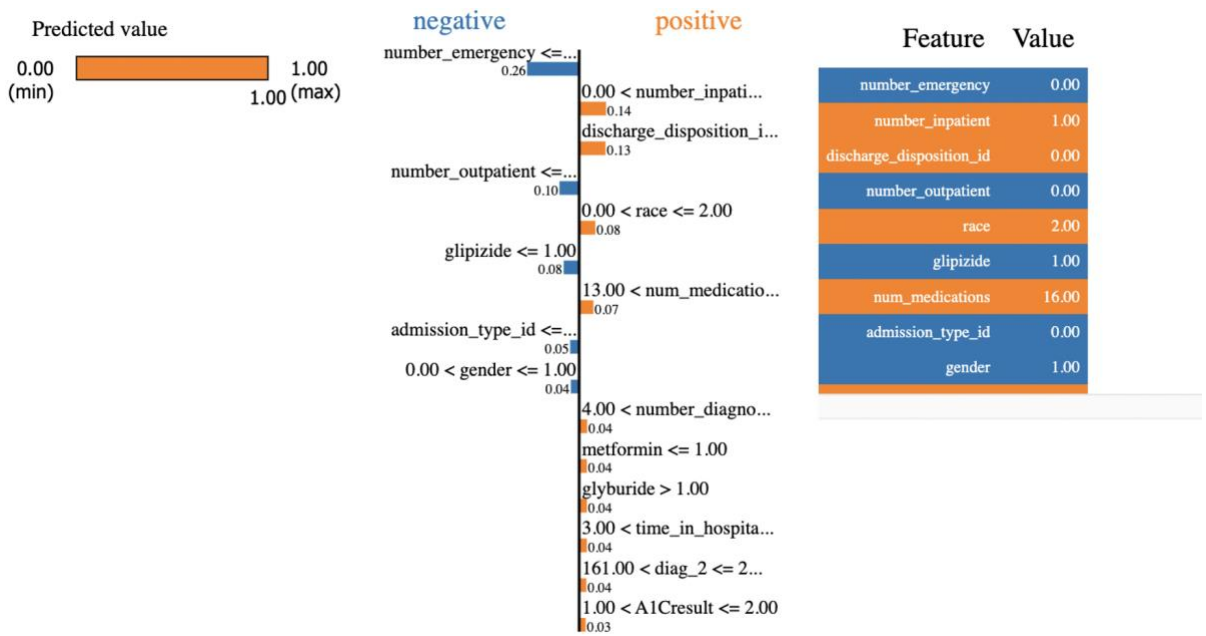

Figure 28 SHAP Implementation in XGB in Second Approach

It is observed in Fig.28 that for a Caucasian individual with a positive readmission decision, the XGBoost attributed a significantly high importance score of 0.08. Even though SHAP cannot be directly linked with bias implications, we can assume that such a high score of importance leading to positive readmission of the patient is aligned with the historical readmission decisions contained within the data, indicative of pre-existing potential disparities. If this is the case, it means that bias comes from potentially unfair hospitals' decision-making (the data itself), and not during XGBoost's training, and that higher readmission rates for the Caucasian rates are shown in the dataset, so the model is accurately trained. A high STP score proving fairness violations, along with the high race importance of a positive prediction, when together, can indicate potential bias, while the SHAP value alone cannot directly be linked to the existence of bias, and only acts as an assisting tool.

4.6.2.2.3 Implementation of ELI5 Library

Weight	Feature
0.2080 ± 0.1367	number_inpatient
0.0759 ± 0.0560	num_medications
0.0716 ± 0.0767	number_diagnoses
0.0676 ± 0.0474	diag_1
0.0649 ± 0.0744	number_emergency
0.0597 ± 0.0386	diag_3
0.0553 ± 0.0293	num_lab_procedures
0.0514 ± 0.0322	diag_2
0.0498 ± 0.0351	age
0.0439 ± 0.0541	number_outpatient
0.0375 ± 0.0278	time_in_hospital
0.0360 ± 0.0294	discharge_disposition_id
0.0257 ± 0.0202	admission_type_id
0.0229 ± 0.0244	num_procedures
0.0213 ± 0.0230	A1Cresult
0.0185 ± 0.0154	insulin
0.0178 ± 0.0170	admission_source_id
0.0169 ± 0.0147	race
0.0142 ± 0.0162	metformin
0.0126 ± 0.0128	gender
	... 4 more ...

