NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF PHYSICS

MARIANA SOFIA FIGUEIREDO PINTO

Bachelor of Science in Biomedical Engineering

BIAS DISCOVERY AND MITIGATION IN MEDICAL ARTIFICIAL INTELLIGENCE

MASTER IN BIOMEDICAL ENGINEERING

NOVA University Lisbon
October, 2023

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF PHYSICS

BIAS DISCOVERY AND MITIGATION IN MEDICAL ARTIFICIAL INTELLIGENCE

MARIANA SOFIA FIGUEIREDO PINTO

Bachelor of Science in Biomedical Engineering

Adviser: Prof. Dr. Hugo Filipe Silveira Gamboa
Associate Professor, NOVA University Lisbon

Co-adviser: Dr. André Valério Raposo Carreiro
Senior Scientist, Fraunhofer Portugal

MASTER IN BIOMEDICAL ENGINEERING

NOVA University Lisbon
October, 2023

BiAs Discovery and Mitigation in medical Artificial iNtelligence

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v6.10.17) [47].

To my dear sister for believing in me even when I couldn't.

Acknowledgements

First, I want to express my gratitude to my advisers. I thank Professor Hugo Gamboa for the opportunity of developing this work under a wonderful environment at Fraunhofer AICOS. This experience allowed me to developed multiple skills in a field I've grown to adore.

For the close guidance, wisdom and support, I thank André Carreiro. Your enthusiasm for the theme and the endless encouragement allowed me to achieve more than what I could expect. To Pedro Madeira, for providing reassurance with your calm nature and good humour. I thank you both for the advice and invaluable contributions, it has been a pleasure to work with you. A special thanks extends to the team members at the Associação Fraunhofer Portugal AICOS, for the warm welcome I was offered at the office.

To my mother and father for understanding when I couldn't go home as often, supporting unconditionally my journey, even if they were not fully aware of what exactly my dissertation was about. To my sister, for aiding me in this and other journeys, and pushing me to a better self.

To my friends, from my hometown and the newly found ones in the big city, for every precious memory, the jokes, and the shared despair.

To everyone that I didn't mention but are just as present and important in my life.

To every setback that I had to surpass and every experience that led me to where I am now. I'm truly grateful I've gotten this far.

“ I have loved the stars too fondly to be fearful of the night. ”

— Sarah Williams

(Poet)

Abstract

Machine-learning systems are used to improve efficiency and quality of results and should uphold an impartiality standard above human decisions. Nevertheless, biases are frequently observed, leading to suboptimal outcomes for specific groups. This problem is amplified in healthcare by the field's complexity, limitations, and implications of its applications. The most common problem is the lack of enough samples to accurately represent the population, resulting in impeding consequences in these specific groups.

Existing methods for evaluating these systems vary from evaluating the global performance of studied groups to comparing similar samples in an instance-based analysis. From the latter approach, the methodology of generating counterfactuals, samples modified to answer "what if..?" scenarios, has gained popularity in recent years, valued for its interpretability and versatility.

However, despite the instance-based perspective it provides, there is a gap in how to properly generalize this methodology. This work extends this approach by exploring novel evaluation metrics supported by a new visualization analogous to the confusion matrix. It also explores the plausibility of the generated counterfactuals, experimenting with the incorporation of domain knowledge. Motivated by a prevalent issue in healthcare - data scarcity - it analyzes the results of performing data augmentation with counterfactuals to mitigate bias without compromising performance.

As a result, this work contributes with a new bias detection and mitigation technique and reports promising results for ensuring more reliable decision-support systems in healthcare.

Keywords: Bias, Machine-Learning, Counterfactuals, Fairness, Augmentation

Resumo

Sistemas de aprendizagem automática são utilizados para melhorar a eficiência e a qualidade dos resultados, e deveriam manter um padrão de imparcialidade acima das decisões humanas. No entanto, é frequente observarem-se enviesamentos, que levam a decisões não adaptadas para grupos específicos. Este problema é extrapolado nos cuidados de saúde devido à complexidade, limitações e implicações nesta aplicação. Um problema recorrente é a falta de amostras suficientes para representar a população devidamente, o que resulta em consequências irreversíveis em grupos específicos.

Os métodos existentes para a avaliação destes sistemas variam desde a avaliação do desempenho global dos grupos em estudo, até a comparação de amostras semelhantes em uma análise instância a instância. A partir desta última abordagem, a metodologia de geração de contrafactuais, amostras modificadas para responder a cenários "e se..?", ganhou popularidade nos últimos anos, valorizada pela sua interpretabilidade e versatilidade.

No entanto, apesar da perspetiva local que proporciona, permanece uma lacuna na forma de generalizar corretamente esta metodologia. Este trabalho expande esta abordagem explorando novas métricas de avaliação apoiadas por uma nova visualização análoga à matriz de confusão tradicional. Também explora a plausibilidade dos contrafactuais gerados, experimentando a incorporação de conhecimento do domínio. Como resposta ao problema prevalente dos cuidados de saúde - escassez de dados - é estudado o aumento de dados com contrafactuais para mitigar o enviesamento sem comprometer o desempenho.

Por conseguinte, este trabalho contribui com uma nova técnica de deteção e mitigação de enviesamentos e apresenta resultados promissores para garantir sistemas de suporte à decisão em saúde mais fiáveis.

Palavras-chave: Viés, Aprendizagem Automática, Contrafactuais, Justiça, Aumento de dados

Contents

List of Figures

Abbreviations

ACC Accuracy *(pp. 37, 38, 53, 58, 59)*
ADASYN Adaptive Synthetic Sampling *(pp. 31, 32)*
AI Artificial Intelligence *(pp. 1, 4, 10)*

BMI Body Mass Index *(pp. 26, 28, 30, 33)*
BSA Body Surface Area *(pp. 26, 28, 30)*

CAD Coronary Artery Disease *(pp. 27, 37)*
CCM Counterfactual Confusion Matrix *(pp. 21, 23, 39, 40, 53)*
CDF Cumulative Distribution Function *(pp. 18, 19)*
CF Counterfactual *(pp. 2, 3, 11, 14, 16–26, 28–40, 53, 58)*
CFE Counterfactual Explanation *(pp. 11, 16, 17)*
cGAN Conditional Generative Adversarial Network *(p. 40)*
CM Confusion Matrix *(pp. 8, 21, 23, 25, 40, 53)*
CMCC Counterfactual Matthew's Correlation Coefficient *(pp. 22, 29, 31, 32, 35, 36, 38, 53, 58, 59)*
CN Consistent Negative *(pp. 21, 22)*
CP Consistent Positive *(pp. 21, 22)*
CR Consistency Rate *(pp. 21, 22, 53)*

DiCE Diverse Counterfactual Explanations *(p. 16)*
DL Deep Learning *(p. 4)*

ECCM Extended Counterfactual Confusion Matrix *(pp. 23, 25, 30–32, 35–37, 39, 40, 53–58)*
ECG Eletrocardiogram *(p. 37)*
ENN Edited Nearest Neighbours *(pp. 35, 40)*
EOdds Equalized Odds *(pp. 13, 32)*
EOpp Equality of Opportunity *(pp. 13, 53, 58)*

FCN False Consistent Negatives *(pp. 23–25)*
FCP False Consistent Positives *(pp. 23–25)*
FDR False Discovery Rate *(p. 8)*
FN False Negative *(pp. 8–10, 13, 22, 26, 27, 29, 30, 37)*
FNR False Negative Rate *(pp. 9, 58)*
FNSR False Negative Switch Rate *(pp. 24, 29, 30, 32, 35, 53)*
FOR False Omission Rate *(p. 8)*
FP False Positive *(pp. 8–10, 13, 22–24, 26, 27, 29, 30, 34, 38, 58)*
FPR False Positive Rate *(pp. 9, 58)*
FPSR False Positive Switch Rate *(pp. 24, 32, 33, 35, 53, 58, 59)*
FSN False Switched Negatives *(pp. 24, 25)*
FSP False Switched Positives *(pp. 24, 25)*

GAN Generative Adversarial Network *(pp. 17, 39)*
GBDT Gradient Boosting Decision Tree *(p. 16)*

ICML International Conference on Machine Learning *(pp. 40, 57)*

JSCD Jensen-Shannon Counterfactual Divergence *(pp. 25, 59)*

MCC Matthew's Correlation Coefficient *(pp. 9, 10, 32, 35, 36, 53)*
ML Machine Learning *(pp. 1, 2, 4–7, 10–12, 23, 32, 39, 40, 57)*

NCR Negative Consistency Rate *(p. 53)*
NN Neural Networks *(p. 17)*
NPV Negative Predicted Vaue *(p. 8)*
NSR Negative Switch Rate *(pp. 22, 24, 30–32, 35–38, 53, 58, 59)*

OTM Optimal Transport Map *(pp. 17, 39)*

PCP Positive Consistent Precision *(pp. 22, 23, 53)*
PCR Positive Consistency Rate *(p. 53)*
PDF Probability Distribution Function *(pp. 19, 33)*
PPV Precision or Positive Predicted Value *(pp. 8, 9, 22, 23, 53)*
PredEq Predictive Equality *(pp. 13, 53, 58)*
PredP Predictive Parity *(pp. 13, 29–31, 34–38, 53)*
PSDR Positive Switch Discovery Rate *(p. 53)*
PSR Positive Switch Rate *(pp. 22, 25, 29, 30, 32, 35–38, 53, 59)*

RMSCD Root Mean Squared Counterfactual Differences *(pp. 26, 59)*
RMSE Root Mean Squared Error *(p. 25)*

SHAP SHapley Additive exPlanations *(pp. 11, 12, 27)*
SN Switched to Negative *(pp. 21, 22)*
SP Switched to Positive *(pp. 21, 22)*
SR Switch Rate *(pp. 22, 29, 32, 53)*

TCN True Consistent Negatives *(pp. 23–25)*
TCP True Consistent Positives *(pp. 23–25)*
TN True Negative *(pp. 8–10, 29, 30, 33, 34, 38, 58)*
TNR Specificity or True Negative Rate *(pp. 9, 22, 29–32, 34–36, 38, 53, 59)*
TNSR True Negative Switch Rate *(pp. 24, 30, 32–35, 37, 38, 53, 59)*
TP True Positive *(pp. 8–10, 13, 24, 29, 30, 33, 37)*
TPR Recall or True Positive Value *(pp. 9, 22, 26, 27, 29–38, 53, 58, 59)*
TPSR True Positive Switch Rate *(pp. 24, 25, 30–35, 37, 38, 53)*
TSN True Switched Negatives *(p. 24)*
TSNR True Switch Negative Rate *(p. 24)*
TSP True Switched Positives *(p. 24)*
TSPR True Switch Positive Rate *(p. 24)*

WIT What-If Tool *(p. 16)*

XAI eXpainable Artificial Intelligence *(pp. 10, 11)*

1 Introduction

1.1 Motivation

In 1943, *Walter Pitts* and *Warren McCulloch* developed a mathematical model to mimic neural networks and emulate the human thought process [50]. This study was motivated by furthering the understanding of the human brain, but soon the interest was broadened to improve computation. Then the curiosity and potential in Artificial Intelligence (AI) grew with new methods, applications and goals, improving the time and efficiency of tasks.

In fact, the mathematical foundation behind Machine Learning (ML) models allows for an unparalleled mechanism of finding hidden patterns within the provided data. But because of the outstanding efficiency, it can easily veil some unwanted tendencies. These tendencies, known as bias, may occur due to the incomplete scenery painted by the data, with misrepresentation of the target population, either as a consequence of mishandling it, or due to the inherent process of optimization that prioritizes performance over othercriteria. And, regardless of origin, bias can lead to ill-adapted models for specific population groups.

The most prominent example of models assimilating bias is the results of *Northpointe's COMPAS*, an ML model that uses defendant's data to assess their risk of committing another crime. When the output was analyzed, it was noted a tendency to impair African Americans, such that, given the same parameters, a man of this ethnicity would have a higher likelihood of a higher risk assessment than a man of another ethnicity [4].

Fairness has since become a prominent topic, especially in high-risk fields such as healthcare, an area where lives are at stake, and hence where bias can be a critical liability.

Such findings were also reported when using psychiatric notes to predict Intensive Care Unit (ICU) mortality and 30-day psychiatric readmission. The constructed model had lower performance depending on the gender, race, and health insurance of the patient. In spite of this, this model serves as an example and proof of the existing prejudice in the original clinical notes [17].

Due to the complex nature of health-related problems, bias can be created even when

there is no evident prejudice the algorithm can mirror. For instance, models concerning pathologies are often limited by a minimal volume of samples of affected individuals from which to learn. Aggravating the problem, every health condition has a myriad of symptoms that correlate with the patient's gender, age, race, family history, environment, and corresponding lifestyle. At the same time, the datasets tend to have under-represented groups which may lead to subpar performance.

A combination of these issues manifests when detecting melanomas in darker skin tones. Studies report a large percentage of incorrect assessments from dermatologists for darker phenotypes. This is rooted in the substantially lower prevalence of this disease for darker skin tones which leads to ignoring the correct possibility before it is too late. ML models also display the same tendency. Due to the small portion of positive tests for this group, the model is better optimized to the more represented group leading to clear discrepancies in performance [23]

These concerns are often addressed through questions such as 'What caused the bias?', 'Is this decision truly because of belonging to this group?', or rather 'What if this person belonged to another group?'. These questions introduce a line of thought based on event causality known as *Counterfactual Thinking* [63]. This logic is applied in ML with variable levels of success, presenting great results for local analysis but with limited evaluation strategies for global evaluations of the model.

1.2 Objectives

In response to the challenges posed by bias in ML, this work presents a methodology that complements existing bias detection and mitigation techniques. The objectives of this study are as follows:

1. Expand upon the notion of Counterfactual (CF) fairness to develop a novel bias detection methodology and plausible CF generation process tailored for classification tasks, focusing on its application to real-world medical datasets.

2. Perform a comprehensive evaluation of the proposed bias detection approach, comparing it with established techniques.

3. Evaluate the impact of domain knowledge on bias detection and mitigation within the proposed methodology.

4. Investigate the effectiveness of data augmentation using counterfactuals as a bias mitigation technique, evaluating it against proposed bias evaluation metrics and state-of-the-art methods.

The framework of this work is depicted in Figure 1.1.

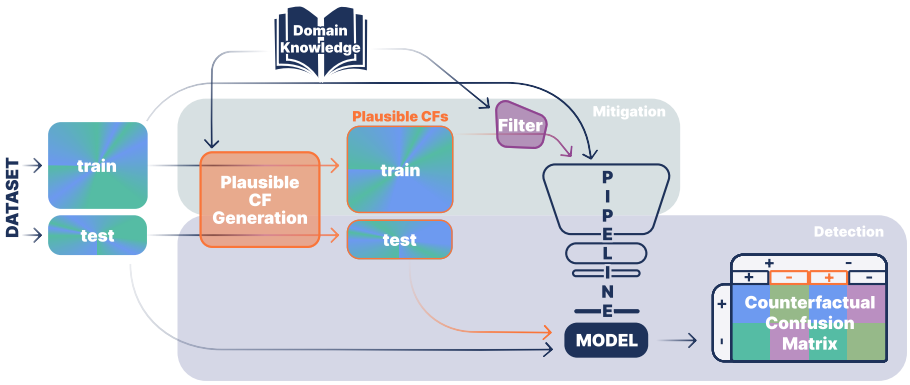


Figure 1.1: Proposed framework for detecting and mitigating bias employing CFs.

1.3 Document Structure

This document is written following the standard structure. Starting with the current Chapter 1 where the context and motivation are presented, as well as the goals it proposes to achieve. In Chapter 2 several concepts are introduced as a foundation for a better understanding of the problem, the current solutions, and the proposed framework. Next, in Chapter 3, the most relevant studies that inspire this work are highlighted.

The second half of the document focuses on the proposed methodology. First, Chapter 4 details the processes for developing the solution. The conducted experiences are presented and analyzed in Chapter 5. Finally, Chapter 6 further discusses the overall applicability and importance of the proposed approach, as well as limitations and points of improvement.

1.4 Declaration of Originality

The research work described in this dissertation was carried out in accordance with the norms established in the ethics code of Universidade Nova de Lisboa. The work described and the material presented in this dissertation, with the exceptions clearly indicated, constitute original work carried out by the author.

2

Theoretical Concepts

This chapter introduces the fundamentals for a clear understanding of the developed work. It presents an outline of the ML pipeline, with a focus on the phases that could potentially introduce bias. The proposed methodology was developed for supervised classification problems, and tested with tabular data and binary labels, but with the potential for expansion to other data typologies.

2.1 Machine Learning

ML is a branch of AI that utilizes data and algorithms to attempt to replicate the way humans learn. An ML model applies mathematical algorithms to a large number of samples to identify trends and produce an output [43]. Deep Learning (DL) is a subset of ML that uses multiple processing layers to extract features from the data [46]. DL models can be used for data of any nature but show special utility for unstructured data that can't be effectively processed through less complex methods. Models can be defined as generative or discriminative [36]. *Generative Models* use the provided data to generate new samples, distinct from the original samples. *Discriminative Models* provide a prediction based on the data it was trained on. For example, optimizing product displacement in a pharmacy, identifying a tumor in an image, or predicting diseases through clinical notes.

Discriminative ML tasks can be divided into unsupervised and supervised learning. *Unsupervised* models are trained on unlabeled data. The model can learn to aggregate samples based on their similarity. They can be employed in medical imaging to identify anomalies. *Supervised learning* refers to problems where the data it was trained on includes the ground-truth label as target for each sample. For example, when predicting a disease, the dataset includes relevant information and if each individual is healthy. Thus, extrapolating the patterns from healthy and unhealthy individuals, the model is able to identify if new patients are healthy or not.

Depending on the label, a supervised problem can be defined as a classification problem for binary, $Yi \in 0, 1$, or multivariate, $Yi \in \mathbb{N}$, labels, or as a regression problem for continuous labels, $Yi \in \mathbb{R}$. A problem can be further defined as multi-label if it attempts

to predict multiple characteristics simultaneously.

As for the data typologies, simpler ML models often work with tabular data. However, there are techniques for unstructured data, like time series and images. It is important to note that tabular features can be extrapolated from time series and even images, with variable degrees of success in the final model.