Figure 29 ELI5 Implementation for Second Approach

Like before, Race does not heavily influence the model's predictions according to this output, but it remains higher than some non-sensitive attributes. For example, its larger impact on the decisions compared to the drug Metformin should be alarming. Race attribute here takes again the 18th place, but time has a higher weight score and deviation. Again, race's importance, in a global explanation for readmissions, being higher than certain medications should raise awareness and consideration.

4.6.2.3 Conclusion

In comparing the two approaches, the implementation of Dalex revealed the Top Codes approach to exhibit fewer potential fairness violations, indicating lower bias. The SHAP (local explanations) and ELI5 (global explanations) findings further supported the Top Codes method by revealing a smaller impact of the Race feature in influencing predictions for both local scale (random instances) and global. It is important to discuss that while the One Code approach showed higher promise during the model performance findings, with the assistance of Fairness libraries and xAI tools, this research was led to adopt the Top Codes approach, for achieving more equitable, balanced, and fair predictive results.

5. DISCUSSION

As we embark on our discussion, it is important to recognize the incorporation of preliminary discussions within the presentation of findings in Section 4. This approach was chosen to

provide immediate context and insights into each discovery, as the initial level of discussion is more contextually relevant to the specific results being presented. In section 5, there is a zoom-out from the individual results, offering a broad reflection.

5.1 Model Performance and Parameters

For the reproducibility of the empirical study, the code for this project has been made available under https://github.com/NickolasB98/TOM_Thesis.

5.1.1 Variation due to random seeds.

The observed variations in model performance and fairness metrics arise from the inherent randomness introduced by specific random seeds in model parameters and data sampling during training and testing. While specific random seeds were intentionally selected for results reproducibility, the potential to alter these seeds exists in principle.

5.1.2 Parameters.

The initial model configurations were maintained with default settings, as the primary aim was not to fine-tune hyperparameters but to evaluate the baseline model performance. An exception to this approach was made in the second series of models where sample weights were adjusted to address class imbalances.

5.2 Addressing the Research Questions

Our investigation provided compelling responses to the main RQ. The study underscored the significance of a comprehensive domain understanding that integrates our dataset, coupled with the implementation of tailored preprocessing approaches. These findings highlight the need for ML developers to incorporate domain-specific knowledge and insights for each case. Specifically, the imbalances within our dataset, particularly concerning diverse racial groups, emerged as factors influencing both model fairness and performance. The gain of domain insights led to data engineering approaches, linking the well-handling of ICD-9 codes, as highly effective in increasing model performance, and mitigating the data imbalances, to a point of achieving fairer predictions. Leaving such biases unaddressed could have serious consequences. For example, if a model systemically underpredicts the risk of hospital readmission for minority groups, these patients may not receive the follow-up care or interventions they require, potentially leading to poorer health outcomes. Thus, addressing biases extends beyond technical accuracy; it is a fundamental ethical and clinical responsibility.

Regarding the second RQ, our study offered a comparison of the Unawareness method as implemented in the initial grouping ICD-9 Codes strategy, with more tailored, aware of the

protected attribute, preprocessing approaches. This comparison implies substantial implications for the trustworthy use of awareness. Not only did the new engineering approaches avoid the loss of biological patient information, but also achieved an enhanced balance between fairness and accuracy. This suggests that incorporating awareness of sensitive attributes with consideration to fairness contributes to more reliable and effective ML predictions in healthcare.

5.3 Limitations and Further Extensions

Enhanced Data Collection:

The dataset utilized for this study poses certain limitations that need to be acknowledged. Specifically, it lacks comprehensive data on glucose variability, which, as identified in existing literature (Hsieh, 2019; Spanakis et al., 2020), is a significant predictor of the target label of readmissions. The only traceable data concerning glucose is indirectly derived from modifications in medication prescriptions, which is an insufficient marker of glucose stability or variability. Future research should focus on obtaining more detailed patient data, including information on glucose variability. A more informative dataset can provide a deeper understanding of patient attributes and contribute to improved model accuracy.

Hyperparameter Optimization and Ensemble Models:

While this study primarily focused on baseline model configurations, future investigations could delve into hyperparameter optimization to fine-tune model performance, with a focus on fairness. However, to achieve higher accuracies utilizing this dataset, more comprehensive data quality should be prioritized.

Proxies & Limitations of the Unawareness Method's Efficiency:

The unawareness method faces a critical limitation when dealing with proxy variables, as shown in (Pessach & Shmueli, 2022). Proxy variables are correlated features with the sensitive attributes, and the unawareness method may overlook these correlations, if the proxies are also not omitted, bias will not be mitigated. During the dataset's analysis, the correlation matrix did not show direct proxies to race. However, there is a need for debiasing techniques that can effectively find and omit potentially less noticeable proxies of sensitive attributes, like the weighted estimator as proposed by (J. Chen et al., 2019).

Focus on Other Bias Mitigation Phases:

This paper concentrated on the pre-processing phase of bias mitigation, like in the case of (Aldrees et al., 2022). Building on this foundation, future researchers are encouraged to

broaden their exploration to include in-processing and post-processing bias mitigation techniques. Examples are shown in Table 2.

Transparency, Interpretability, and Link of xAI with Bias:

A limitation of Dalex was its inability to build comparison plots between the RF models with the different sample weights, as the library asks for totally distinctive models as parameters.

A limitation of SHAP is that it relies solely on individual feature importance. A feature may not have a large individual SHAP value, but it may have other significance when in combination with other features. Further xAI tools (e.g. ANCHORS) can be explored, that study the significance of feature combinations (Ribeiro et al., 2018). Lastly, as highlighted in the results, there is no direct link between xAI tools insights and the presence of bias. The next step would be to further explore if there is an association between these tools and the confirmation of whether a model can be proved as biased.

5.4 Broader Implications

Our research findings have several broader implications for the development of equitable ML models in clinical settings. Firstly, by closely examining how sensitive attributes are distributed within the dataset and considering the specific domain where the ML application is employed, researchers can develop effective strategies to find and address unintentional bias, either with data analysis and correlations or with the assistance of ML algorithms.

Secondly, our findings underscore the importance of effectively handling imbalanced data, a common challenge in healthcare datasets. Our study demonstrated that balancing the data can significantly impact model performance and fairness, linking this to (Clemmensen & Kjærsgaard, 2022), who stress the potential of balanced datasets in promoting equitable healthcare predictions. By employing pre-processing techniques prioritizing information-rich data, omitting data that cause the overshadowing effect on under-represented groups, or with the utilization of sampling tools, researchers can effectively fix imbalances, thereby enhancing the overall generalizability of their models.

Thirdly, our study highlights the limitations of Fairness through Unawareness. Our results showed that this method is not preferred in bias mitigation, leading to information loss. Instead, we advocate for gaining the know-how of optimal use of biological features like race and skin color especially in fields like healthcare. Consequently, our findings aligned with the emerging perspective in literature as per (Bogen et al., 2020), and (R. J. Chen et al., 2021) advocating for a more informed approach to handling sensitive attributes, in cases where biological data is crucial.

Additionally, we found the implementation of xAI tools to bear significant relevance in our decisions, also linking to the findings of (Mathews, 2019; Pawar et al., 2020). These tools provided the necessary transparency to interpret model decisions and visualize insights for our protected attribute. xAI tools in healthcare can be very significant for giving insights to healthcare employees, or for informing patients about the parameters leading to ML's decisions.

Lastly, the deployment of ML applications in clinical settings raises ethical concerns, regarding the potential for discrimination led by unmitigated bias, linking to the findings of (I. Y. Chen et al., 2021) It is important to establish transparent and accountable processes for the development, evaluation, and deployment of ML in healthcare. Patients should be informed about the use of ML algorithms in their care and provided with opportunities to give feedback. Ongoing monitoring and evaluation are essential to ensure that ML algorithms continue to perform as expected and do not perpetuate or exacerbate existing biases.

6. CONCLUSION

This research investigated the issue of fairness in machine learning models for healthcare predictive modeling. As machine learning becomes more widely used in healthcare, it is essential to ensure that predictive models are fair in their output. This study provides valuable insights and contributions to this emerging field.

First, the study demonstrated the importance of data preprocessing in promoting fairness and model accuracy in healthcare predictive classifiers. By carefully examining the dataset, and gaining proper domain knowledge, we were led to employ tailored data engineering, and disparities in model treatment across diverse patient groups were addressed.

Secondly, the study advocates for the conscientious utilization of sensitive attributes within a well-structured preprocessing framework, highlighting the need for a balanced approach to leveraging such attributes, and avoiding their omissions.