2.1.1 Biases in the ML Pipeline

After the base definitions for ML models, this section presents each step of the creation of a supervised classification model for tabular data and potential biases that arise from each stage(see Figure 2.1), following the review works of *Mehari* and *Suresh* [51, 72].

2.1.1.1 Data Aquisition

In this stage, there must be a balance between the quantity of information that can be acquired in one acquisition and the context limitations, such as the feasibility of acquisition techniques for the target population, time constraints, and available equipment. Due to the reduced sheer volume of potential participants in clinical studies, the acquisition is often conducted for a long period, sometimes in different locations, revealing the necessity for a rigorous protocol.

The limitations incurred in this stage can often lead to the introduction of bias. If the acquired data doesn't include relevant features or there are incongruous acquisition processes across different locations, the model will reflect *sampling bias*. If a particular group is ill-represented in the dataset, not reflecting real-world characteristics, it incurs *measurement bias*. Finally, *representation bias* is an easier-tracked bias that happens when one group is less represented than another, which is very common in healthcare as different diseases are tied to different phenotypes.

2.1.1.2 Preprocessing

This stage encompasses various essential tasks, starting with *feature extraction*. Since ML algorithms are typically designed to work with numerical data, text data must undergo a process known as *type casting* to be converted into a numeric format. For features with multiple categories that lack a continuous relationship between them, an *encoding* process is often applied. In this process, features are decomposed into new features.

In a process known as *Feature Engineering*, an analysis of the content of features is conducted, which culminates in simplifying, dropping, and/or combining different features, according to the essential information they provide. Features may be discarded due to excessive missing values or severely unbalanced distribution of values or because they contain redundant information already provided by another feature. The redundancy of the dataset is often assessed through correlation maps, a visual representation of the *Pearson correlation coefficient* for each set of features, and other correlation coefficients.

Labels are also processed in this stage, though they are only included in specific steps. A recurrent problem in ML is an *Unbalanced Dataset*, meaning that the prevalence of an outcome is substantially higher than the others. Knowing this, the problem can often be simplified from a multi-class problem to a binary problem. For example, if the objective is identifying a set of coronary diseases and the number of samples for each condition is not satisfactory, then the problem can be simplified to identifying if the patient has *any* coronary disease.

The dataset is then dividedinto two main groups: the training and the testing set. The training set is used as input so that the model can learn, while the testing set is reserved for evaluating the model's performance on data it has never encountered during training. Often, there is a third set, the validation set, which will be explored further in Section 2.1.1.4. The final steps of preprocessing involve handling missing values and normalizing the range of feature values. Feature normalization is an important step that allows to remove the influence of differences in value ranges preventing disproportionate contributions based on the feature's range. Both of these procedures are initially applied to the training set. This helps in identifying the appropriate strategies for imputing missing values and defining the minimum and maximum values for feature scaling. Once these parameters are learned from the training set, they are then applied to the other sets to ensure consistency and fairness in model evaluation.

2.1.1.3 Feature Selection

Feature Selection is done because of what is known as *The Curse of Dimensionality*. While more information is beneficial for improving performance, the time and power required for model development increase exponentially with each dimension. At the same time, a large number of dimensions results in samples sparsed across the data space, which is not desirable [52]. Thus, there is a need for a good ratio between the number of features and the number of samples. A good rule of thumb for the number of features is using one tenth of the number of samples, but it depends on the task. Dimensionality reduction techniques can be categorized into filter methods, wrapper methods, and embedded methods. For filter methods, the correlation between features is calculated and the least independent features are removed. In wrapper methods, the best features are selected according to the evaluation of models with different subsets of features. In embedded methods, new features are created based on the combination of existing ones. If an essential feature is removed in any step of feature selection it may induce a bias known as *omitted variables*, where crucial information for a determined group was not included in the model.

2.1.1.4 Fitting and Optimization

After completing the data-cleaning phase, the model is prepared for training. This crucial step encompasses three key processes: selecting potential models, fitting them to the training data, and evaluating performance. Among supervised models, the most

commonly employed ones are *Naive Bayes*, *Decision Trees*, *Logistic Regression*, *K-Nearest Neighbors*, and *Support Vector Machine*. These algorithms can be used individually or combined to form more robust ensembles. For example, *Random Forest* algorithms generate several decision trees trained on random subsets of samples and features, through a *bagging (Bootstrap Aggregation)* technique [13]. Each of these trees predicts an outcome and the final decision is selected through majority voting. Because of this, random forests tend to be more robust and less prone to overfitting, compared to decision trees.

As an overall rule, the selected model should adhere to *Occam's Razor* principle, which can be interpreted in context as: *for similar performance, a simpler solution is preferred to a complex one*. Striking the right balance between model simplicity and performance is key to ensuring that the chosen model achieves good generalization without unnecessary computational or runtime overhead [56].

Each algorithm has a set of adjustable hyperparameters, whose optimization is pivotal to creating a reliable model [70]. However, there is a delicate balance to strike: since models are tailored to the training data, overfitting often occurs [14]. *Overfitting* is when a model performs exceptionally well on the training data but poorly on testing data. When overfitting occurs, fine-tuning the algorithm's hyperparameters is necessary. The objective is to achieve comparable performance metrics on both the training and testing datasets.

To grant optimal hyperparameter values while preventing overfitting, the training set can be further split, adding a validation set. During the optimization process, the model is iteratively trained and evaluated using this validation set. The validation set may consist of either a fixed group of samples or through *Cross Validation* in which multiple groups that rotate in sequence, facilitating the selection of the model that exhibits optimal performance across all iterations. It is important to note that the testing set must not be used as a validation set, as its sole purpose is to evaluate the final model.

Another relevant category of optimization techniques in ML includes post-processing methods, which focus on adjusting the model's predictions rather than modifying the model architecture itself. One of the most common methods for algorithms that produce probabilities involves shifting the prediction threshold. The threshold represents the probability limit at which a prediction changes. In binary classification, the default threshold is typically set at 50%, but this value can be adjusted to optimize performance. There are also other techniques in this category, such as calibration. Calibration involves training a new method to adjust the predicted probability distribution so that it closely resembles the distribution of the actual training outcomes. If the performance is too low, it may be necessary to try more complex algorithms and revisit the feature engineering step. In some instances it may be necessary to adapt the problem itself, switching from multi-class to a binary classification.

As mentioned, optimization may lead to over-adapting the model to the most represented group, introducing *learning bias*. On the other hand, an oversimplified model may as well subdue the less represented group, constituting *aggregation bias*. As a final typology, bias can even be introduced in the labels, which is the case for *historical bias*. In

this case, the original classification of the labels was skewed by real-world discrimination.

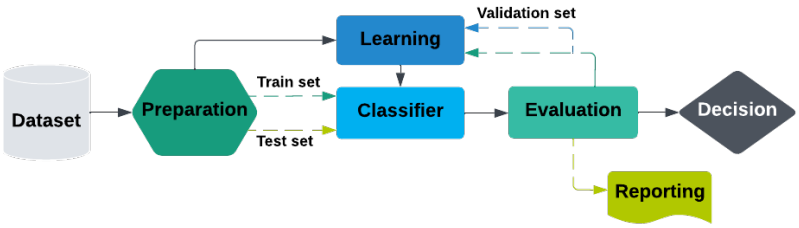


Figure 2.1: A typical Machine Learning pipeline.

2.1.2 Performance Metrics

During the evaluation stage, the performance is inferred based on the number of correct and incorrect predictions in relation to the ground truth. Focusing on binary classification tasks, each prediction, positive (P) or negative (N), is assessed as either true (T) or false (F) following the nomenclature: True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). A Confusion Matrix (CM) is a useful analytical tool that allows a direct comparison between these outcomes, offering a better understanding of the model's shortcomings, represented in Figure 2.2. For a more in-depth analysis, there are multiple metrics for evaluation based on these values:

- Accuracy: The overall rate at which the model is right.

$$\frac{TP + TN}{Total} \in [0, 1]$$

- Prevalence and Predicted Prevalence: Rate of real and predicted positive outcomes.

$$Prevalence = \frac{TP + FN}{Total} \in [0, 1], \quad PredictedPrevalence = \frac{TP + FP}{Total} \in [0, 1]$$

- Parameters that measure the rate at which the model correctly predicts a positive or negative outcome out of the total positive or negative predicted outcomes:
  - Precision or Positive Predicted Value (PPV) and Negative Predicted Vaue (NPV) The rate at which the model is right when it predicts positive and negative values.

$$PPV = \frac{TP}{TP + FP} \in [0, 1], \qquad NPV = \frac{TN}{FN + TN} \in [0, 1]$$

  - False Discovery Rate (FDR) and False Omission Rate (FOR) The rate at which the model is wrong when it predicts positive and negative values.

$$FDR = \frac{FP}{TP + FP} \in [0, 1], \qquad FOR = \frac{FN}{FN + TN} \in [0, 1]$$

- Parameters that measure the rate at which the model correctly predicts a positive or negative outcome out of the total real positive or negative outcomes:
  - Recall or True Positive Value (TPR), and Specificity or True Negative Rate (TNR) The rate at which the model correctly predicts positive and negative values.

$$TPR = \frac{TP}{TP + FN} \in [0,1], \qquad TNR = \frac{TN}{TN + FP} \in [0,1]$$

  - False Positive Rate (FPR) and False Negative Rate (FNR) The rate at which the model incorrectly predicts positive and negative values.

$$FPR = \frac{FP}{TN + FP} \in [0,1], \qquad FNR = \frac{FN}{TP + FN} \in [0,1]$$

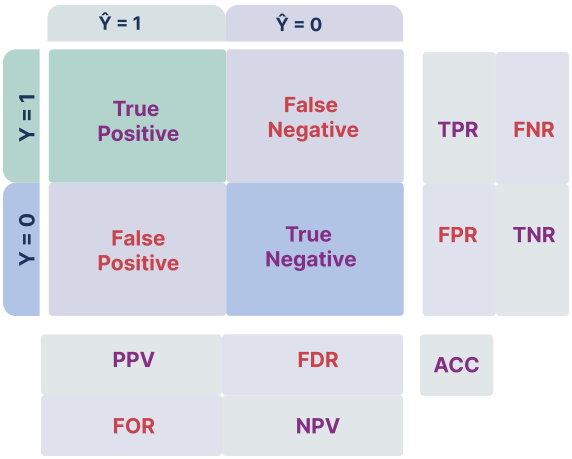


Figure 2.2: Confusion Matrix with different metrics calculated based on the table values. Adapted from [19].

The performance of the model is improved by reducing the number of FP and FN, known as type I and type II errors. The main metrics used to evaluate these errors are precision and recall, respectively. While the goal is to maximize both of these metrics, it is not possible to do so simultaneously. This is known as the *Recall/Precision Trade-off* [30]. Different problems require a different equilibrium, depending on the costs of FP and FN in that specific scenario. If both errors are equally costly, there are additional metrics that measure the distribution of the errors, such as the *F1-Score*, that corresponds to the harmonic mean of precision and recall.

When employing evaluating performance, it is important to consider the class imbalance. For unbalanced datasets, metrics that count samples with different ground truths are not reliable. For instance, PPV is not suitable when there are substantially less positive samples since even a small portion of FP can overstate the TPs. Metrics such as TPR and TNR are exempt from this problem but the accuracy is not. To solve this problem there is the *Balanced Accuracy*, which gives the mean of TPR and TNR, as well as the Matthew's

Correlation Coefficient (MCC), claimed for its robustness and concordance with the other performance metrics [18].

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \in [-1, 1]$$

2.1.3 Data Augmentation

Despite dedicated efforts to build ML models, the inherent characteristics of a dataset can sometimes result in sub-optimal model performance. This challenge is particularly pronounced when dealing with unbalanced datasets, or datasets that lack diversity among the samples. In such scenarios, the solution often seems simple: collect more data. However, collecting additional data is not always feasible due to constraints such as data collection costs, time constraints, or even the unavailability of certain data points.

Recognizing the need for a practical solution, several data augmentation techniques have emerged. It is a valuable approach that expands the sample space through various techniques from oversampling the minority class to generating synthetic data [26, 55]. These techniques have proven to be effective in improving model performance, especially in the face of data limitations. By augmenting the dataset with synthetic data, ML models can better capture underlying patterns, reduce biases, and improve their ability to generalize. However, the choice of technique and its success still depends on the specific problem and dataset at hand.

A key challenge in synthetic data generation techniques is validating the quality and reliability of synthetic samples [55, 73]. The lack of universally accepted validation methods remains a concern, as ensuring that synthetic data accurately represents the real-world dataset is vital for trustworthy ML results.

2.1.4 Explainability

While ML models can perform exceptionally well, they often lack the interpretability needed to identify the reasons behind a particular prediction. Understanding the "why"is crucial for further research and for validating the model's accuracy.

eXpainable Artificial Intelligence (XAI) is one of AI research and methods development that excel at extracting the reasoning behind each decision. XAI plays a key role in identifying errors within the model, which is essential for any developer. It also provides transparency and accountability to users, ensuring compliance. This section focuses on more general methods of explainability, even though bias and fairness are tightly intertwined with XAI [3, 54].

XAI techniques can be distinguished by the scope of the explanation, the model specificity, and at what stage they are employed. Concerning the scope of the explanation, XAI methods can be defined as global for techniques that aim to provide an overall understanding of the model's behavior across the entire dataset; or local for an instance-based analysis.

One of the primary concerns in XAI is determining the most important features. The umbrella term *Feature Importance* covers these methods. These techniques are used not only for the aforementioned reasons, but also to build a new architecture that uses only the most important features, thereby reducing dimensionality, improving efficiency, and potentially improving performance.

SHapley Additive exPlanations (SHAP) are widely recognized as a global explanation methodology for assessing feature importance [40]. Rooted on *Coalitional Game Theory*, the algorithm attempts to sort the features by their contribution to the prediction. This is achieved by calculating the *marginal contribution* of a feature, given by the difference between the average prediction with and without said feature.

Counterfactual Explanation (CFE) is a technique used in ML to provide local explanations through 'what-if' scenarios. These explanations involve generating alternative data points from a given sample by systematically modifying feature values while keeping other aspects unchanged, called CFs. The goal is to understand how changes in individual features or conditions would affect the model's output or decision [75]. CFE are particularly valuable in scenarios such as Clinical Decision Support (CDS) systems. When combined, these tools help healthcare professionals comprehend why a specific treatment option was recommended over others, by presenting alternative scenarios and highlighting the necessary changes in patient characteristics for different outcomes [68].

2.1.5 Domain Knowledge

ML models utilize data to generate outcomes. However, they are limited by the representations in the data and, in some cases, there may be certain domain implications not included in the data or model architecture. For this reason, it is imperative to aid the process through human knowledge, more specifically *Domain Knowledge*.

Domain knowledge is applied in virtually every step in the ML pipeline. In practice, it refers to using specific industry knowledge to ensure better decisions in the criteria of performance and fairness. It is essential for feature engineering and selection, performance evaluation, synthetic data validation as well as for explainability tasks.

2.2 Bias

As a broad concept, bias in ML can be defined as the tendency of a model to make incorrect predictions in a way that may be prejudicial to a particular group [57]. Because bias is intrinsically related to fairness, there is not a clear and cut-through definition that can be applied to every case as the latter is in itself an argued topic in the field of philosophy. For this reason, the proposed solutions to detect and mitigate biases vary greatly according to their contextual application. Currently, the large scope of research strives to improve existing methods as well as challenge the accepted notions, hoping to achieve a concise framework on how to prevent unreliable models. This section presents

the current detection and mitigation methodologies as well as their caveats and points of improvement.

2.2.1 Bias Detection

When analyzing the results of a model, it can be easy to identify some patterns in the labels that may point to potential bias. Nevertheless if the type of bias is not properly identified, trying to compensate for these results may be unfair in itself [2]. This may not be an easy task, as bias can originate throughout any step of the ML pipeline, from data acquisition and pre-processing to model training, evaluation, and deployment.

The first step for identifying bias is to define the sensitive feature, meaning the feature associated with the type of group we want to study. Often, this feature is related to sensitive information such as gender, race, and age, but it can refer to any information [61]. Another similarly used term is protected attribute but is often restricted to fairness applications. Sensitive features can be defined *a priori* or through statistical analysis that may point out underrepresented groups or through *post-hoc* methods such as SHAP values that may identify a feature as more relevant than expected.

Analogous to the importance of choosing the most suitable performance metrics, defining the right fairness criteria the model is expected to achieve is crucial. The two main broad definitions are **Group Fairness** and **Individual Fairness**, referring to global and local methodologies, respectively [1]. Group fairness focuses on ensuring that fairness is achieved at a group level, whereas individual fairness focuses on ensuring that individuals who have similar characteristics or feature values are treated equally by the model, regardless of their group provenance.

2.2.1.1 Group Fairness

The most prevalent methodology for detecting bias is known as group fairness. Existing fairness metrics under this umbrella term are referred to as *measure parities* as they compare performance parameters between the data sectioned by the subgroups of the sensitive feature. These metrics can be further divided by the fairness definition employed: *Independence*, *Separation*, and *Sufficiency*, [7].

By the independence criterion, a model is considered fair if the outcome is independent of the sensitive feature. Statistically, this can be achieved if the probability of an event conditioned by a subgroup is the same for every subgroup.

Definition 1 (Independence Criterion) *A predictor $\hat{Y}$ obeys under the independence criterion given the sensitive attribute A if*

$P(\hat{Y} = y|A = a) = P(\hat{Y} = y|A = b)$ *for all* $y \in \mathbb{N}$ *and for any* $a, b \in A$ *where* $a \neq b$.

For classification problems, the used fairness metric is *Demographic Parity*, also known as *Statistical Parity* which compares the predicted prevalence among subgroups, meaning

[1] Not to be confused with error bias, as in the bias-variance trade-off.

it requires that the probability of the model predicting a positive outcome is the same independently of the subgroup.

Several caveats and assumptions must be considered based on this definition. This criterion is only suitable when the sensitive information does not impact the output of the decision-making process. In the medical field, different conditions are often associated with different races, genders, and ages, so this measure is not sufficient to guarantee that the model is unbiased. However, when the information is irrelevant, one may question why the sensitive trait information was included in the first place. Although it may seem logical, the concept of *fairness through unawareness*, which advocates for the exclusion of sensitive information from the learning phase, has been countered by several arguments because the measure is ultimately ineffective. This is due to the correlation of the sensitive feature with other features, and while singly the correlation may not be significant, when multiple features are combined, the sensitive information can be inferred by the model. Ultimately, a model with the sensitive feature has similar characteristics as a model without it. Moreover, including the information can be beneficial for mitigation countermeasures and increasing model transparency.