In conclusion, this study contributes to the growing body of knowledge on fairness in healthcare predictive modeling and emphasizes its broader implications. The paper advocates for the responsible and equitable use of machine learning in healthcare, emphasizing the need for comprehensive data exploration, effective imbalanced data management, and advanced data engineering.

References

Adadi, A., & Berrada, M. (2020). *Explainable AI for Healthcare: From Black Box to Interpretable Models*. https://doi.org/https://doi.org/10.1007/978-981-15-0947-6_31

Ahmed, M., & Kabir, M. M. J. (2019). Dimensionality reduction in predicting hospital readmissions of diabetics. *2019 5th International Conference on Advances in Electrical Engineering, ICAEE 2019*, 155–159. https://doi.org/10.1109/ICAEE48663.2019.8975432

Aldrees, A., Poland, C., & Irshad, S. A. (2022). Auditing Algorithms: Determining Ethical Parameters of Algorithmic Decision-Making Systems in Healthcare. *Communications in Computer and Information Science*, *1577 CCIS*, 296–309. https://doi.org/10.1007/978-3-031-04447-2_20

Allen, R., & Masters, D. (2020). Artificial Intelligence: the right to protection from discrimination caused by algorithms, machine learning and automated decision-making. *ERA Forum*, *20*(4), 585–598. https://doi.org/10.1007/s12027-019-00582-w

Alloghani, M., Aljaaf, A., Hussain, A., Baker, T., Mustafina, J., Al-Jumeily, D., & Khalaf, M. (2019). Implementation of machine learning algorithms to create diabetic patient re-admission profiles. *BMC Medical Informatics and Decision Making*, *19*. https://doi.org/10.1186/s12911-019-0990-x

Aspinall, P. J., & Jacobson, B. (2007). Why poor quality of ethnicity data should not preclude its use for identifying disparities in health and healthcare. *Quality and Safety in Health Care*, *16*(3), 176–180. https://doi.org/10.1136/qshc.2006.019059

Barton, M., Hamza, M., & Guevel, B. (2023). Racial Equity in Healthcare Machine Learning: Illustrating Bias in Models With Minimal Bias Mitigation. *Cureus*. https://doi.org/10.7759/cureus.35037

Baumann, E., & Rumberger, J. L. (2018). *State of the Art in Fair ML: From Moral Philosophy and Legislation to Fair Classifiers*. http://arxiv.org/abs/1811.09539

Bogen, M., Rieke, A., & Ahmed, S. (2020). Awareness in practice: Tensions in access to sensitive attribute data for antidiscrimination. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 492–500. https://doi.org/10.1145/3351095.3372877

Caton, S., & Haas, C. (2020). Fairness in Machine Learning: A Survey. *ArXiv*. https://doi.org/https://doi.org/10.48550/arXiv.2010.04053

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. In *Journal of Artificial Intelligence Research* (Vol. 16). https://doi.org/https://doi.org/10.48550/arXiv.1106.1813

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical Machine Learning in Healthcare. *Annual Review of Biomedical Data Science Annu. Rev. Biomed. Data Sci*, *4*, 123–144. https://doi.org/10.1146/annurev-biodatasci-092820

Chen, J., Kallus, N., Mao, X., Svacha, G., & Udell, M. (2019). Fairness under unawareness: Assessing disparity when protected class is unobserved. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 339–348. https://doi.org/10.1145/3287560.3287594

Chen, R. J., Chen, T. Y., Lipkova, J., Wang, J. J., Williamson, D. F. K., Lu, M. Y., Sahai, S., & Mahmood, F. (2021). *Algorithm Fairness in AI for Medicine and Healthcare*. http://arxiv.org/abs/2110.00603

Clemmensen, L. H., & Kjærsgaard, R. D. (2022). Data Representativity for Machine Learning and AI Systems. *ArXiv*, *2203.04706v2*. http://arxiv.org/abs/2203.04706

Deho, O. B., Joksimovic, S., Li, J., Zhan, C., Liu, J., & Liu, L. (2023). Should Learning Analytics Models Include Sensitive Attributes? Explaining the Why. *IEEE Transactions on Learning Technologies*, *16*(4), 560–572. https://doi.org/10.1109/TLT.2022.3226474

Di Camillo, B., Nicosia, G., Buffa, F., & Lo, B. (2020). Guest editorial data science in smart healthcare: Challenges and opportunities. In *IEEE Journal of Biomedical and Health Informatics* (Vol. 24, Issue 11, pp. 3041–3043). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/JBHI.2020.3028398

Dixon, B. E., Gibson, P. J., Frederickson Comer, K., & Rosenman, M. (2015). Measuring Population Health Using Electronic Health Records: Exploring Biases and Representativeness in a Community Health Information Exchange. *Studies in Health Technology and Informatics*, *216*, 1009. https://doi.org/10.3233/978-1-61499-564-7-1009

EEOC. (1978). *EEOC, Uniform Guidelines on Employee Selection Procedures. The U. S. Equal Employment Opportunity Commission (EEOC)*. https://www.govinfo.gov/content/pkg/CFR-2017-title29-vol4/xml/CFR-2017-title29-vol4-part1607.xml

Elreedy, D., & Atiya, A. F. (2019). A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. *Information Sciences*, *505*, 32–64. https://doi.org/10.1016/j.ins.2019.07.070

Fabris, A., Messina, S., Silvello, G., & Susto, G. A. (2022). Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, *36*(6), 2074–2152. https://doi.org/10.1007/s10618-022-00854-z

Feldman, K., Faust, L., Wu, X., Huang, C., & Chawla, N. V. (2017). *Beyond Volume: The Impact of Complex Healthcare Data on the Machine Learning Pipeline*. https://doi.org/10.1007/978-3-319-69775-8_9

Feng, Q., Du, M., Zou, N., & Hu, X. (2022). *Fair Machine Learning in Healthcare: A Review*. http://arxiv.org/abs/2206.14397

Fletcher, R. R., Nakeshimana, A., & Olubeko, O. (2021). Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health. *Frontiers in Artificial Intelligence*, *3*. https://doi.org/10.3389/frai.2020.561802

G. Sartor. (2020). The impact of the General Data Protection Regulation (GDPR) on artificial intelligence. *EPRS | European Parliamentary Research Service*. https://doi.org/10.2861/293

Gajane, P., & Pechenizkiy, M. (2017). On Formalizing Fairness in Prediction with Machine Learning. *ArXiv*, *abs/1710.03184*. http://arxiv.org/abs/1710.03184

Gaonkar, B., Cook B.A, K., & Macyszyn M.D, L. (2020). Ethical Issues Arising due to Bias in Training A.I. Algorithms in Healthcare and Data Sharing as a Potential Solution. *AI Ethics Journal*, *1*(1), 1–11. https://doi.org/https://doi.org/10.47289/AIEJ20200916

Gervasi, S. S., Chen, I. Y., Smith-Mclallen, A., Sontag, D., Obermeyer, Z., Vennera, M., & Chawla, R. (2022). The Potential For Bias In Machine Learning And Opportunities For Health Insurers To Address It. *Health Affairs*, *41*(2), 212–218. https://doi.org/10.1377/hlthaff.2021.01287

Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., & Ranganath, R. (2020). A Review of Challenges and Opportunities in Machine Learning for Health. *AMIA Jt Summits Transl Sci Proc.*, 191–200. https://doi.org/https://doi.org/10.48550/arXiv.1806.00388

Giovanola, B., & Tiribelli, S. (2023). Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. *AI and Society*, *38*(2), 549–563. https://doi.org/10.1007/s00146-022-01455-6

Gurupur, V., & Wan, T. T. H. (2020). Inherent bias in artificial intelligence-based decision support systems for healthcare. In *Medicina (Lithuania)* (Vol. 56, Issue 3). MDPI AG. https://doi.org/10.3390/medicina56030141

Han, X., Chi, J., Chen, Y., Wang, Q., Zhao, H., Zou, N., & Hu, X. (2023). FFB: A Fair Fairness Benchmark for In-Processing Group Fairness Methods. *ArXiv*, *2306.09468v1*. http://arxiv.org/abs/2306.09468

Hsieh, C. J. (2019). High Glucose Variability Increases 30-Day Readmission Rates in Patients with Type 2 Diabetes Hospitalized in Department of Surgery. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-50751-7

Johnson, A. E. W., Ghassemi, M. M., Nemati, S., Niehaus, K. E., Clifton, D., & Clifford, G. D. (2016). Machine Learning and Decision Support in Critical Care. *Proceedings of the IEEE*, *104*(2), 444–466. https://doi.org/10.1109/JPROC.2015.2501978

Kamikubo, R., Wang, L., Marte, C., Mahmood, A., & Kacorri, H. (2022, October 22). Data Representativeness in Accessibility Datasets: A Meta-Analysis. *ASSETS 2022 - Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. https://doi.org/10.1145/3517428.3544826

Kleinberg, G., Diaz, M. J., Batchu, S., & Lucke-Wold, B. (2023). Racial underrepresentation in dermatological datasets leads to biased machine learning models and inequitable healthcare HHS Public Access (PMID 36619609). In *J Biomed Res (Middlet)* (Vol. 2022, Issue 1).