Specifically, achieving this criterion through demographic parity can in itself induce some dangerous bias due to the ground truth not being considered. Requiring a model with the same proportion of predicted positives for each subgroup ideally entails a high number of TP, but the same can be achieved by a high number of FP. A glaring issue is when the model is unable to correctly sort the positive outcomes for a specific subgroup, meaning the higher portion of the positive predictions are incorrectly marked with high rates of FP and FN. For this reason, Demographic Parity is often used for cases where the ground truth is untrustworthy or for ensuring equality measures but it is not reliable in most cases.

The separation criterion requires that the error rate among subgroups is the same. In so, a given model is considered fair if, given an outcome, the probability of being correct is the same for each subgroup.

Definition 2 (Separation Criterion) *A predictor $\hat{Y}$ obeys under the separation criteria given the sensitive attribute A if*

$P(\hat{Y} = y|Y = y, A = a) = P(\hat{Y} = y|Y = y, A = b)$ *for all* $y \in \mathbb{N}$ *and for any* $a, b \in A$ *where* $a \neq b$.

Following this rule for both possible outcomes, *Hardt et. al* introduced Equalized Odds (EOdds) as well as less constrained rules for positive and negative outcomes: Equality of Opportunity (EOpp), and Predictive Equality (PredEq), respectively [34].

The sufficiency criterion entails that the model doesn't rely on the sensitive feature to make a prediction. Accordingly, a model is deemed fair if the probability of the model being correct given a prediction is consistent for every subgroup. Statistically, this corresponds to ensuring that the precision is the same across subgroups through a metric known as Predictive Parity (PredP).

Definition 3 (Sufficiency Criterion) *A predictor $\hat{Y}$ obeys under the sufficiency criteria given the sensitive attribute A if*

$P(Y = y|\hat{Y} = y, A = a) = P(Y = y|\hat{Y} = y, A = b)$ *for all $y \in \mathbb{N}$ and for any $a, b \in A$ where $a \neq b$.*

It is important to note that the aforementioned metrics are not compatible. *Garg et al.* studied this topic and concluded that while it is possible to conjugate some of these metrics, it is under unfeasible constraints in most combinations [27, 29]. Hence why the choice of the bias metric depends on the particular problem.

Group Fairness measures prevail due to its low computational cost and high interpretability. While the low complexity may pose some limitations, its simplicity ensures some interpretability. In various points, they are a good starting point for bias analysis. In this work, some of these characteristics were assimilated in what may be considered a more complete analysis. Appendix A presents a summary table with the group fairness metrics.

2.2.1.2 Individual Fairness

Following the limitations of group fairness, the concept of Individual Fairness was introduced, proposing that each sample should be analyzed with similar instances and not as a whole group. This notion addresses how inconsistent the groups can be, hiding reasons why the discrepancies were noted.

Following the concept of CFs employed in the field of explainability, *Russel et al.* introduced the notion of counterfactual fairness, further adapted in consequent works [45, 64]. A model is counterfactually fair if, for any sample, by switching the value of the sensitive feature, assigning a new subgroup, the prediction doesn't flip, formally defined as:

Definition 4 (Counterfactual Fairness [45]) *A predictor $\hat{Y}$ is counterfactually fair given the sensitive attribute A and any observed variables X if*

$P(\hat{Y}_{a\to b} = y|X = x) = P(\hat{Y}_{b\to a} = y|X = x)$ *for all $y \in \mathbb{N}$ and for any $a, b \in A$ where $a \neq b$.*

CFs are employed for causal inference, being closely related to the concept of independence in fairness. This relates to its main caveat - for CF evaluation to be effective, it requires that the feature doesn't contribute to the outcome, which is undesirable in several fields, as aforementioned. To attain this shortcoming, this work employs a methodology to generate plausible CFs through statistical information and domain knowledge. With this, the sensitive feature is not the only feature changed as other correlated features are also altered based on the feasibility of the values for the given group and label. This methodology is expanded upon in Section 4.1.

Individual Fairness is still a disputed methodology. *Will Fleisher* argues the impracticability and insufficiency of the methodology, further defending that the analysis may induce human biases [25].

2.2.2 Bias Mitigation

Depending on the source of bias, there are different approaches to mitigate bias. Some common misconceptions are that increasing the dataset and removing the sensitive feature from the features are reliable ways of mitigating bias.

While adding more samples can be beneficial, it is not always a panacea for eliminating data bias. However, when dealing with representation bias, it may be a valid approach. Additionally, data augmentation techniques embedded with domain knowledge may be useful for addressing other types of bias.

Removing the sensitive feature from the features often leads to more problems in the data. On one hand, because sensitive features can be important for the outcome; for example, in healthcare, gender can be crucial, thus removing it can hinder the results. On the other hand, in a problem where it should not be relevant, removing it would not correct the misassigned samples, and due to correlations between features, the group may still be inferred and lead to biased results [58].

Mitigation techniques focus on manipulating the data to achieve better representations and on constraining the model architecture or predictions, being applied during preprocessing, in-processing, or post-processing.

Preprocessing approaches focus on creating a more balanced dataset. This can be achieved by manipulating the data, such as resampling by group, applying weights to the features, removing correlations, and other nuanced methodologies [16]. These techniques are useful for mitigating data bias and have the advantage of not requiring alterations in the model. However, they may not be as successful as other methods in improving fairness metrics.

In-processing approaches involve adjusting the algorithm to consider unfair behavior and require that at least one of the aforementioned notions of fairness is satisfied. This can be achieved through fairness constraints by adding conditions to the algorithm and optimization problem. For example, by altering the loss function or penalizing instances when the required fairness metric is not met [78]. Another known in-processing technique used with neural networks is "Adversarial Debiasing,"where the model maximizes its performance while minimizing an adversary model tasked with predicting the sensitive feature value [79].

Post-processing methods encompass both white-box and black-box approaches to modify the model once it has learned from the data. White-box methods alter the model's internals by manipulating weights and imposing rules after training. Because they are model-specific and require a deep understanding of the model, they have been less explored, being replaced by in-processing approaches. Black-box approaches alter the predictions through processes akin to post-processing optimization processes [57, 34].

3

State of the Art

Over the years, different libraries and frameworks have emerged for bias detection and mitigation. This chapter addresses current relevant tools, methodologies, and studies regarding this thesis topic.

To ensure transparency, bias-detection methods need to be user-friendly. In this regard, the bias audit toolkit *Aequitas* is a good example of an easy-to-use platform to assess group fairness with clear guided steps and thorough descriptions of the results [65]. The What-If Tool (WIT) is another easy-to-use interface that allows to explore 'what if' scenarios, allowing for a counterfactual analysis of the results [76]. *Amazon SageMaker* is a subscription-based service that provides tools under *Clarify* for detecting bias and monitoring deployed models to detect changes in the distribution of data [33]. For developers, the most used *Python* libraries are *Fairlearn*, a project created by *Fairlearn Organizations* [9]; and *AI Fairness 360*, a toolkit created by *IBM* [8].

Fair GBM is another prominent methodology for bias mitigation through fairness constraints proposed by Feedzai [21]. It is based on the Light GBM, a time-efficient Gradient Boosting Decision Tree (GBDT), introduced by *Ke et al.* [48].

CF generation is a straightforward methodology, though, as with any synthetic data, the validity of CFs is an important analysis in any study that employs them. With the intent of increasing robustness in *CF generation*, it is essential to ensure the CF's plausibility [5]. In the context of CFE, there are several studies for generating plausible counterfactuals employing metrics such as *sparsity* and *diversity* [69], as well as improving the *consistency* of the generated CF [11]. Diverse Counterfactual Explanations (DiCE) is a prevalent technique for generating CFs that satisfies two properties: diversity and feasibility given the user context and constraints, [53]. The method is able to generate a diverse set of counterfactuals that effectively approximate local decision boundaries providing meaningful insights into the model's predictions. *Human-in-the-Loop* processes are also prominently employed to generate CFs, improving robustness by not solely relying on data distributions and employing domain expertise [37].

However, based on the definition 2.2.1.2 of counterfactual fairness, the research is not as vast in the current literature. The accumulated knowledge is not transferable between

application to CFE and counterfactual fairness due to the essentially opposite goal of the approaches. The goal of a CFE generation algorithm is to find the closest (plausible) CF in which the prediction flips, whereas the goal of CFs in counterfactual fairness is to flip the sensitive feature and only then verify if there is any change in the prediction. For this reason, generation models for CFE have a clear optimization goal, flipping the prediction, and fairness'CFs do not. The work of [10] presents a notable contribution in this regard, introducing 'fliptests' that account for flips from positive to negative or vice versa for each subgroup, further constraining by features as well as ground truth based on the case study. However, the application of this evaluation focuses on understanding which features changed in the CF generation process instead of measuring model prediction's flips. The generation process uses Optimal Transport Maps (OTMs) or approximations through Generative Adversarial Networks (GANs) to convert distributions between subgroups, effectively achieving feasible CFs. However, these methods come with shortcomings, they are computably demanding, and, while effective in reading and learning the distributions, easily prone to replicating statistical relations that are not necessarily desirable or even representative of real-world populations. This is, however, used to their advantage, by identifying the features that change the most for a given *fliptest* to understand underlying tendencies the model may use to conduct its predictions. *Goethals et al.* propose another metric based on these flips, the *PreCoF* which measures the difference in portion of these flips [31]. Albeit with similar methods and goals, this thesis work aims for a more straightforward evaluation to which plausible CFs are fundamentally required, so it was developed a controlled CF generation process, described in Section 4.1.

Motivated by ensuring continual auditing *Maughan et al.* devised a method to evaluate CF after deployment, suited for Neural Networks (NN) [49]. The proposed metric 'Predictive Sensitivity' measures the dot product of the gradient of the feature's contributions in the task classifier and in predicting the protected attribute (auxiliary model). Both factors are calculated by the difference in weight of the feature when flipping the protected attribute. This measure is proposed as a proxy to causal inference and higher values indicate CF bias. During training, a base acceptable value for predictive sensitivity is established, and, after deployment, predictions are monitored to detect if the model is striving away from the acceptable range. This method does not require an explicit CF generation process being a cost-efficient approach, although training an additional model. As restrictions, this method requires white box access to auditing and is limited to NNs.

Russel et al. proposed a method to make fair predictions by integrating multiple causal models [64]. In that study, the model is generated with an optimization task of achieving, as the authors define it, an *Approximation Counterfactual Fairness*. This metric is calculated based on the difference of flipped instances between each sequential causal model. In the context of data augmentation, several studies explored the benefits of utilizing CFs for solving attenuating the class imbalance [59, 74], as well as testing which generation processes ensure better results [38].

4

Methodology

According to the goals established in Section 1.2, this chapter describes the developed methodology for generating plausible CFs, followed by the proposed framework for bias analysis and mitigation. To conclude, it presents relevant insights into the used datasets and the considerations taken when employing the aforementioned processes.

4.1 Plausible Counterfactuals Generation

This section describes a method for CF generation that addresses the limitation posed by *Russel*'s CFs, wherein the outcome must remain independent of the sensitive feature.

The goal of this process is to generate a corresponding CF that belongs to the same outcome, additionally changing characteristics highly correlated to the sensitive feature under analysis. For instance, inherent biological differences between men and women imply that a mere change in the feature 'sex' might be insufficient for crafting a plausible CF. The proposed algorithm considers the statistical distribution of each group of the sensitive feature relative to every possible outcome within the training set, leveraging conditional logic to adjust features.

The algorithm starts by handling continuous variables. The process (considering only one label) is roughly illustrated in Figure 4.1. In a binary classification task, and without loss of generalization, it begins by computing the Cumulative Distribution Functions (CDFs) $P \in \{P_{A-}, P_{A+}, P_{B-}, P_{B+}\}$ of values v derived from the continuous feature f_i of the training set X, for each subgroup $g \in \{A, B\}$ and label $y \in \{0, 1\}$, where 0 corresponds to the negative(-) class and 1 to the positive(+) class. For a sample s_i in the sample space S from subgroup A that has value $v_i \in v$ for a continuous feature f_i and label $y = 1$, the goal is to generate a CF in the subgroup B with a value $v_i' \in v$, while retaining the label $y = 1$. First, the algorithm assesses the probability of the event $v \leq v_i$ within the P_{A+} distribution (Figure 4.1B). Formally, for sample $s_i \in S_{A+}$, and the closest not exceeding feature value $v_{i-1} \in X_{A+}$, the probability of the event v_i is given by Equation 4.1.

$$P_{A+}(v_i) = P_{A+}(v_{i-1}) + (P_{A+}(v_{i+1}) - P_{A+}(v_{i-1}))\frac{v_i - v_{i-1}}{v_{i+1} - v_{i-1}} \tag{4.1}$$

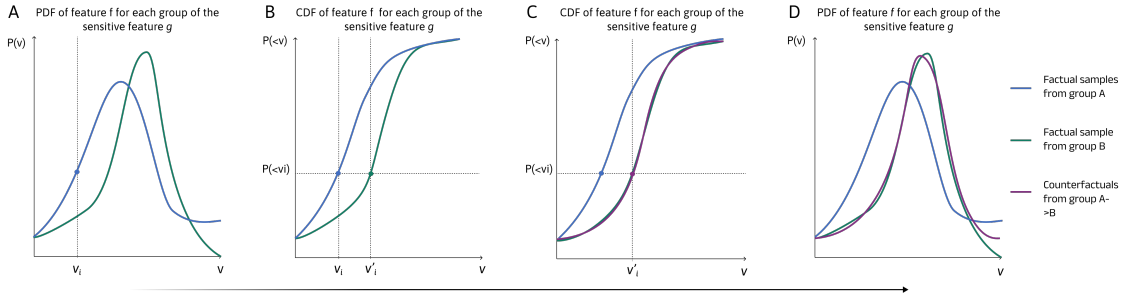


Figure 4.1: Sequential steps for adjusting the value v_i of the continuous feature f_i throughout the generation of CF from group B from samples from group A, given the sensitive feature $g \in A, B$. Note that the process further sections the samples by label.
A - Estimate PDF of the continuous feature sectioned by subgroups and label; B - Estimate CDF for both distributions marking the corresponding value v'_i in the other group CDF; C - Resulting CF CDF(purple) when repeating the process for every sample; D - Resulting PDF from the CFs(purple).

Then, it finds the corresponding value v'_i in the distribution P_{B+} from $P_{A+}(v_i)$ (Figure 4.1C), formally defined in Equation 4.2. When faced with values or probabilities not represented in the training set, the algorithm employs interpolation to calculate the approximate value or probability.

$$v'_i = v'_{i-1} + (v'_{i+1} - v'_{i-1}) \frac{P_{A+}(v_i) - P_{B+}(v'_{i-1})}{P_{B+}(v'_{i+1}) - P_{B+}(v'_{i-1})} \tag{4.2}$$

For categorical features, a similar approach is adopted. However, when finding the corresponding feature values of the CF, instead of interpolation, the method selects the nearest value from the new distribution, considering the discrete nature of the feature. It is important to note that this process is specifically suited for features - whether continuous or categorical - that exhibit a sequential relationship among values, such as 'age' or 'burn degree'. For other categorical features, it is preferable to one-hot encode them and treat them as binary. Importantly, this method is able to treat encoded features as a group, ensuring the absence of impossible combinations. To provide a clearer picture with an example, consider the task of generating a CF of a male patient based on a healthy female patient's data. It would be implausible if the testosterone levels in the CF remained consistent with the original female patient's levels, given testosterone's role as a sex hormone. A reduced testosterone level, for instance, might be associated with a particular health condition. In addressing this, the proposed method finds the testosterone concentration of the initial female sample within its distribution, say, within the lower quartile. It then identifies an analogous value in the male distribution that also falls within the same quartile. This ensures that females with comparatively lower testosterone levels correspond to males with similar relative levels, and the converse holds true.

The process of flipping binary features is based on the probability of the original

value, conditioned for each group. Initially, for each binary feature, the method tests the probability of the original feature value v_i occurring in the new group. Should this probability fall beneath a predetermined threshold, then the feature value is flipped. This step prevents the emergence of 'impossible' feature combinations. For example, consider an original sample representing a pregnant woman, aiming to generate a male CF. Since the attribute of pregnancy is intrinsically female-exclusive, the probability of a pregnant individual being female is 1, while the counterpart for a male is 0. Consequently, the algorithm modifies the feature value from 'pregnant' to 'not pregnant' in the male CF.

Subsequently, an iterative test for the remaining binary features initiates. For a sample s labeled as 1 from subgroup A with value v_i for the binary feature $f_i \in F$, where F is the set of binary features, the method calculates the difference between the probability of that event occurring for subgroup A and B. If the difference surpasses a predefined threshold (0.5 by default) the feature value flips, remaining unchanged otherwise. In notation,

$$v_i' = \begin{cases} |v_i - 1| & \text{if} |P(v_i|g = A, y = 1) - P(v_i|g = B, y = 1)| \geq 0.5 \\ v_i & \text{if} |P(v_i|g = A, y = 1) - P(v_i|g = B, y = 1)| < 0.5 \end{cases} \tag{4.3}$$

Acknowledging that this change might not be plausible, the procedure iteratively refines itself, incorporating additional constraints on subsequent feature flips. In the second level where f_i flipped, other features are tested, comparing probabilities conditioned on the new feature value and on the sensitive feature.

$$v_j' = \begin{cases} |v_j - 1| & \text{if} |P(v_j|g = B, f_i = v_i', y = 1) - P(v_j|g = B, f_i = |v_i' - 1|, y = 1)| \geq 0.5 \\ v_j & \text{if} |P(v_j|g = B, f_i = v_i', y = 1) - P(v_j|g = B, f_i = |v_i' - 1|, y = 1)| < 0.5 \end{cases} \tag{4.4}$$

As the iterations progress, the pool of samples adhering to these constraints diminishes. Consequently, the probabilities are less credible with respect to the representability of the population. Thus, it is crutial to consider the size of the dataset and select an appropriate *depth* - an hyperparameter determining the maximum number of iterations. Because there are features that can be established as unrelated to the sensitive feature, it is possible to select the potential features and keep the others unchanged. The main components of this methodology are comprised in Algorithm 1, found in Appendix B.

This approach can yield multiple CFs for each sample as it evaluates each feature in parallel, depending on the hyperparameters.

4.2 Counterfactual Analysis

Albeit the extensive works following the notion of counterfactual fairness, there was not a concise set of metrics easily employed for an in-debt global analysis of where and why a decision was flipped [71]. The analysis seems to be constrained to how many samples flipped, not taking into account in which context these flips occurred. For instance,

knowing if the model tends to flip the outcome from positive to negative at a high rate for a certain group paints a clearer picture of potential biases. Similarly, if originally incorrectly predicted samples are now correctly predicted by switching the subgroup, it may indicate the model is better adapted to the new subgroup. There is valuable information in knowing which samples have flipped the outcome.

4.2.1 The Counterfactual Confusion Matrix

The prominent CM allows for easy visualization of the combinations of the ground truth and the predictions, containing all the necessary information for evaluating the performance of a classification problem. Inspired by this instrument, the proposed Counterfactual Confusion Matrix (CCM) adds two dimensions to the analysis: the sensitive feature values and the resulting prediction for the counterfactual samples. As follows, to reduce complexity in more linear cases, there are two versions of the matrix, one that does not include the ground truth, and an extended version.

Recalling the definition of counterfactual fairness defined in 2.2.1.2, a model is deemed fair if the prediction of a CF is the same as the prediction of the original sample, for each CF generated by flipping the value of the sensitive feature of each sample in the dataset. Because of the assumption of independence when generating the CFs, considering the pair $\{s, s_{CF}\}$ for a sample $s \in S$, the true value for each element of a pair is the same. Thus, it is possible to directly compare each pair with a criterion of **consistency**, evaluating if the prediction changes for the CF sample. There are four possible combinations of results for a binary problem, which can be counted and presented just like the combination between predictions versus real values can be summarized in a CM. In this case, the reference is the original prediction, so it occupies the spot of the ground truth in rows, while the CF predictions are placed in the columns, as illustrated in Figure 4.2.

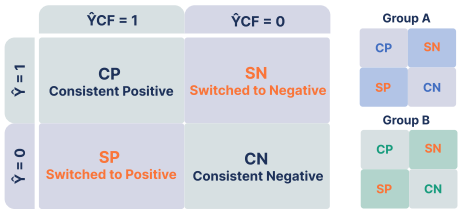


Figure 4.2: The Counterfactual Confusion Matrix.

When analyzing a counterfactual matrix, the base parameters account for the CFs predicted as positive (P) or negative (N) which are consistent (C) with the original sample or switch (S) its value following the nomenclature: Consistent Positive (CP), Consistent Negative (CN), Switched to Positive (SP), and Switched to Negative (SN).

Once again drawing the comparison with the CM, there are several metrics that can be extrapolated. Parallel to 'accuracy' in the CM, the model can be evaluated based on the overall consistency of the predictions, $ConsistencyRate(CR) = \frac{CP+CN}{CP+CN+SP+SN}, CR \in [0, 1]$.

This metric is already used in the literature as it measures the portion of the samples that have the same prediction that the corresponding CFs [20]. Nevertheless, since in bias evaluation the goal is to detect undesired behavior (switches in the prediction), the complementary metric $SwitchRate(SR) = 1 - CR, SR \in [0, 1]$ is more relevant for this task. As a general intuition, a counterfactually fair model has a CR close to 1. Especially if the model has great performance, it is essential that this coefficient is high. For instance, if the model has good performance and no apparent disparity displayed by the group fairness metrics, a high SR indicates the model is 'overfitted' to the specific distributions of the subgroups studied as represented in the dataset. It can serve as a proxy to identify if the dataset has a sufficient representation of the population. It is important to note that these interpretations have to be followed by careful data analysis to attest to the theory. CR and SR have the same shortcomings of accuracy. While accuracy has limited application for unbalanced datasets, CR and SR are limited by the positive and negative proportion of the original predictions. For this reason, the *Counterfactual Matthew's Correlation Coefficient (CMCC)* $= \frac{CP \times CN - SP \times SN}{\sqrt{(CP+SP)\times(CP+SN)\times(CN+SP)\times(CN+SN)}}, CMCC \in [-1, 1]$, is more robust to different portions of positives and negatives, sustaining a similar interpretation as the CR.

Building from CR and SR, it is important to access in which direction the change occurs, from negative to positive or from positive to negative. In practice, this can be calculated through ratios that mirror the TPR and TNR. The ratio that indicates the portion of samples originally predicted as positive for which the CFs are predicted as negative, $NegativeSwitchRate(NSR) = \frac{SN}{SN+CP}, NSR \in [0, 1]$, and as positive, *Positive Switch Rate (PSR)* $= \frac{SP}{SP+CN}, PSR \in [0, 1]$, indicate the tendencies the model has when predicting different groups. These metrics are more useful when calculated for each subgroup, similarly to the strategy employed by group fairness parities, using the notation $a \rightarrow b$ to note the original group a and the new group in the CF sample. Thus, $\text{NSR}_{a \rightarrow b}$ indicates the portion of samples from the group a initially predicted as positive for which the CFs (group b) are predicted as negative. Intuitively, high values for this metric suggest that the model associates group b with negative outcomes. And just like group fairness, it is valuable to compare them to properly understand if the model is biased or simply unable to handle new samples, introducing CF parity measures. To illustrate, considering the task is to evaluate bias based on the sex of the subjects, m for men and w for women, that has $PSR_{w \rightarrow m} = 0.40$, the interpretation would be the model has a tendency to associate men with positive outcomes, however, if $PSR_{m \rightarrow w} = 0.35$, meaning the difference is only 0.05, then the conclusion is not as straightforward. In this case, it is necessary to consider other CF metrics which will be discussed further in the next section, that consider the overall performance of the model.

Analogously to the PPV, the $PositiveConsistentPrecision(PCP) = \frac{CP}{CP+SP}, PCP \in [0, 1]$ assesses the portion of CFs predicted as positive that are consistent with the original corresponding samples. This metric is relevant in tasks where an increased number of positives is undesirable. For example, if FP has a higher cost than FN, if the model predicts

if patients are ready for discharge, PPV is the preferred metric. When using ML to assess the compatibility of an organ donor, FPs leads to the organ being rejected, affecting the donor, the recipient, and other potential recipients that would be actually compatible [32]. If a model has great performance for this task but a group has high PCP then it can be assumed the model is ill-suited to the new subgroup.

4.2.2 The Extended Counterfactual Confusion Matrix

The previous counterfactual confusion matrix does not include ground truth information. Nonetheless, there is a valuable insight in exploring if there are more switches in originally true or false predictions. In the proposed Extended Counterfactual Confusion Matrix (ECCM), the model's predictions are compartmentalized in the columns by prediction of the original sample and prediction of the counterfactual sample. For a column where $\hat{Y} = 0$ and $\hat{Y}_{CF} = 1$, each cell can be read as 'samples initially predicted as 0 that were predicted as 1 once the value of the sensitive feature was flipped'. The sum of each column corresponds to the values of the CCM.

The ground truth is shown in the rows, as depicted in Figure 4.3. Optionally, one can include the combination with the values for the sensitive feature. For example, the cell where $Y = 0$, $Group = A$, $\hat{Y} = 0$, and $\hat{Y}_{CF} = 1$, can be read as 'samples labeled as 0 from the group A initially predicted as 0 and predicted as 1 once switched to the group B'.

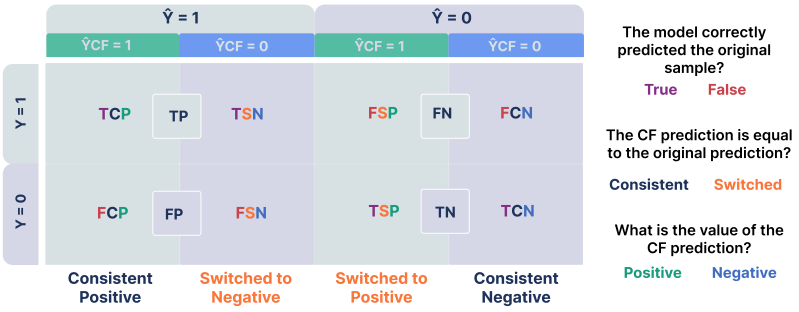


Figure 4.3: The Extended Counterfactual Confusion Matrix.

The inclusion of the ground truth allows to understand in which instances the model has most difficulties. For instance, if the switches occur mostly on initially correctly predicted samples, it indicates the model is not well adapted to the new (counterfactual) group. But if the changes occur in initially incorrectly predicted samples, then the conclusion is the exact opposite. To measure this, the ECCM allows to derive new metrics for each quadrant of the underlying CM. The base parameters convey if the prediction for the original sample is correct (T) or not (F), if the prediction is positive (P) or negative (N), and if the CF prediction switched (S) or is consistent (C). Thus, there are eight combinations for each group: True Consistent Positives (TCP), False Consistent Positives (FCP), True Consistent Negatives (TCN), False Consistent Negatives

(FCN), True Switched Positives (TSP), False Switched Positives (FSP), True Switched Negatives (TSN), False Switched Negatives (FSN). From them, it is possible to measure the portion of samples that switch from true to false, $TrueNegativeSwitchRate(TNSR) = \frac{TSP}{TSP+TCN}, TNSR \in [0,1]$ and $TruePositiveSwitchRate(TPSR) = \frac{TSN}{TSN+TCP}, TPSR \in [0,1]$, or from false to true, $FalseNegativeSwitchRate(FNSR) = \frac{FSP}{FSP+FCN}, FNSR \in [0,1]$ and $FalsePositiveSwitchRate(FPSR) = \frac{FSN}{FSN+FCP,FPSR} \in [0,1]$. Once again compartmentalizing by group, these ratios allow to identify the tendencies the model follows, and eventually assess which kind of samples need to be added to the dataset. For example, a high $\text{TPSR}_{a \to b}$, indicates that a significant portion of samples from group a, which were correctly predicted as positive, had their corresponding CFs predicted as negative. This suggests that the model may not have been trained with sufficient data from group b where the ground truth is positive. These metrics are usually paired by outcome, TPSR and FPSR, and TNSR and FNSR. If, for example, TPSR and FPSR are both high, it may indicate that the model associates the new group to negative outcomes, serving as a proxy of NSR. If only the TPSR is high, with a suboptimal original recall (large portion of FPs), it may correspond to a low NSR. This case is noteworthy as it exemplifies that the model is not able to correctly predict samples that originally it would, due to the change to a new group. This suggests that the new group may fall outside the typical representation found in the original dataset. It is also important to note that these pairs have opposite bias interpretations: TPSR and TNSR point to bias against the new group, as the model was not able to correctly predict the CFs; whereas FPSR and FNSR indicate bias against the original group, as changing the group was sufficient for the model to change the predictions to a correct label.

Additionally, $TrueSwitchPositiveRate(TSPR) = \frac{TSP}{TSP+FSP}, TSPR \in [0,1]$ and *True Switch Negative Rate (TSNR)*= $\frac{TSN}{TSN+FSN}, TSNR \in [0,1]$ measure the portion of originally correct instances that switch in relation to all the switched to positive or negative samples. These metrics are useful to understand if the switches are benefiting, or impairing the original group. High TSPR and TSNR indicate bias against the new group while low values point to bias against the original group, following the same reasoning applied to the TPSR versus FPSR interpretation. However, this pair is for seldom use because it is limited by the dataset outcome distribution as well as the original performance. For instance, a good model has significantly less FPs than TPs, thus the portion of instances that switch from TP, TPSR, may be considerably lower than the FPSR, while the summative result of TSN may be much higher than FSN. However, knowing these limitations and in context with other metrics, both can provide some insight.

The proposed metrics cover the analysis for binary sensitive features, although they can also be used for categorical ones. In this instance, one can approach the problem not only pair by pair but also in an 'one-versus-all' strategy. For example, given three groups, A, B, and C, to study bias, may be relevant to study the switches from $A \to B$ and $A \to C$, as well $A \to B \vee C$. Metrics that suggest bias against the original group, e.g FPSR, are suited for the $A \to B \vee C$ case, while metrics that suggest bias against the new group,

e.g. TPSR, are better suited to pair distinction, $A \rightarrow B$ and $C \rightarrow B$, or the inverse of the 'one-versus-all' strategy, $A \vee C \rightarrow B$. In these situations, the number of generated CFs increases, leading to a higher computational cost. For g groups, a dataset with N samples generates at least $(g-1) \times g \times N$ CFs.

Similar to the CM performance parameters, these metrics are also applicable to multiclass problems. Akin to sensitive features with multiple groups, it may be interesting to group outcomes to certain analyses. Considering three possible outputs I,K, and J, metrics that evaluate clear tendencies to a specific output, e.g PSR, are suited for the 'one-versus-all' strategy, in this case, exploring the portion of samples that were predicted as K or J but switched to I. If the behavior is attested from both K and J groups, it better supports the theory that the model associates the new group with I than if it only occurs for samples initially predicted as K, but both are important considerations.

The ideal ECCM represents a case where every CF is consistent with its original sample. Visually, this corresponds to having both the inner columns empty. However, an ideal ECCM does not necessarily correspond to an ideal CM, or vice-versa. An ideal CM corresponds to an ECCM with the FCP, FSP, FCN, and FSN cells empty. An ideal model in both criteria is represented by an ECCM which only has values in the TCP and TCN cells. These scenarios are represented in Figure 4.4.

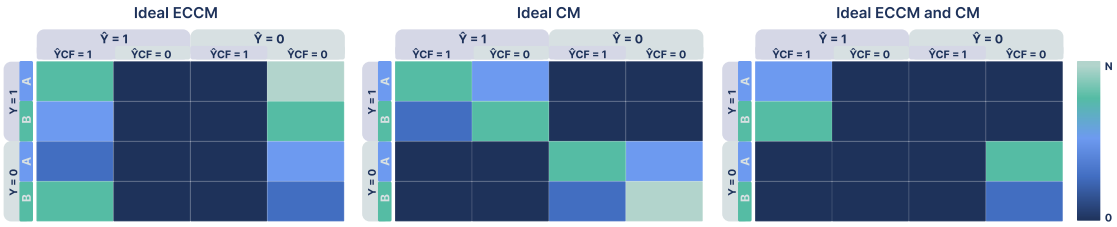


Figure 4.4: Edge cases of ECCMs: (1) for a not ideal model for performance but that is CF fair; (2) of an ideal model for performance (ideal CM), but not in terms of CF fairness; (3) for an ideal model both in terms of performance and CF fairness.

The presented metrics fail to evaluate the confidence the model has in the model predictions, both original and CF. Most classification models output a probability score that is then converted to a discrete output. The closer the sample score is to the prediction threshold, the less certain the model is about the prediction. When evaluating bias it may be valuable to compare the probability distributions of the groups to the probability distribution of the corresponding CFs. For instance, a model that is CF ideal may be less certain for CFs than the original samples, which in itself may indicate some bias.

Considering that an unbiased model is not affected by group membership, the probabilistic distribution of a group and its CFs should overlap. Taking this into account, using metrics such as the Jensen-Shannon Counterfactual Divergence (JSCD), allows to measure the distance between the distributions and assess the similarity in the predictions. Additionally, establishing the original group scores as the reference, analog to the Root

Mean Squared Error (RMSE), the Root Mean Squared Counterfactual Differences (RMSCD)= $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(s_i - s_i')^2}$ measures the (root) cumulative 'error', i.e., the score difference for each sample i of the original, s_i, and the CFs, s_i' sample.

A summary of the proposed metrics can be found in Table C.1 in Appendix C.

4.3 Use Cases

Although the proposed method is model-agnostic, there are several important considerations regarding the used data. Here, the datasets used to test the methodology are presented, establishing the goals of the use case and highlighting important details for a thorough model evaluation.

To ensure a fair evaluation of each model's architecture and have more robust metrics, it was applied a cross-validation technique. For each model, the dataset was sectioned in five train-test partitions, known as folds. Each fold was used to train the model, essentially generating 5 models. The evaluation process combines the results of the five test sets. Because the test sets don't overlap from each fold to the other, the resulting matrixes are obtained summing the individual matrixes of each fold. This way, the dataset is all used for testing without occurring leakage with the training set. The larger volume of samples allow to more reliable conclusions of the methods as a whole.

4.3.1 CardioFollow.AI

CardioFollow.AI is a project aiming to provide continuous remote care to patients post-hospital discharge from cardiac surgery [62, 66, 22]. One of the tasks of interest is to predict consequent health complications. With this goal, a dataset was created with patients' information and registered complications.

For the task under study, a higher weight is placed on positive outcomes. FPs, while undesirable, primarily result in resource allocation and potential patient anxiety. These outcomes generally have minimal long-term impact on patient health. In contrast, FNs are of greater concern, as they can lead to severe health consequences, including life-threatening situations, due to delayed intervention. As such, the primary objective is to optimize the predictive model's ability to identify positive outcomes promptly, enabling the timely implementation of preventive measures to safeguard patient health. This translates into finding the best TPR to minimize FNs.

The dataset includes general demographic information such as age, sex, height, and weight; preoperative risk factors such as Body Mass Index (BMI) and Body Surface Area (BSA), smoking history, diabetes, hypertension, and hypercholesterolemia; pre-existing renal, pulmonary, vascular, and neurological pathology or dysfunction; procedural specifics including cardiac rhythm, number of diseased coronary arteries, and cardiac catheterization results; the patient's condition at the time of the procedure, the urgency, the reason for the procedure, the type and specifications of the procedure, and the type of

implant; morbidity scale assessment score summarizing the complications experienced; and general chronological information regarding the duration of the procedure, the pre- and post-operative periods, and the time since the last consultation. Although the original label distinguishes between complications, the target variable represents only the presence/absence of postoperative health complications. The dataset comprises 5649 samples, of which only 418 represent patients who suffered from complications. The dataset is considered unbalanced with a distribution of 0.074 positives to 0.926 negatives.

Currently, there is a deployed model optimized with a minimum set of eight features comprising total time of stay (1), of that time: in intensive care (2); time to surgery (3) time from surgery until discharge (4); number of complications (5); time in bypass (6); implant size (7); and type of diabetes treatment (if any) (8). The time measurements are adequate proxies for inferring urgency and how the patient recovered, whereas the other features serve to assess the gravity of the occurrence. While other parameters are relevant for this task, due to the relatively small dataset for the total possible combinations of risk factors, procedures and complications, through careful experimentation with SHAP values, these are the features that paint the best overall scenario.