Li, Q., & Mao, Y. (2014). A review of boosting methods for imbalanced data classification. *Pattern Analysis and Applications*, *17*(4), 679–693. https://doi.org/10.1007/s10044-014-0392-8

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable ai: A review of machine learning interpretability methods. In *Entropy* (Vol. 23, Issue 1, pp. 1–45). MDPI AG. https://doi.org/10.3390/e23010018

Lu, H., & Uddin, S. (2022). Explainable Stacking-Based Model for Predicting Hospital Readmission for Diabetic Patients. *Information (Switzerland)*, *13*(9). https://doi.org/10.3390/info13090436

Lum, K. (2017). Limitations of mitigating judicial bias with machine learning. In *Nature Human Behaviour* (Vol. 1, Issue 7). Nature Publishing Group. https://doi.org/10.1038/s41562-017-0141

Mahmoud, M., Bader, M., & McNicholas, J. (2023). Short-Term and Long-Term Readmission Prediction in Uncontrolled Diabetic Patients using Machine Learning Techniques. *International Journal of Advanced Science and Technology* , 680–688. https://doi.org/10.5220/0011926000003414

Makhlouf, K., Montréal, M. M., Zhioua, S., & Palamidessi, C. (2021). *On the Applicability of Machine Learning Fairness Notions*. https://api.semanticscholar.org/CorpusID:235244122

Mathews, S. M. (2019). Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A Literature Review. *Advances in Intelligent Systems and Computing*, *998*, 1269–1292. https://doi.org/10.1007/978-3-030-22868-2_90

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. *ArXiv*, *1908.09635v3*. http://arxiv.org/abs/1908.09635

Oskouei, R. J., & Bigham, B. S. (2017). Over-sampling via under-sampling in strongly imbalanced data. *International Journal of Advanced Intelligence Paradigms*, *9*(1), 58–66. https://doi.org/10.1504/IJAIP.2017.081179

Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Peixoto, R. M., Guimarães, G. A. S., Cruz, G. O. R., Araujo, M. M., Santos, L. L., Cruz, M. A. S., Oliveira, E. L. S., Winkler, I., &

Nascimento, E. G. S. (2023). Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods. In *Big Data and Cognitive Computing* (Vol. 7, Issue 1). MDPI. https://doi.org/10.3390/bdcc7010015

Pawar, U., O'shea, D., Rea, S., & O'reilly, R. (2020). *Incorporating Explainable Artificial Intelligence (XAI) to aid the Understanding of Machine Learning in the Healthcare Domain*. https://doi.org/10.13140/RG.2.2.24754.02246

Pessach, D., & Shmueli, E. (2020). Algorithmic Fairness. *Koret Foundation Grant for Smart Cities and Digital Living 2030.* http://arxiv.org/abs/2001.09784

Pessach, D., & Shmueli, E. (2022). A Review on Fairness in Machine Learning. *ACM Computing Surveys*, *55*(3). https://doi.org/10.1145/3494672

Petersen, E., Holm, S., Ganz, M., & Feragen, A. (2023). The path toward equal performance in medical machine learning. In *Patterns* (Vol. 4, Issue 7). Cell Press. https://doi.org/10.1016/j.patter.2023.100790

Pot, M., Kieusseyan, N., & Prainsack, B. (2021). Not all biases are bad: equitable and inequitable biases in machine learning and radiology. In *Insights into Imaging* (Vol. 12, Issue 1). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1186/s13244-020-00955-7

Quaresmini, C., & Primiero, G. (2023). *Data quality dimensions for fair AI*. http://arxiv.org/abs/2305.06967

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, *169*(12), 866–872. https://doi.org/10.7326/M18-1990

Rasheed, K., Qayyum, A., Ghaly, M., Al-Fuqaha, A., Razi, A., & Qadir, J. (2021). *Explainable, Trustworthy, and Ethical Machine Learning for Healthcare: A Survey*. https://doi.org/https://doi.org/10.1016/j.compbiomed.2022.106043