The base algorithm used is Random Forest. After training, a post-processing optimization technique was applied, to ensure the best possible recall. This process, *Threshold Optimization*, creates several data subsets and, from the probability score, establishes the optimum value for recall within acceptable values for precision. The experiments in this work are based on this model's architecture.

4.3.2 Heart Disease

Heart Disease is a public dataset from the UCI Machine Learning Repository [6]. It contains clinical and non-invasive test records from 303 patients at the Cleveland Clinic, 425 patients at the Hungarian Institute of Cardiology, 200 patients at the Veterans Administration Medical Center in Long Beach, California, and 143 patients from the University Hospitals in Zurich and Basel, Switzerland [24]. The goal is to predict Coronary Artery Disease (CAD) attested by angiography results, graded by severity from 0-4 following the Coronary Artery Disease Reporting and Data System (CAD-RADS) [44]. An alternative goal is to simply detect the presence/absence of vessel occlusion. The patients went through three non-invasive tests: exercise electrocardiogram, thallium scintigraphy, and cardiac fluoroscopy. The dataset comprises 66 features, from demographic information, 'age' and 'sex', risk factors, family history, diabetes, cholesterol, and smoking history. Personal identifiable data, as the name, ID and social security number were excluded.

During preprocessing, the dataset was substantially reduced to 581 samples and 38 features, due to a high rate of missing values. To grant better results, the task under analysis is the binary identification of CAD. Because FNs sustain a higher risk for the patients than FPs, the preferred performance parameter is recall (TPR). The prevalence of the disease in the processed dataset is 0.629, displaying a slight unbalance between labels.

5

Results & Discussion

This chapter describes the most relevant experiments to support this work's conclusions. Each experiment is described and discussed following the framework for analyzing a dataset, constricted by its particularities and context application. For this reason, the sections are organized by use case.

5.1 CardioFollow.AI Dataset

The first step for analyzing bias in the dataset is selecting the sensitive feature to investigate. The first experience considers a protected attribute, the 'sex' of the patient. Further experiments explore biases relating to the risk factor 'smoking'.

5.1.1 Bias for the sensitive feature 'sex'

The proposed method for generating CFs (c.f. Section 4.1), ensures minimum changes in the original sample. However, it mostly replicates the already preexisting patterns in the original dataset. In this section, two sets of generated CFs are analyzed: one in which all the features have the potential to change, and another with only feasible changes, based on Domain Knowledge. In this experience, instead of the deployed model described in Section 5.1, a model following the same architecture but trained with all the extracted features is used. This is done to include features directly related to the sensitive feature, such as 'BMI', 'BSA', 'Weight', 'Height', and 'Creatinine'. If not included, the generated CFs incur the risk of not suffering any alterations and, for this reason, there are no possible interpretations to draw from. The training set is built from the contributions of 1557 women and 2416 men with approximately the same prevalence, approximately 7%. In addition to comparing different CFs sets for bias analysis, different augmentation approaches are studied. The different metrics for the resulting models are summarized in Table 5.1 to allow an easier discussion. Each metric is extrapolated from the sum of the five matrixes correspondent to each fold generated from cross-validation. These sets are validated through real-world knowledge.

Analysing Bias - CFs without Domain Knowledge

Before delving into CFs, the first step is to evaluate the base model performance and group fairness metrics. In this specific task, where the TPR is optimized, the most suited group fairness metric is PredP. The base model has an overall TPR of 74.6%. Sectioning by the sensitive feature, the male subgroup has 72.3% TPR and the female subgroup has 78.0% TPR. In terms of PredP this corresponds to a $5.7p.p.$ difference, indicating slight bias against men. As for the negative outputs, the overall TNR is 48.8%, meaning the model predicts just as much FP as TN. This statistic is relatively consistent in both sexes with 49.4% for males and 47.8% for females.

Evaluating the counterfactual fairness of the model with CFs generated without Domain Knowledge, there are more prediction changes when flipping male to female than the other way around, $SR_{M\to F}$ = 11.7% and $SR_{F\to M}$ = 8.5%, and there is better consistency regardless of the outcome as well, as seen by $CMCC_{F\to M}$ = 83.2% against $CMCC_{M\to F}$ = 77.2%. By itself, this is indicative of a more considerable bias in female patients, but other metrics support a tendency to associate female patients with positive outcomes, such as $PSR_{M\to F}$ = 19.7%. A large portion of these positive flips are correct predictions, $FNSR_{M\to F}$ = 36.7%, suggesting that more than one-third of the FNs predicted in male samples were correctly predicted as TPs, once changed to CF females. This seems to be counter-intuitive as it would be expected that changing from men to women would lead to less FNs, to match the original female TPR. Nonetheless, this result can unveil bias that would only be revealed with new predictions on samples less similar to the training dataset. However, it is important to note how these CFs were generated. Because they were created without considering Domain Knowledge, the modifications incurred tend to display the typical sample of the other subgroup. For example, one heavily correlated feature with sex is 'Smoking', a risk factor much more common in men than women in this dataset. 47.0% of men and only 9.9% of women are smokers, meaning 88.0% of smokers are men. Based on these statistics, when generating a CF for a male smoker, the result is most likely a woman who does not smoke. Logically, 'Smoking' should be unrelated to the sex of the patient and, as an addictive trait and risk factor, it should be kept unchanged in the CF.

Analysing Bias - CFs with Domain Knowledge

Considering this setback, by resorting to Domain Knowledge, the second set of CF was generated only allowing changes in the features that should be related to the patient's sex. The existing preconditions and risk factors should be left untouched since they are the most critical points of similarity in this task. The main features that can be justifiably switched are related to general biological differences between men and women. Men tend to have lower fat storage, and higher density bones and are generally taller than women [67, 60]. Men tend to have higher muscle mass, resulting in slightly higher creatinine levels than women. Creatinine is a waste product produced by muscle metabolism and excreted by the kidneys and high levels can indicate kidney disfunction. Thus, for the same blood concentration, women may face a more severe condition than men [1]. The potential

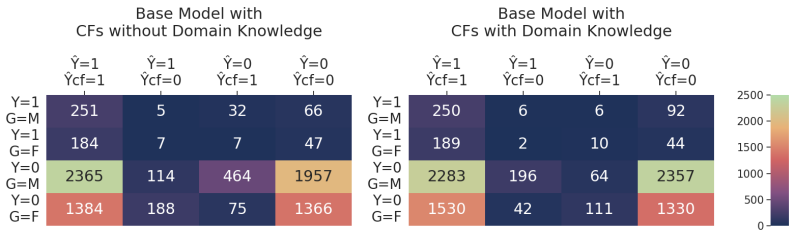


Figure 5.1: ECCM for the base model with CFs generated with (right) and without (left) domain knowledge.
Y - Ground Truth (0 for 'recovery without complications' and 1 for 'post-surgery complications'); G - Group (M for 'male' and F for 'female'); $\hat{Y}$ - Prediction for the factual(original) sample; $\hat{Y}_{CF}$ - Prediction for the CF sample.

features for generating the CFs are 'BMI', 'BSA', 'Weight', 'Height', and 'Creatinine'.

Contrary to the tendency displayed with CFs generated without Domain Knowledge, this set reveals a tendency to associate male samples with positive outcomes, given by $PSR_{F\rightarrow M}$ = 8.1%, with a significant portion being from switches from FNs to TPs, $FNSR_{F\rightarrow M}$ = 18.5%. At the same time, it is noted a tendency to associate negative outcomes to female samples, $NSR_{M\rightarrow F}$ = 7.4%. This is the exact opposite of the other set of CFs, emphasizing the importance of a careful CF generation process. Assuming these CFs as plausible, these results point to a slight bias against women. To attempt to mitigate this issue, the dataset was augmented with the CFs generated for the train set in an attempt to teach the model to better understand the differences in how to predict female and male samples.

The ECCM for CFs with and without Domain Knowledge can be visualized in Figure 5.1, where one can visually observe less samples in the middle columns when Domain Knowledge is used, suggesting lower counterfactual bias.

Mitigating Bias - Data Augmentation with CFs

Augmenting the dataset with CFs has the purpose of teaching the model how it should interpret each instance of symptoms and characteristics in the case of the patient being male or female.

Applying the method to the training set did not achieve better results in the criteria of performance, slightly decreasing the TPR at 73.1% and only increasing TNR by 1.9$p.p.$ to 50.7%. It was able to mitigate some of the discrepancies between men and women in TPR, TPR_M = 71.8% and TPR_F = 75.1%, resulting in a $PredP$ = 3.3$p.p.$. However, the difference is so subtle, it poses doubt about its actual improvement, especially considering the drawbacks of data augmentation. Analyzing the CFs changes, the method is proved as not effective in this particular instance. The model displays a higher tendency to associate male samples with positive outcomes $PSR_{F\rightarrow M}$ = 10.0%, with a larger portion occurring from TNs to FPs, $TNSR_{F\rightarrow M}$ = 9.6%. On the other hand, the tendency to associate female samples with negative outcomes seems to be more prominent as well, $NSR_{M\rightarrow F}$ = 11.7%, with a higher portion resulting from TPs to FNs, $TPSR_{M\rightarrow F}$ = 3.2%. This is an unexpected

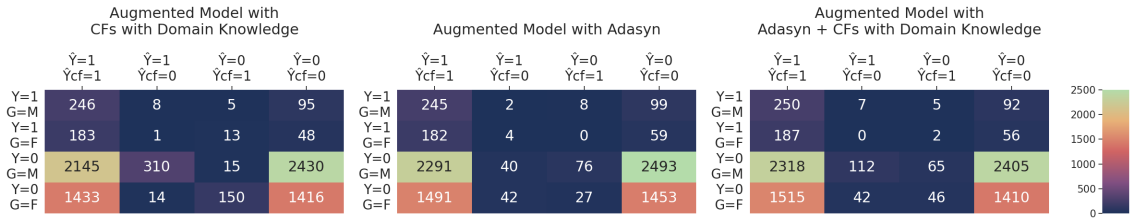


Figure 5.2: ECCM for base model trained on the CardioFollow.AI data, considering 'Sex' as sensitive feature, and using data augmentation with CFs(left), ADASYN(middle), and a combination of both(right).
Y - Ground Truth (0 for 'recovery without complications' and 1 for 'post-surgery complications'); G - Group (M for 'male' and F for 'female'); $\hat{Y}$ - Prediction for the factual(original) sample; $\hat{Y}_{CF}$ - Prediction for the CF sample.

result as it would be expected that augmentation with CFs would decrease CF bias. Although the metrics' values do not infer substantial bias, this experience reveals that, based on context, including all the CFs may not be beneficial.

Mitigating Bias - Data Augmentation with Adaptive Synthetic Sampling (ADASYN)

It is established the potential problem of this task lies in insufficient samples, hinting at augmentation as a viable option. There are several techniques for data augmentation, and the ADASYN method is well-regarded for its ability to oversample less populated distributions by using density-based sample saturation to generate additional synthetic samples [35].

Proceeding with this method, the overall performance remained similar, but the group performance became more discrepant in terms of TPR, $TPR_M = 69.8\%$ and $TPR_F = 75.9\%$, $PredP = 6.2p.p.$ and TNR, $TNR_M = 52.4\%$ and $TNR_F = 49.1\%$. Nevertheless, there are fewer switches in the CF predictions, $CMCC = 95.3\%$. The $NSR_{M\to F} = 1.6\%$ and the $TPSR_{M\to F} = 0.8\%$ are considerably lower. This method led to a worse equilibrium between male and female samples in terms of group fairness. It is important, however, to retain the properties that allowed for better CFs parameters.

Mitigating Bias - Data Augmentation with CFs oversampling with ADASYN

Trying to improve the overall performance, the dataset was augmented with CFs and then oversampled with ADASYN. CFs serves to increase variability in the data, for example, with more female patients that smoke, while the oversampling technique allows to fill less populated feature spaces. As a result, the performance slightly improved in relation to the other augmentation methods, $TPR = 74.1\%$. Analyzing by group, the TPR improved slightly for men and decreased for women, $TPR_F = 76.3\%$ and $TPR_M = 72.6\%$, with $PredP = 3.6p.p.$. For negative outcomes, the values improve in relation to the original model but are worse than the other tested techniques. As for the CF analysis, there are fewer flips, with the overall CMCC improving from 89.7% to 93.4%, mitigating all the mentioned metrics in the base model by at least $3.0p.p.$.

Table 5.1: Performance and CF metrics referring to the base model and the model after data augmentation with ADASYN, plausible CFs and a combination of both.

	Augmentation Method											
	Base			Plausible CFs			ADASYN			CFs with ADASYN		
MCC(%)↑	12.0	13.6	10.9	12.2	14.3	10.9	12.0	13.2	11.1	12.2	13.0	11.5
TPR(%)↑	**74.6**	**78.0**	**72.3**	**73.1**	**75.1**	**71.8**	**72.3**	**75.9**	**69.8**	**74.1**	**76.3**	**72.6**
TNR(%)↑	**48.8**	**47.8**	**49.4**	50.7	52.0	49.9	51.2	**49.1**	**52.4**	49.6	48.3	50.4
CMCC(%)↑	**89.7**	89.9	89.8	87.9	89.4	87.7	**95.3**	95.5	95.2	**93.4**	94.4	92.8
SR(%)↑	5.1	5.1	5.2	6.1	5.5	6.4	2.3	2.2	2.4	3.3	2.8	3.6
PSR(%)↓	4.6	**8.1**	2.8	4.4	**10.0**	0.8	2.6	1.8	3.1	2.9	3.2	2.7
NSR (%)↓	5.5	2.5	**7.4**	7.8	0.9	**11.7**	2.0	2.7	**1.6**	3.6	2.4	4.4
TPSR(%)↓	1.8	1.1	2.3	2.1	0.5	**3.2**	1.4	2.2	**0.8**	1.6	0.0	2.7
TNSR(%)↓	4.5	7.7	2.6	4.1	**9.6**	0.6	2.5	1.8	3.0	2.8	3.2	2.6
FPSR(%)↓	5.9	2.7	7.9	8.3	1.0	12.6	2.1	2.7	1.7	3.9	2.7	4.6
FNSR(%)↓	10.5	**18.5**	6.1	11.2	21.3	5.0	4.8	0.0	7.5	4.5	3.5	5.2
	Total	$F \rightarrow M$	$M \rightarrow F$	**Total**	$F \rightarrow M$	$M \rightarrow F$	**Total**	$F \rightarrow M$	$M \rightarrow F$	**Total**	$F \rightarrow M$	$M \rightarrow F$

This use case displays how the generation of CFs heavily influences the results and the importance of integrating Domain Knowledge to ensure a reliable evaluation. As for mitigation exploration, this experience allowed to delve into the implications posed by CF augmentation, as well as the potential for combining these samples with other methods to allow for a more robust model in terms of performance and bias.

The ECCM for each of the augmentation techniques can be visualized in Figure 5.2, complemented by the most relevant metrics summarized in Table 5.1.

5.1.2 Bias for the sensitive feature 'smoking'

In ML, bias is typically linked to fairness, particularly regarding protected attributes. However, analyzing other features may also yield critical insights into the model's behaviour. Smoking is a well-established risk factor that increases the probability of several lung and heart diseases [28]. As such, it would be reasonable to assume that the model would perform better for individuals who exhibit this characteristic. However, the model underperforms for smokers, hinting at a dataset representational bias. The TPR for the testing subset of non smokers is 74.6% and 71.4% for smokers. In this use case, different sets of CFs were used. The base model mirrors the architecture described in Section 4.3.1, noting that the sensitive feature is not included in training.

Analysing Bias - CFs without Domain Knowledge

Because smoking is correlated with most of the complications and preconditions it seems reasonable to allow all the features to switch when generating them. The first set of CFs was generated allowing changes in every feature.

While $EOdds = 3.2p.p.$ ($TPR_{NS} = 74.6\%$ - $TPR_S = 71.4\%$) is not strictly significant, the resulting ECCM displayed in Figure 5.4 reveals a slight bias, associating non smokers with positive outcomes (post-surgery complications), supported by a $TNSR_{S \rightarrow NS} = 22.8\%$. Simultaneously, smokers are more linked to negative outcomes (smooth recovery), evidenced by $TPSR_{NS \rightarrow S} = 13.7\%$. Suspecting the potential bias due to the lack of representation of

specific samples, the next step is to attempt mitigation through augmentation with CFs. From this, it is expected to mitigate both group and CF bias. Retraining the model with the dataset augmented by the train set CFs, the results show improvement in performance for smokers, TPR_S = 72.8%, and non smokers, TPR_S = 7.70%, both accompanied by a slight decrease in the negative outcomes, where total TNR decreases from 50.2% to 49.1%.

When reevaluating the CF metrics, it was noted that the value of $TNSR_{S \to NS}$ = 12.1% was slightly mitigated, as well as $TPSR_{NS \to S}$ = 0.3%. Nevertheless, the value of $FPSR_{NS \to S}$ = 57.4% is significantly higher than the base model without augmentation. When there is augmentation, a high flux from initially incorrectly predicted values to correctly predicted values may indicate some overfitting. This occurs because the CFs for the test group are generated based on the train samples, thus they are inevitably closer to the train set than the original samples. For this reason, the main concern is the switches in originally TP and TN.

Analysing Bias - CFs with Domain Knowledge by removing correlation with the 'sex' feature

As it was introduced before, 'smoking' is highly correlated with male patients in this dataset(88.0% of smokers are male). Moreover, the PDF of the patient's height is skewed to higher values in relation to non smokers (see Figure 5.3), which can influenced by the former correlation with sex. While height is not strictly related to smoking and could be easily excluded from the CF generation process, there are features related to both sex and smoking habits. For example, one side effect of smoking is the loss of appetite, which leads to smokers having on average a lower BMI. Weight can also be tied to the individual's sex, with men being generally heavier. For this reason, when switching from smoker to non smoker, because most smokers are men, whereas non smokers have approximately the same sex representation (52.3% female and 47.7% male), the tendency is to decrease the weight when the opposite would be closer to a real-world scenario. With these considerations, the generated CFs for smokers and non smokers were created based on an augmented version of the dataset with male and female CFs. These auxiliary CFs were previously described as including a selected set of features based on Domain Knowledge. Statistically, combining the original training set with the 'sex' CFs, aids in eliminating correlations between the sex of the patient and every feature that was not included in the generating process. As smoking is part of the non-included features, now, for every male smoker there is an equivalent female smoker. Analyzing once again the distribution of heights for the augmented training set, the mean value for smokers and non smokers is now aligned as it would be expected, as illustrated in Figure 5.3.

With this second set of CFs for smokers, the resulting CF analysis indicates once again a tendency to associate non smokers with positive outcomes to higher degree, $TNSR_{S \to NS}$ = 0.545.

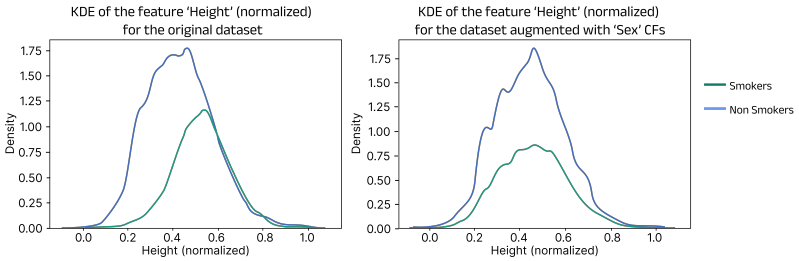


Figure 5.3: Kernel Density Estimation of the normalized 'Height' of smokers and non smokers in the original dataset(Left) and in the dataset augmented with 'Sex' CFs(Right).

Mitigation Bias - Data Augmentation using CFs without Selection

A first experiment with all generated CFs revealed, as shown in Table 5.2, reduced switches from TN_S to FP_{NS} ($TNSR_{S\rightarrow NS}$ = 43.1%, $TPSR_{NS\rightarrow S}$ = 2.1%) but an aggravated TPR disparity between groups. This suggests a model seemingly more counterfactually fair, whereas it is still not able to reliably predict complications for smokers.

Mitigation Bias - Data Augmentation using CFs with Selection

Since 'smoking' is a risk factor, intuitively, there are some logical implications that can be drawn. For instance, it is expected that if a non-smoker had complications, then if this patient smoked, they would also have complications. Similarly, if a smoker didn't have complications post-surgery, then if they didn't smoke, they would most likely not have complications as well. However, the remaining scenarios are more challenging to deduce the result. As a consequence, it seems apparent the validity in CF from patients who smoke and didn't have complications to patients who never smoked, but not for patients who smoke and had complications to patients who don't smoke, since smoking may be (one of) the key factor(s). This hypothesis was tested by conducting augmentation with only selected CFs that obey these rules. This method essentially teaches the model about the added risk of smoking. Because of this direct influence, including all these CFs is not viable as the model will associate all patients that smoke to consequent complications. After some trial and error, it was concluded that including a small fraction of 10% was enough to mitigate some of the existing CF bias, increasing the performance for positive outcomes for smokers (TPR_S = 71.9%) and non-smokers (TPR_{NS} = 75.7%). The PredP remained low at $3.8p.p.$. The model's overall performance also improved slightly to positive outcomes (TPR = 74.3%), but decreased for negative outcomes to a TNR= 49.5%. The metrics are summarized in the Table 5.2.

This experience exemplifies some of the challenges of dealing with a dataset with limited representation for different subgroups and highlights the necessity of taking into consideration Domain Knowledge when generating CFs. Furthermore, it showed that, even if the sensitive feature 'smoke' is not included for model training, some of the generated plausible CFs can have different predictions than the original samples. This is only possible because the features that are included have changed in the process, using the correlations of these features to the developer's advantage. This sets the limits of the

CFs applications, being applicable in any situation as long as there are features in the model related to the sensitive feature.

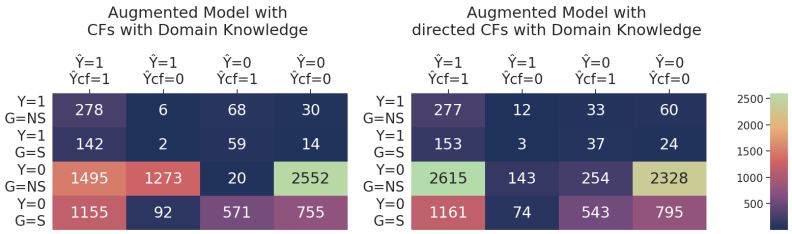


Figure 5.4: ECCM for the base model trained on the CardioFollow.AI data, considering 'Smoking' as sensitive feature, and using data augmentation with CFs generated by removing correlation with the feature 'Sex'. Left: Model augmented with all the CFs; Right: Model augmented with directed CFs.
Y - Ground Truth (0 for 'recovery without complications' and 1 for 'post-surgery complications'); G - Group (S for 'Smoker' and NS for 'Non Smoker'); $\hat{Y}$ - Prediction for the factual(original) sample; $\hat{Y}_{CF}$ - Prediction for the CF sample.

Table 5.2: Performance and CF metrics for the base model trained on the CardioFollow.AI data, considering the sensitive feature 'Smoking', and using data augmentation with CFs generated by removing correlation with the feature 'Sex'.

	Augmentation Method								
	Base Model			All CFs			Directed CFs		
MCC (%)↑	12.1	13.0	11.8	10.6	9.6	11.3	12.2	12.8	12.0
TPR (%)↑	**73.5**	71.4	74.6	71.5	**66.4**	**74.4**	**74.3**	**71.9**	**75.7**
TNR (%)↑	**50.2**	52.8	49.0	49.3	51.6	48.2	**49.5**	52.0	48.4
CMCC (%)↑	71.1	51.1	85.3	51.7	52.2	58.4	74.8	56.8	84.5
TNSR (%)↓	20.2	**54.5**	2.4	15.2	**43.1**	0.8	20.3	**40.6**	9.8
FNSR (%)↓	28.9	72.6	**0.0**	74.3	80.8	**69.4**	45.5	60.7	**35.5**
TPSR (%)↓	8.0	2.6	**10.9**	1.9	1.4	**2.1**	3.4	1.9	**4.2**
FPSR (%)↓	9.1	1.2	12.6	34.0	7.4	46.0	5.4	6.0	5.2
	Total	$S \rightarrow NS$	$NS \rightarrow S$	**Total**	$S \rightarrow NS$	$NS \rightarrow S$	**Total**	$S \rightarrow NS$	$NS \rightarrow S$

5.1.3 Complementarity between CF and Traditional Mitigation Techniques

Instead of the Random Forest architecture presented before, the algorithm Fair GBM was computed with and without mitigation [21]. The undersampling technique Edited Nearest Neighbours (ENN) [77] was applied to balance the classes and achieve a similar performance to the previous model. With this, the model has TPR = 76.6% and does not have significant group bias, as per a $PredP$ = 3.2$p.p.$. Knowing this, it is not expected that Fair GBM will achieve different outcomes. The aim of this experiment is to display how biases can be veiled in seemingly fair models and cannot be weeded out by group fairness mitigation methods. However, there is significant CF bias as detected before in the Random Forest model. In fact, there is a tendency to associate negative outputs with smoking, $NSR_{NS \rightarrow S}$ = 11.3%, and positive outcomes with non smokers, $PSR_{S \rightarrow NS}$ = 17.5% (Table

5.3). After mitigation, the model performance and CF metrics remained similar, improving slightly overall, as displayed in Figure 5.5.

As a complementary device to measure CF bias, assuming the CFs as valid and truthful to the real world representations, it is possible to treat them as real samples and evaluate their performance. For instance, in this case, while the TPR is similar for both groups of the original data, the results for the CFs paint a different picture. For CFs assigned as smokers the $TPR_{S_{CF}}$ = 63.9% and for non smokers it is $TPR_{NS_{CF}}$ = 94.0%, pointing to group bias ($PredP$ = 30.1$p.p.$). The mitigation approach was not able to reduce this discrepancy, as seen by a $TPR_{S_{CF}}$ = 63.1% and $TPR_{NS_{CF}}$ = 94.0%. A better example on how the mitigation of group fairness may induce and/or veil CF bias is described in I.

As a final consideration, while this caveat of fairness constraints may be pointed out, the proposed method is not a replacement for these techniques. Group fairness is a concrete measure of bias whereas CFs are more abstract and require careful validation which was not fully explored in this body of work. With this, the fusion of methodologies is encouraged to achieve results well suited for the represented distributions, group bias, and the less represented distributions, CF bias.

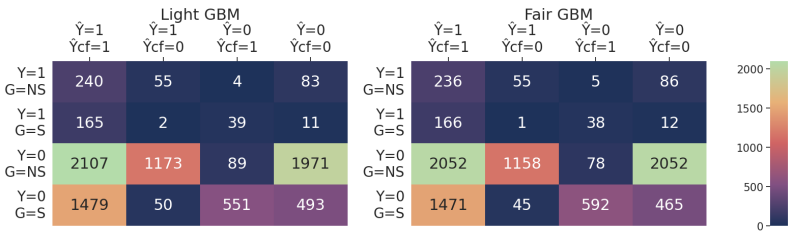


Figure 5.5: ECCM for the models trained on the CardioFollow.AI data, considering 'Smoking' as sensitive feature, and using CFs generated by removing correlation with the feature 'Sex'. Left: Model augmented with all the CFs; Right: Model augmented with directed CFs.
Y - Ground Truth (0 for 'recovery without complications' and 1 for 'post-surgery complications'); G - Group (S for 'Smoker' and NS for 'Non Smoker'); $\hat{Y}$ - Prediction for the factual(original) sample; $\hat{Y}_{CF}$ - Prediction for the CF sample.

Table 5.3: Performance and CF metrics referring to the Fair GBM model with (right) and without fairness constraints (left).

	Light GBM			Fair GBM		
MCC (%)↑	8.3	7.5	8.9	8.8	7.8	9.5
TPR (%)↑	**76.6**	73.1	79.1	76.8	73.5	79.1
TNR (%)↑	39.2	40.9	38.1	40.0	41.1	39.3
CMCC (%)↑	78.8	79.1	79.1	81.0	80.9	81.5
PSR (%)↓	11.8	**17.5**	8.0	10.5	15.9	7.0
NSR (%)↓	9.0	5.1	**11.3**	8.1	4.7	10.2
	Total	$S_{\rightarrow}NS$	$NS_{\rightarrow}S$	**Total**	$S_{\rightarrow}NS$	$NS_{\rightarrow}S$

5.2 Heart Disease Dataset

In this experiment, bias associated with 'Sex' is evaluated. Concerning the correlation of the features with the sex of the individual, there are subtle differences between electrocardiograms for men and women due to hormonal levels, especially the impact of estrogen and testosterone in cardiac functions, as well as anatomical differences such as the size of the heart [41]. As an example, women tend to have higher resting heart rates due to having on average smaller hearts, needing higher frequency to pump enough blood. Additionally, the protective effect of estrogen in arteries is heavily documented, explaining the increased risk of cardiac complications after menopause for women [12].

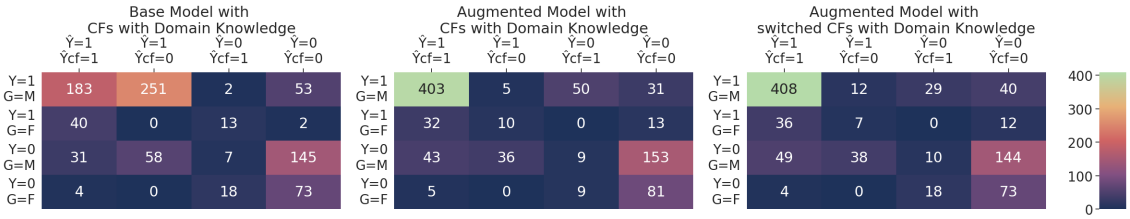


Figure 5.6: ECCM for the base model trained on the Heart Disease dataset, considering the sensitive feature 'Sex', and using data augmentation with CFs generated with Domain Knowledge. Left: Base Model; Middle: Model augmented with all the CFs; Right: Model augmented with switched CFs.
Y - Ground Truth (0 for 'Free from CAD' and 1 for 'Signs of CAD'); G - Group (M for 'male' and F for 'female'); $\hat{Y}$ - Prediction for the factual(original) sample; $\hat{Y}_{CF}$ - Prediction for the CF sample.

Analysing Bias - CFs with Domain Knowledge

This dataset contains several attributes extracted from the Eletrocardiogram (ECG), which were taken into account for generating CFs. In contrast, features such as 'age' and pre-existing conditions and risk factors were kept unchanged.

The base model is a Logistic Regression and despite the small sample size, the results, shown in Table 5.4, are satisfactory with an $Accuracy(ACC)$ = 81.5% and TPR = 87.1%. Nonetheless, there is considerable bias noted by a difference in TPR against women, $PredP$ = 16.1$p.p.$. This discrepancy is apparent in the majority of the cross-validation scheme, reinforcing the existence of model bias. Drawing the ECCM (c.f. Figure 5.6), the results display a clear tendency to associate female patients with negative outcomes, $TPSR_{M\rightarrow F}$ = 57.8% and $NSR_{M\rightarrow F}$ = 59.1%, as well as a tendency of associating positive outcomes to men, $TNSR_{F\rightarrow M}$ = 19.8% and $PSR_{F\rightarrow M}$ = 29.2%, supporting the group fairness assessment.

Mitigating Bias - Augmentation with all the CFs

To mitigate this bias, the dataset was supplemented with all the CFs generated in the first instance. Using this augmented set, the CF bias was mostly mitigated, but, unexpectedly, it introduced an opposite effect leading TP_F to switch to FN_M, as per a

Table 5.4: Performance and CF metrics for the base model and models trained on the dataset augmented with all the CFs and only the CFs that switched.

	Base Model			Augmentation Method: All CFs			Augmentation Method: Switched CFs		
ACC (%) ↑	**81.5**	87.3	80.3	79.8	88.0	78.1	**80.5**	89.3	78.6
TPR (%) ↑	**87.1**	72.7	88.8	82.7	76.4	83.4	**85.1**	78.2	85.9
TNR (%) ↑	72.3	95.8	63.1	75.0	94.7	67.2	72.9	95.8	63.9
CMCC (%) ↓	**33.1**	64.4	35.8	71.5	70.4	68.7	**72.2**	64.4	71.7
PSR (%) ↓	12.8	**29.2**	4.4	19.7	8.7	24.3	17.5	17.5	17.5
NSR (%) ↓	54.5	0.0	**59.1**	9.6	21.3	8.4	10.3	14.9	9.9
TPSR (%) ↓	53.0	0.0	**57.8**	3.3	**23.8**	1.2	4.1	**16.3**	2.9
TNSR (%) ↓	10.3	**19.8**	4.6	07.1	10.0	5.6	11.4	19.8	6.5
	Total	$F_{\rightarrow}M$	$M_{\rightarrow}F$	**Total**	$F_{\rightarrow}M$	$M_{\rightarrow}F$	**Total**	$F_{\rightarrow}M$	$M_{\rightarrow}F$

$TPSR_{F\rightarrow M}$ = 23.8%. While high, considering the $TPSR_{M\rightarrow F}$ for the base model was substantially higher and the other mentioned metrics decreased, the augmentation can be deemed successful in the mitigation of CF bias.

Moreover, since the performance decreased, there is room to explore other iterations that hopefully achieve similar results without duplicating the sample space. A logical conclusion is to select which CFs to include. There are different available routes to achieve this, such as selecting the most unique CFs to avoid redundancy, or selecting the CFs that support the Domain Knowledge, stirring the model in a predetermined direction as it is displayed in experience Section 5.1.2. Finally, there is a simple approach of including CFs that show a different outcome than the original sample, called 'switched CFs'.

Mitigating Bias - Augmentation with switched CFs

Since the model is already able to correctly predict a given set of pairs of samples and their CFs, there seems to be no need to add information about the CF equivalent of these samples. Thus, only the incorrectly predicted, or switched CFs, should be added. This method is not infallible, due to the unpredictability of training a model, but it can achieve good results in specific scenarios. Compared to the original model, it was successful, in this case, in mitigating CF, the CMCC increased from 33.1% in the original model to 72.2%, and group bias, as per a $PredP$ = 7.7$p.p.$, with minimal loss in total performance, as seen by an ACC = 80.5% and TPR = 85.1%. In relation to augmentation with all the CFs the introduced bias displayed that severely increased the $TPSR_{F\rightarrow M}$, is much more tenuous, $TPSR_{F\rightarrow M}$ = 16.3%. In contrast, it was not able to correct the flux of TN_F to FP_M, maintaining the $TNSR_{M\rightarrow F}$ value at 19.8%. Overall, given this classification context application that poses recall as the priority, opting to add only switched CFs proved to be more fruitful in mitigating bias in the priority class. Along with the fact that the contested metric in this case, $TNSR_{M\rightarrow F}$, is largely less grave in value than the $TPSR_{F\rightarrow M}$ introduced by augmentation with all the CFs.

Note about Cross-Validation and individual folds

When analyzing individual folds there is a stark contrast in the effect of the data the model was trained on. The additional matrixes can be viewed in Appendix D.

6

Conclusions & Future Work

This work has successfully achieved its overarching goals, advancing the field of fairness in ML and bias detection within the context of classification tasks, particularly in the critical domain of medical datasets. This chapter reviews the applications and contributions of this work, providing insights into the applications, prospects and limitations of the proposed methods.

The proposed method for generating plausible CFs expands on the notion introduced by *Russell et al.*, enhancing the credibility of the approach [64]. By leveraging dataset distributions, the generated CFs are firmly rooted in real-world examples. The use of outcome-based sectioning ensures consistency with the ground truth, a prerequisite for CF fairness. Demonstrating the importance of domain knowledge, this method introduces variability for realistic changes. Selecting the most feasible potential features related to the sensitive feature under study is crucial to ensure plausibility of the CFs. In turn, because of the changes in related features, the proposed CF evaluation method can be employed even without the sensitive feature, as demonstrated in experience 5.1.2. This opens possibilities for exploring more viable models without significantly expanding the feature space.

Comparatively to existing methods, such as *FlipTest* that presents OTMs and GANs for CF generation [10], this work proposes a controlled generation process, which relies more superficially on the dataset distributions allowing deviations from the given examples (factuals). The decrease in complexity has its shortcomings, as the correlation between features is not as well utilized, on the other hand, assuming some level of independence between features allows to generate samples outside of the representations that may be plausible in context. As advantages, this method it is less computationally demanding, as it does not require training any additional model, and it is suited to some degree of customization from the user. Given the simplicity of the generation and evaluation processes, this framework is still applicable after deployment.