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-Precision Model-Agnostic Explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence : Vol. Vol. 32 No. 1 (2018)*. https://doi.org/https://doi.org/10.1609/aaai.v32i1.11491

Rojas, J. C., Fahrenbach, J., Makhni, S., Cook, S. C., Williams, J. S., Umscheid, C. A., & Chin, M. H. (2022). Framework for Integrating Equity Into Machine Learning Models: A Case Study. In *Chest* (Vol. 161, Issue 6, pp. 1621–1627). Elsevier Inc. https://doi.org/10.1016/j.chest.2022.02.001

Rolf, E., Worledge, T., Recht, B., & Jordan, M. I. (2021). *Representation Matters: Assessing the Importance of Subgroup Allocations in Training Data*. http://arxiv.org/abs/2103.03399

Salih, A., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Menegaz, G., & Lekadir, K. (2023). *Commentary on explainable artificial intelligence methods: SHAP and LIME*. http://arxiv.org/abs/2305.02012

Santoso, N., Wibowo, W., & Himawati, H. (2019). Integration of synthetic minority oversampling technique for imbalanced class. *Indonesian Journal of Electrical Engineering and Computer Science*, *13*(1), 102–108. https://doi.org/10.11591/ijeecs.v13.i1.pp102-108

Saxena, A., & Chandra, S. (2021). Artificial Intelligence and Machine Learning in Healthcare. In *Artificial Intelligence and Machine Learning in Healthcare*. Springer Singapore. https://doi.org/10.1007/978-981-16-0811-7

Shang, Y., Jiang, K., Wang, L., Zhang, Z., Zhou, S., Liu, Y., Dong, J., & Wu, H. (2021). The 30-days hospital readmission risk in diabetic patients: predictive modeling with machine learning classifiers. *BMC Medical Informatics and Decision Making*, *21*. https://doi.org/10.1186/s12911-021-01423-y

Skirpan, M., & Gorelick, M. (2017). *The Authority of "Fair" in Machine Learning*. http://arxiv.org/abs/1706.09976

Spanakis, E. K., Singh, L. G., Siddiqui, T., Sorkin, J. D., Notas, G., Magee, M. F., Fink, J. C., Zhan, M., & Umpierrez, G. E. (2020). Association of glucose variability at the last day of hospitalization with 30-day readmission in adults with diabetes. *BMJ Open Diabetes Research and Care*, *8*(1). https://doi.org/10.1136/bmjdrc-2019-000990

Srinivasan, R., & Chander, A. (2021). Biases in AI systems. In *Communications of the ACM* (Vol. 64, Issue 8, pp. 44–49). Association for Computing Machinery. https://doi.org/10.1145/3464903

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., & Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-00349-y

UCI Diabetes Dataset, Clore, J., Cios, K., Strack, B., & DeShazo, J. (2014). Diabetes Data Set. In *BioMed Research International*. BioMed Research International. https://doi.org/https://doi.org/10.24432/C5230J.

Wojtusiak, J. (2021). Reproducibility, transparency and evaluation of machine learning in health applications. *HEALTHINF 2021 - 14th International Conference on Health Informatics; Part of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2021*, 685–692. https://doi.org/10.5220/0010348306850692

Yan, S., Kao, H. Te, & Ferrara, E. (2020). Fair Class Balancing: Enhancing Model Fairness without Observing Sensitive Attributes. *International Conference on Information and Knowledge Management, Proceedings*, 1715–1724. https://doi.org/10.1145/3340531.3411980

Yu, Z., Chakraborty, J., & Menzies, T. (2021). *FairBalance: How to Achieve Equalized Odds With Data Pre-processing*. http://arxiv.org/abs/2107.08310

Yuan, M., Kumar, V., Ahmad, M. A., & Teredesai, A. (2021). *Assessing Fairness in Classification Parity of Machine Learning Models in Healthcare*. http://arxiv.org/abs/2102.03717

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). *Mitigating Unwanted Biases with Adversarial Learning*. http://arxiv.org/abs/1801.07593

Zhu, M., Xia, J., Jin, X., Yan, M., Cai, G., Yan, J., & Ning, G. (2018). Class weights random forest algorithm for processing class imbalanced medical data. *IEEE Access*, *6*, 4641–4652. https://doi.org/10.1109/ACCESS.2018.2789428

Žliobaitė, I., & Custers, B. (2016). Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, *24*(2), 183–201. https://doi.org/10.1007/s10506-016-9182-5