The proposed CCM allows for an in-depth bias evaluation of developed models. It is extremely flexible to different tasks, and while it has only been demonstrated in binary classification tasks for binary sensitive features, it can be easily employed without loss of generalization to categorical features and multi-class tasks. The ECCM allows to associate

switches with the label and correctness of the model, providing a more nuanced view on which samples the model is less certain, better accounting for the model error. As limitations, the metrics extrapolated from the extended version are limited by the portion of samples of the CM to which they belong, so the drawn conclusions should account for the sample pool size. On the other hand, compared to methods that measure causal inference, which are mostly model defined, as an agnostic tool, this method has a greater range of applications.

The experiences effectively convey the complementarity between group fairness and the proposed CF metrics. The group fairness metrics grant a practical and objective reference for bias, while the CF metrics explore bias outside the most common distributions in the dataset, highlighting the advantages of using them simultaneously. Additionally, this work displays the flexibility of CFs for mitigation techniques, whether by using all the generated samples, selecting those best supported by literature or domain knowledge, or even by adding only the samples for which the model was not consistent.

Given the potential of CFs for fairer decision-support systems, there are several venues to extend this work in future research. Starting with the generation process, generative models could improve the robustness of the CFs. Potentially introducing a control element in the process is also viable either by Human-in-the-Loop techniques, resorting to experts to review the validity of the CF, or through a Conditional Generative Adversarial Network (cGAN). Additionally, instead of relying on the training set distributions, work can be done to use real-world statistics instead, when available. In this context, techniques such as transfer learning could also be valuable for the task.

As mitigation techniques, the proposed metrics could be employed as constraints in the model fitting process inspired by existing algorithms relying on group fairness criteria. For augmentation purposes, it would be interesting to curate CFs included in the learning stage. Because the CFs are close relatives to the original sample, this step would be crucial to reduce data redundancy. This could be achieved by iterative processes that add random batches and evaluate based on CF and/or standard performance metrics. Another option is adapting undersampling techniques which maximate the variability of samples such as ENN. And, once again, domain expertise may be crucial to ensure only the best CFs are used.

The counterfactual setting has been proved beneficial to improve models' robustness, outside of the fairness context. Given an appropriate CF generation process targeting specific aspects (e.g., filling out-of-distribution regions), the ECCM and its metrics could be used to evaluate how well the model responds towards more responsible decision-support systems, benefiting society as whole.

As material contributions to this field, this thesis resulted in an article introducing the CCM and the derived bias evaluation metrics, with a poster displayed at 2023 International Conference on Machine Learning (ICML) "DMLR Workshop: Data-centric ML Research"[15]; and an article including the generation process submitted to Journal of Data-centric Machine Learning Research.

Bibliography

[1] S. Agrawal et al. "Sex, Myocardial Infarction, and the Failure of Risk Scores in Women." In: *Journal of women's health* 24.11 (2015-08), pp. 859–61. DOI: 10.1089/jwh.2015.5412 (cit. on p. 29).

[2] S. Alelyani. "Detection and Evaluation of Machine Learning Bias". en. In: *Applied Sciences* 11.14 (2021-07), p. 6271. ISSN: 2076-3417. DOI: 10.3390/app11146271 (cit. on p. 12).

[3] S. Ali et al. "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence". In: *Information Fusion* 99 (2023-05), p. 101805. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2023.101805 (cit. on p. 10).

[4] J. Angwin et al. *Machine Bias*. en. 2016-05. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (visited on 2023-01-06) (cit. on p. 1).

[5] A. Artelt and B. Hammer. *Fairness and Robustness of Contrasting Explanations*. 2021. DOI: 10.48550/arXiv.2103.02354 (cit. on p. 16).

[6] A. Asuncion and D. Newman. *UCI Machine Learning Repository*. https://archive.ics.uci.edu/. 2007. DOI: 10.24432/C52P4X (cit. on p. 27).

[7] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. 2019. URL: http://www.fairmlbook.org (cit. on p. 12).

[8] R. K. E. Bellamy et al. *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*. 2018-10. DOI: 10.48550/arXiv.1810.01943. (Visited on 2023-02-12) (cit. on p. 16).

[9] S. Bird et al. *Fairlearn: A toolkit for assessing and improving fairness in AI*. Tech. rep. MSR-TR-2020-32. Microsoft, 2020-05. URL: https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/ (cit. on p. 16).

[10] E. Black, S. Yeom, and M. Fredrikson. "FlipTest: Fairness Testing via Optimal Transport". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. New York, NY, USA: Association for Computing Machinery, 2020, 111–121. ISBN: 9781450369367. DOI: 10.1145/3351095.3372845 (cit. on pp. 17, 39).

[11] E. Black et al. "Consistent Counterfactuals for Deep Models". In: *CoRR* abs/2110.03109 (2021). DOI: 10.48550/arXiv.2110.03109 (cit. on p. 16).

[12] S. Bokhari and S. R. Bergmann. "The Effect of Estrogen Compared to Estrogen Plus Progesterone on the Exercise Electrocardiogram". In: *Journal of the American College of Cardiology* 40.6 (2002), pp. 1092–1096. DOI: 10.1016/s0735-1097(02)02111-3 (cit. on p. 37).

[13] L. Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001-10), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324 (cit. on p. 7).

[14] K. P. Burnham and D. R. Anderson, eds. *Model Selection and Multimodel Inference*. en. New York, NY: Springer, 2004. ISBN: 978-0-387-95364-9. DOI: 10.1007/b97636 (cit. on p. 7).

[15] A. Carreiro et al. "The Matrix Reloaded: A Counterfactual Perspective on Bias in Machine Learning". In: Data-centric Machine Learning Research (DMLR) Workshop at the International Conference of Machine Learning (ICML). 2023. URL: https://dmlr.ai/assets/accepted-papers/85/CameraReady/ICML_2023_DMLR_Workshop_Counterfactual_Confusion_Matrix.pdf (cit. on pp. 40, 57).

[16] H. Chen et al. *On Learning and Testing of Counterfactual Fairness through Data Preprocessing*. 2022-02. DOI: 10.48550/arXiv.2202.12440 (cit. on p. 15).

[17] I. Chen, P. Szolovits, and M. Ghassemi. "Can AI Help Reduce Disparities in General Medical and Mental Health Care?" In: *AMA Journal of Ethics* 21.2 (2019), E167–E179. DOI: 10.1001/amajethics.2019.167 (cit. on p. 1).

[18] D. Chicco, N. Tötsch, and G. Jurman. "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation". In: *BioData Mining* 14.1 (2021-02), p. 13. ISSN: 1756-0381. DOI: 10.1186/s13040-021-00244-z (cit. on p. 10).

[19] W. Commons. *File:Preventive Medicine - Statistics Sensitivity TPR, Specificity TNR, PPV, NPV, FDR, FOR, ACCuracy, Likelihood Ratio, Diagnostic Odds Ratio 2 Final wiki.png — Wikimedia Commons, the free media repository*. [Online; accessed 30-September-2023]. 2020. URL: https://commons.wikimedia.org/w/index.php?title=File:Preventive_Medicine_-_Statistics_Sensitivity_TPR,_Specificity_TNR,_PPV,_NPV,_FDR,_FOR,_ACCuracy,_Likelihood_Ratio,_Diagnostic_Odds_Ratio_2_Final_wiki.png&oldid=506826597 (cit. on p. 9).

[20] G. Cornacchia et al. *Counterfactual Reasoning for Bias Evaluation and Detection in a Fairness under Unawareness setting*. 2023. arXiv: 2302.08204 [cs.LG] (cit. on p. 22).

[21] A. F. Cruz et al. *FairGBM: Gradient Boosting with Fairness Constraints*. 2023-03. DOI: 10.48550/arXiv.2209.07850 (cit. on pp. 16, 35, 58).

[22] I. Curioso et al. "Addressing the Curse of Missing Data in Clinical Contexts: A Novel Approach to Correlation-Based Imputation". In: *Journal of King Saud University - Computer and Information Sciences* 35 (2023), pp. 1–12 (cit. on p. 26).

[23] R. Daneshjou et al. "Disparities in Dermatology AI Performance on a Diverse, Curated Clinical Image Set". In: *Science Advances* 8 (2022), eabq6147. DOI: 10.1126/sciadv.abq6147 (cit. on p. 2).

[24] R. C. Detrano et al. "International application of a new probability algorithm for the diagnosis of coronary artery disease." In: *The American journal of cardiology* 64 5 (1989), pp. 304–10. URL: https://api.semanticscholar.org/CorpusID:23545303 (cit. on p. 27).

[25] W. Fleisher. "What's Fair about Individual Fairness?" In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '21. New York, NY, USA: Association for Computing Machinery, 2021-07, pp. 480–490. ISBN: 978-1-4503-8473-5. DOI: 10.1145/3461702.3462621 (cit. on p. 14).

[26] J. Fonseca and F. Bacao. "Tabular and Latent Space Synthetic Data Generation: A Literature Review". English. In: *Journal of Big Data* 10 (2023-07), pp. 1–37. ISSN: 2196-1115. DOI: 10.1186/s40537-023-00792-7 (cit. on p. 10).

[27] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. *On the (im)possibility of fairness*. arXiv:1609.07236 [cs, stat]. 2016-09. URL: http://arxiv.org/abs/1609.07236 (visited on 2023-02-09) (cit. on p. 14).

[28] G. Gallucci et al. "Cardiovascular risk of smoking and benefits of smoking cessation". In: *Journal of Thoracic Disease* 12.7 (2020). ISSN: 2077-6624. URL: https://jtd.amegroups.org/article/view/37685 (cit. on p. 32).

[29] P. Garg, J. Villasenor, and V. Foggo. *Fairness Metrics: A Comparative Analysis*. arXiv:2001.07864 [cs]. 2020-01. DOI: 10.48550/arXiv.2001.07864. URL: http://arxiv.org/abs/2001.07864 (visited on 2023-02-19) (cit. on p. 14).

[30] A. Géron. *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems*. Sebastopol, CA: O'Reilly Media, 2017. ISBN: 978-1491962299 (cit. on p. 9).

[31] S. Goethals, D. Martens, and T. Calders. "PreCoF: counterfactual explanations for fairness". In: *Machine Learning* (2023). ISSN: 1573-0565. DOI: 10.1007/s10994-023-06319-8 (cit. on p. 17).

[32] N. Gotlieb, A. Azhie, D. Sharma, et al. "The Promise of Machine Learning Applications in Solid Organ Transplantation". In: *npj Digital Medicine* 5 (2022), p. 89. DOI: 10.1038/s41746-022-00637-2. URL: 10.1038/s41746-022-00637-2 (cit. on p. 23).

[33] M. Hardt et al. "Amazon SageMaker Clarify: Machine Learning Bias Detection and Explainability in the Cloud". In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. arXiv:2109.03285 [cs]. 2021-08, pp. 2974–2983. DOI: 10.1145/3447548.3467177 (cit. on p. 16).

[34] M. Hardt et al. "Equality of Opportunity in Supervised Learning". In: *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc., 2016. URL: https://papers.nips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html (visited on 2023-02-09) (cit. on pp. 13, 15).

[35] H. He et al. "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning". In: *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. 2008, pp. 1322–1328 (cit. on p. 31).

[36] I. Hull. "Generative Models". In: 2021-01, pp. 307–330. ISBN: 978-1-4842-6372-3. DOI: 10.1007/978-1-4842-6373-0_9 (cit. on p. 4).

[37] D. Kaushik, E. H. Hovy, and Z. C. Lipton. "Learning the Difference that Makes a Difference with Counterfactually-Augmented Data". In: *ArXiv* abs/1909.12434 (2019). URL: https://api.semanticscholar.org/CorpusID:203591519 (cit. on p. 16).

[38] D. Kaushik et al. "Explaining The Efficacy of Counterfactually-Augmented Data". In: *ArXiv* abs/2010.02114 (2020) (cit. on p. 17).

[39] G. Ke et al. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf (cit. on p. 58).

[40] Y. Kim and Y. Kim. "Explainable heat-related mortality with random forest and SHapley Additive exPlanations (SHAP) models". In: *Sustainable Cities and Society* 79 (2022), p. 103677. ISSN: 2210-6707. DOI: 10.1016/j.scs.2022.103677. URL: https://www.sciencedirect.com/science/article/pii/S2210670722000117 (cit. on p. 11).

[41] A. A. Knowlton and A. R. Lee. "Estrogen and the Cardiovascular System". In: *Pharmacology Therapeutics* 135.1 (2012), pp. 54–70. DOI: 10.1016/j.pharmthera.2012.03.007 (cit. on p. 37).

[42] R. Kohavi and B. Becker. *Adult (Census Income) Dataset*. https://archive.ics.uci.edu/ml/datasets/Adult. UCI Machine Learning Repository. 1996 (cit. on p. 57).

[43] J. R. Koza et al. "Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming". In: *Artificial Intelligence in Design '96*. Ed. by J. S. Gero and F. Sudweeks. Dordrecht: Springer Netherlands, 1996, pp. 151–170. ISBN: 978-94-009-0279-4. DOI: 10.1007/978-94-009-0279-4_9. URL: 10.1007/978-94-009-0279-4_9 (cit. on p. 4).

[44] P. Kumar and M. Bhatia. "Coronary Artery Disease Reporting and Data System: A Comprehensive Review". In: *Journal of Cardiovascular Imaging* 30.1 (2022). Epub 2021 Mar 23, pp. 1–24. DOI: 10.4250/jcvi.2020.0195. URL: 10.4250/jcvi.2020.0195 (cit. on p. 27).

[45] M. J. Kusner et al. *Counterfactual Fairness*. 2017. DOI: 10.48550/ARXIV.1703.06856. URL: https://arxiv.org/abs/1703.06856 (cit. on p. 14).

[46] Y. LeCun, Y. Bengio, and G. Hinton. "Deep learning". en. In: *Nature* 521.7553 (2015-05). Number: 7553 Publisher: Nature Publishing Group, pp. 436–444. ISSN: 1476-4687. DOI: 10.1038/nature14539. URL: https://www.nature.com/articles/nature14539 (visited on 2023-02-21) (cit. on p. 4).

[47] J. M. Lourenço. *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. 2021. URL: https://github.com/joaomlourenco/novathesis/raw/master/template.pdf (cit. on p. ii).

[48] B. Matthews. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". In: *Biochimica et Biophysica Acta (BBA) - Protein Structure* 405.2 (1975), pp. 442–451. ISSN: 0005-2795. DOI: 10.1016/0005-2795(75)90109-9. URL: https://www.sciencedirect.com/science/article/pii/0005279575901099 (cit. on p. 16).

[49] K. Maughan, I. C. Ngong, and J. P. Near. "Prediction Sensitivity: Continual Audit of Counterfactual Fairness in Deployed Classifiers". In: *ArXiv* abs/2202.04504 (2022). URL: https://api.semanticscholar.org/CorpusID:246680201 (cit. on p. 17).

[50] W. S. McCulloch and W. Pitts. "A logical calculus of the ideas immanent in nervous activity". en. In: *The bulletin of mathematical biophysics* 5.4 (1943-12), pp. 115–133. ISSN: 1522-9602. DOI: 10.1007/BF02478259. URL: 10.1007/BF02478259 (visited on 2023-02-09) (cit. on p. 1).

[51] N. Mehrabi et al. "A Survey on Bias and Fairness in Machine Learning". en. In: *ACM Computing Surveys* 54.6 (2022-07), pp. 1–35. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3457607. URL: https://dl.acm.org/doi/10.1145/3457607 (visited on 2023-02-07) (cit. on p. 5).

[52] G. Miner et al. "Feature Selection and Dimensionality Reduction". In: *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Academic Press, 2012, pp. 929–934. ISBN: 9780123869791. DOI: 10.1016/B978-0-12-386979-1.00038-4. URL: https://www.sciencedirect.com/science/article/pii/B9780123869791000384 (cit. on p. 6).

[53] R. K. Mothilal, A. Sharma, and C. Tan. "Explaining machine learning classifiers through diverse counterfactual explanations". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, 2020-01. DOI: 10.1145/3351095.3372850 (cit. on p. 16).

[54] S. T. Mueller et al. "Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI". In: *CoRR* abs/1902.01876 (2019). arXiv: 1902.01876. URL: http://arxiv.org/abs/1902.01876 (cit. on p. 10).

[55] S. I. Nikolenko. *Synthetic Data for Deep Learning*. 2019. arXiv: 1909.11512 [cs.LG] (cit. on p. 10).

[56] B. Nisbet. "Tutorial E - Feature Selection in KNIME". In: *Handbook of Statistical Analysis and Data Mining Applications (Second Edition)*. Ed. by R. Nisbet, G. Miner, and K. Yale. Second Edition. Boston: Academic Press, 2018, pp. 377–391. ISBN: 978-0-12-416632-5. DOI: 10.1016/B978-0-12-416632-5.00027-X. URL: https://www.sciencedirect.com/science/article/pii/B978012416632500027X (cit. on p. 7).

[57] E. Ntoutsi et al. "Bias in data-driven artificial intelligence systems—An introductory survey". en. In: *WIREs Data Mining and Knowledge Discovery* 10.3 (2020). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1356, e1356. ISSN: 1942-4795. DOI: 10.1002/widm.1356. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1356 (visited on 2023-01-08) (cit. on pp. 11, 15).

[58] D. Pedreschi, S. Ruggieri, and F. Turini. *Discrimination-aware data mining*. Journal Abbreviation: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Pages: 568 Publication Title: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2008-08. DOI: 10.1145/1401890.1401959 (cit. on p. 15).

[59] G. Pombo et al. "Equitable modelling of brain imaging by counterfactual augmentation with morphologically constrained 3D deep generative models". In: *Medical Image Analysis* 84 (2023), p. 102723. ISSN: 1361-8415. DOI: 10.1016/j.media.2022.102723. URL: https://www.sciencedirect.com/science/article/pii/S1361841522003516 (cit. on p. 17).

[60] M. L. Power and J. Schulkin. "Sex differences in fat storage, fat metabolism, and the health risks from obesity: possible evolutionary origins". In: *British Journal of Nutrition* 99.5 (2008), 931–940. DOI: 10.1017/S0007114507853347 (cit. on p. 29).

[61] C. Reddy et al. "Benchmarking Bias Mitigation Algorithms in Representation Learning through Fairness Metrics". en. In: (2021-03). URL: https://openreview.net/forum?id=xEpUl1um6V (visited on 2023-02-21) (cit. on p. 12).

[62] B. Ribeiro et al. "Unravelling Heterogeneity: A Hybrid Machine Learning Approach to Predict Post-Discharge Complications in Cardiothoracic Surgery". In: (2023), p. 12 (cit. on p. 26).

[63] N. J. Roese and K. Epstude. "Chapter One - The Functional Theory of Counterfactual Thinking: New Evidence, New Challenges, New Insights". In: *Advances in Experimental Social Psychology*. Ed. by J. M. Olson. Vol. 56. Academic Press, 2017-01, pp. 1–79. DOI: 10.1016/bs.aesp.2017.02.001. URL: https://www.sciencedirect.com/science/article/pii/S0065260117300187 (visited on 2023-09-22) (cit. on p. 2).

[64] C. Russell et al. "When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. URL: https://papers.nips.cc/paper/2017/hash/1271a7029c9df08643b631b02cf9e116-Abstract.html (visited on 2023-02-09) (cit. on pp. 14, 17, 39).

[65] P. Saleiro et al. *Aequitas: A Bias and Fairness Audit Toolkit*. arXiv:1811.05577 [cs]. 2019-04. URL: http://arxiv.org/abs/1811.05577 (visited on 2023-02-11) (cit. on p. 16).

[66] R. Santos et al. "A Risk Prediction Framework to Optimize Remote Patient Monitoring Following Cardiothoracic Surgery". In: (2023), p. 6 (cit. on p. 26).

[67] S. H. Schlecht, E. M. R. Bigelow, and K. J. Jepsen. "How Does Bone Strength Compare Across Sex, Site, and Ethnicity?" In: *Clinical Orthopaedics and Related Research* 473 (2015), 2540–2547. DOI: 10.1007/s11999-015-4229-6. URL: 10.1007/s11999-015-4229-6 (cit. on p. 29).

[68] P. Schulam and S. Saria. *Reliable Decision Support using Counterfactual Models*. 2018. arXiv: 1703.10651 [stat.ML] (cit. on p. 11).

[69] B. Smyth and M. T. Keane. *A Few Good Counterfactuals: Generating Interpretable, Plausible and Diverse Counterfactual Explanations*. 2021. arXiv: 2101.09056 [cs.AI] (cit. on p. 16).

[70] S. Sra, S. Nowozin, and S. J. Wright. *Optimization for Machine Learning*. The MIT Press, 2011. ISBN: 026201646X (cit. on p. 7).

[71] I. Stepin et al. "A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence". In: *IEEE Access* 9 (2021), pp. 11974–12001. DOI: 10.1109/ACCESS.2021.3051315 (cit. on p. 20).

[72] H. Suresh and J. Guttag. "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle". en. In: *Equity and Access in Algorithms, Mechanisms, and Optimization*. – NY USA: ACM, 2021-10, pp. 1–9. ISBN: 978-1-4503-8553-4. DOI: 10.1145/3465416.3483305 (cit. on p. 5).

[73] J. Taub, M. J. Elliot, and G. M. Raab. "Creating the Best Risk-Utility Profile : The Synthetic Data Challenge". In: 2019. URL: https://api.semanticscholar.org/CorpusID:204747180 (cit. on p. 10).

[74] M. Temraz and M. T. Keane. "Solving the class imbalance problem using a counterfactual method for data augmentation". en. In: *Machine Learning with Applications* 9 (2022-09), p. 100375. ISSN: 26668270. DOI: 10.1016/j.mlwa.2022.100375. URL: https://linkinghub.elsevier.com/retrieve/pii/S2666827022000652 (visited on 2023-09-21) (cit. on p. 17).

[75] S. Wachter, B. Mittelstadt, and C. Russell. "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR". In: *Harvard Journal of Law & Technology* 31.2 (2018-10). Available at SSRN: https://ssrn.com/abstract=3063289 or http://dx.doi.org/10.2139/ssrn.3063289 (cit. on p. 11).

[76] *What-If Tool*. original-date: 2018-09-07T20:26:10Z. 2023-01. URL: https://github.com/PAIR-code/what-if-tool (visited on 2023-01-23) (cit. on p. 16).

[77] D. Wilson. "Asymptotic Properties of Nearest Neighbor Rules Using Edited Data". In: *IEEE Transactions on Systems, Man, and Cybernetics* 2.3 (1972), pp. 408–421 (cit. on p. 35).

[78] M. B. Zafar et al. "Fairness Constraints: A Flexible Approach for Fair Classification". In: *Journal of Machine Learning Research* 20.75 (2019), pp. 1–42. URL: http://jmlr.org/papers/v20/18-262.html (cit. on p. 15).

[79] B. H. Zhang, B. Lemoine, and M. Mitchell. "Mitigating Unwanted Biases with Adversarial Learning". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '18. New York, NY, USA: Association for Computing Machinery, 2018-12, pp. 335–340. ISBN: 978-1-4503-6012-8. DOI: 10.1145/3278721.3278779. URL: 10.1145/3278721.3278779 (visited on 2023-02-21) (cit. on p. 15).

A

Group Fairness Metrics

Table A.1: Group Fairness Metrics.

Name	**Formula**
Demographic Parity	$P(\hat{Y} = 1 \mid A = a) = P(\hat{Y} = 1 \mid A = b)$
Predictive Parity	$P(\hat{Y} = 1 \mid Y = 1, A = a) = P(\hat{Y} = 1 \mid Y = 1, A = b)$
Equality of Opportunity	$P(\hat{Y} = 1 \mid Y = 1, A = a) = P(\hat{Y} = 1 \mid Y = 1, A = b)$
Equalized Odds	$P(\hat{Y} = 1 \mid Y = 1, A = a) = P(\hat{Y} = 1 \mid Y = 1, A = b)$ $P(\hat{Y} = 1 \mid Y = 0, A = a) = P(\hat{Y} = 1 \mid Y = 0, A = b)$
Predictive Equality	$P(\hat{Y} = 1 \mid Y = 0, A = a) = P(\hat{Y} = 1 \mid Y = 0, A = b)$
Overall Accuracy Equality	$P(\hat{Y} = Y \mid A = a) = P(\hat{Y} = Y \mid A = b)$
Conditional Use Accuracy Equality	$P(\hat{Y} = 1 \mid Y = 1, A = a) = P(\hat{Y} = 1 \mid Y = 1, A = b)$ $P(\hat{Y} = 0 \mid Y = 0, A = a) = P(\hat{Y} = 0 \mid Y = 0, A = b)$
Treatment Equality	$P(\hat{Y} = 1 \mid Y = 0, A = a) = P(\hat{Y} = 0 \mid Y = 1, A = a)$ $P(\hat{Y} = 1 \mid Y = 0, A = b) = P(\hat{Y} = 0 \mid Y = 1, A = b)$

B

Counterfactual Generation Alghorithm

Algorithm 1 Counterfactual Generation

procedure GENERATECOUNTERFACTUALS(Samples, PotentialFeatures=[BinaryFeatures, CategoricalFeatures, Continuous], TrainingSet, SensitiveFeature, minThreshold, maxDepth)
    Initiate $Samples_{CF}$ as a copy of $Samples$
    for each feature f_i not in $BinaryFeatures$ **do**
        Calculate $P_g^{y_i}$ of f_i in the $TrainingSet$ for each group in $SensitiveFeature$
        and label y_i
        for each sample s_i in $Samples_{CF}$ **do**
            if f_i is in $ContinuosFeatures$ **then**
                Interpolate to find v_i' in CDF $P_{g_{CFi}}^{y_i}$ based on v_i in $P_{g_i}^{y_i}$
            else if f_i is in $CategoricalFeatures$ **then**
                Find closest v_i' in CDF $P_{g_{CFi}}^{y_i}$ based on v_i in $P_{g_i}^{y_i}$
            end if
        end for
    end for
    for each sample s_i in $Samples_{CF}$ **do**
        Flip value of $SensitiveFeature$
        $s_i = FlipBinaryFeatures(s_i, BinaryFeatures, TrainingSet_{y=y_i},$
        $SensitiveFeature, minThreshold, maxDepth)$
    end for
    return $Samples_{CF}$
end procedure
procedure FLIPBINARYFEATURES(Sample, BinaryFeatures, TrainingSet, SensitiveFeature, minThreshold, maxDepth, iter=0)
    Initiate $Sample_{CF}$ as a copy of $Sample$
    for each binary feature f_i in $BinaryFeatures$ **do**
        if $iter < maxDepth$ **then**
            Calculate the difference in conditional probability of the value v_i of f_i
            given the original, g_i, and new group, g_{CFi}, of the $SensitiveFeature$
            $ProbDiff = |P(v_i|g_i) - P(v_i|g_{CFi})$
            if $ProbDiff > minThreshold$ **then**
                Flip the binary feature value
                Start new iteration
                $iter_i = iter + 1$
                $Sample_{CF} = FlipBinaryFeatures(Sample_{CF}, BinaryFeatures - f_i,$
                $TrainingSet[SensitiveFeature = g_{CFi}], f_i, minThreshold, maxDepth, iter_i)$
            end if
        end if
    end for
    return $Sample_{CF}$
end procedure

C

Extended Counterfactual Confusion Matrix Metrics

Table C.1: Summary of the CF metrics extrapolated from the CCM and ECCM and its interpretation.

			Group or Global Metric			Parity (A - B)		
CCM Metric	Range	Highest Bias Value	Intuition	Impairec Group	CM Analog	Range	Impaired Group	Group Fairness Analog
CR = 1 - SR	[0,1]	0	The model relies heavily on the distributions represented in the dataset for each group of the sensitive feature, indicating potential bias	New Group	ACC	[-1,1]	[-1,0[A]0,+1] B	Overall Accuracy Parity
PCR = 1 - NSR	[0,1]	0	The model associates the new group with negative outcomes, indicating bias	New Group	TNR	[-1,1]	[-1,0[A]0,+1] B	PredEq
NCR = 1 - PSR	[0,1]	0	The model associates the new group with positive outcomes, indicating bias	Original Group	TPR	[-1,1]	[-1,0[A]0,+1] B	EOpp
PCP = 1 - PSDR	[0,1]	0	The model associates the new group with positive outcomes, indicating bias	Original Group	PPV	[-1,1]	[-1,0[A]0,+1] B	PredP
CMCC	[-1,1]	0	The model relies heavily on the distributions represented in the dataset for each group of the sensitive feature	New Group	MCC	[-2,2]	[-2,0[A]0,+2] B	-

Including the Ground Truth

			Group or Global Metric		Parity (A - B)	
CCM Metric	Range	Highest Bias Value	Intuition	Impaired Group	Range	Impaired Group
TPSR	[0,1]	1	The model is not able to predict correctly positive outcomes based on the group. It indicates the need for more variability in samples with positive ground truth for the new group	New Group	[-1,1]	[-1,0[A]0,+1] B
TNSR	[0,1]	1	The model is not able to predict correctly negative outcomes based on the group. It indicates the need for more variability in samples with negative ground truth for the new group	New Group	[-1,1]	[-1,0[A]0,+1] B
FPSR	[0,1]	1	The model is not able to predict correctly negative outcomes based on the group. It indicates the need for more variability in samples with negative ground truth for the original group	Original Group	[-1,1]	[-1,0[B]0,+1] A
FNSR	[0,1]	1	The model is not able to predict correctly positive outcomes based on the group. It indicates the need for more variability in samples with positive ground truth for the new group	Original Group	[-1,1]	[-1,0[B]0,+1] A

D

Cross Validation Folds

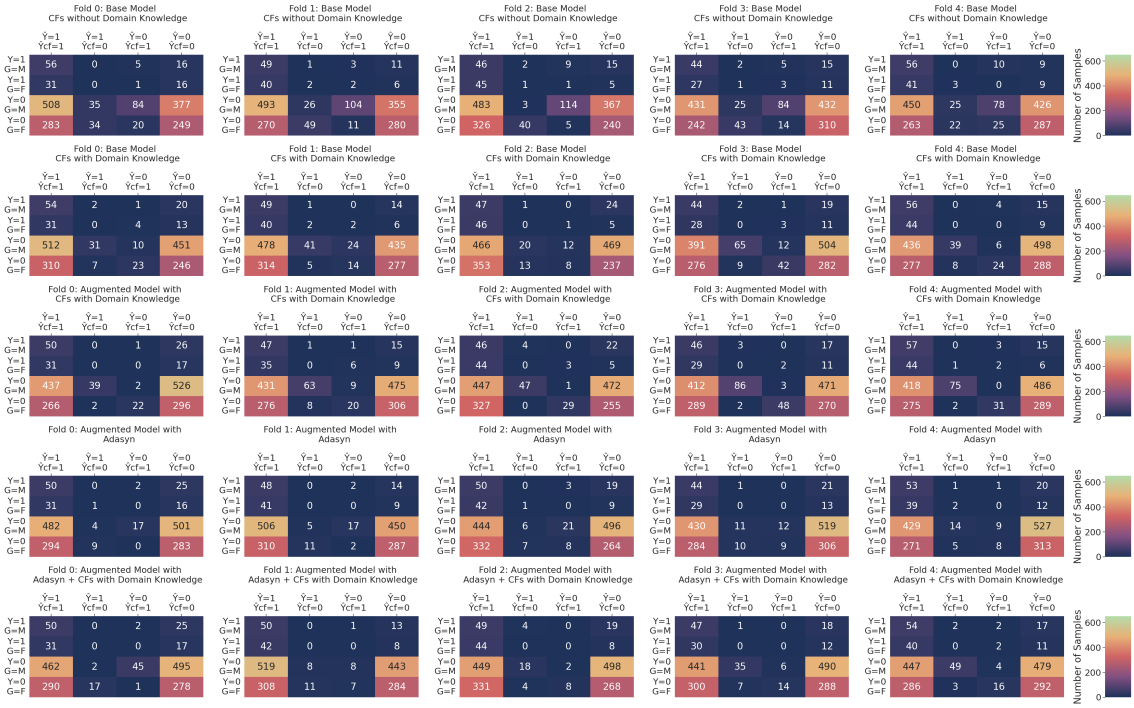


Figure D.1: ECCM for each fold of the trained model to assess bias for the sensitive feature 'sex' in the Cardio Follow.AI dataset, with respect to Section 5.1.1.

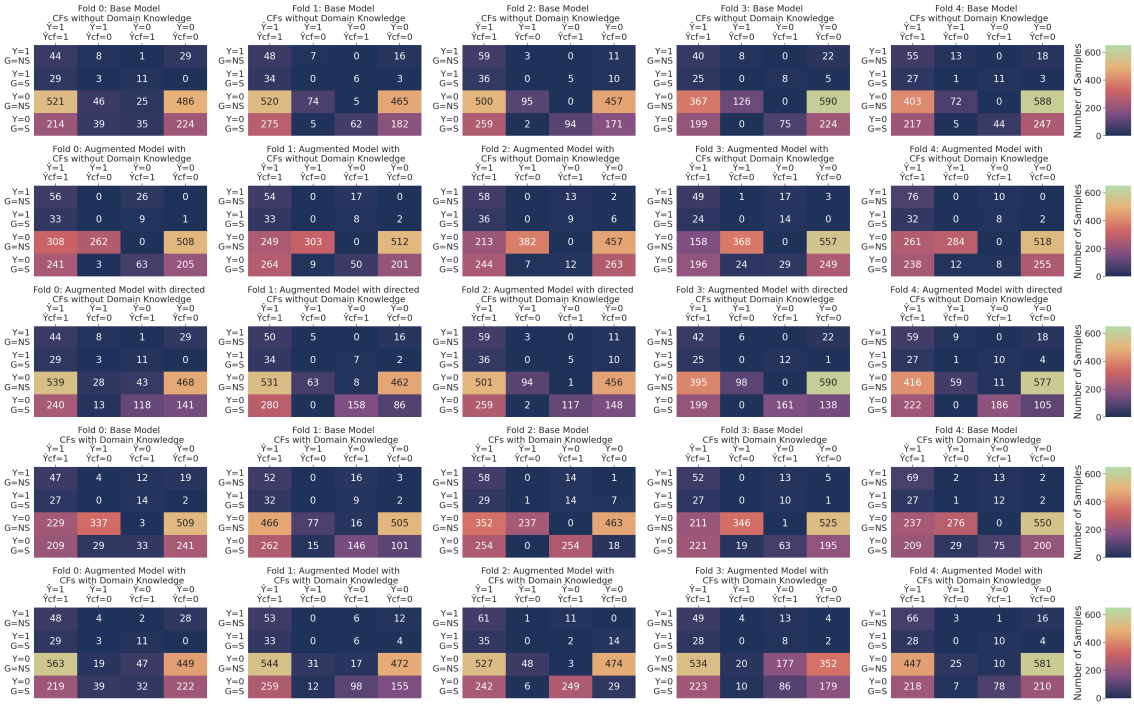


Figure D.2: ECCM for each fold of the trained model to assess bias for the sensitive feature 'smoking' in the Cardio Follow.AI dataset, with respect to Section 5.1.2.

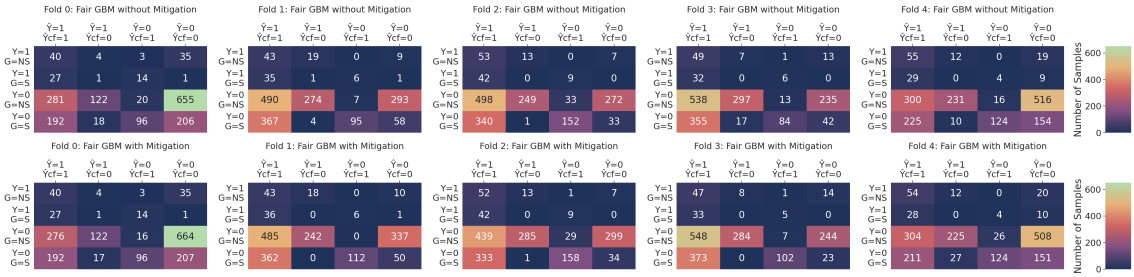


Figure D.3: ECCM for each fold of the trained model to assess bias for the sensitive feature 'smoking' in the Cardio Follow.AI dataset, with respect to Section, with respect to Section 5.1.3.

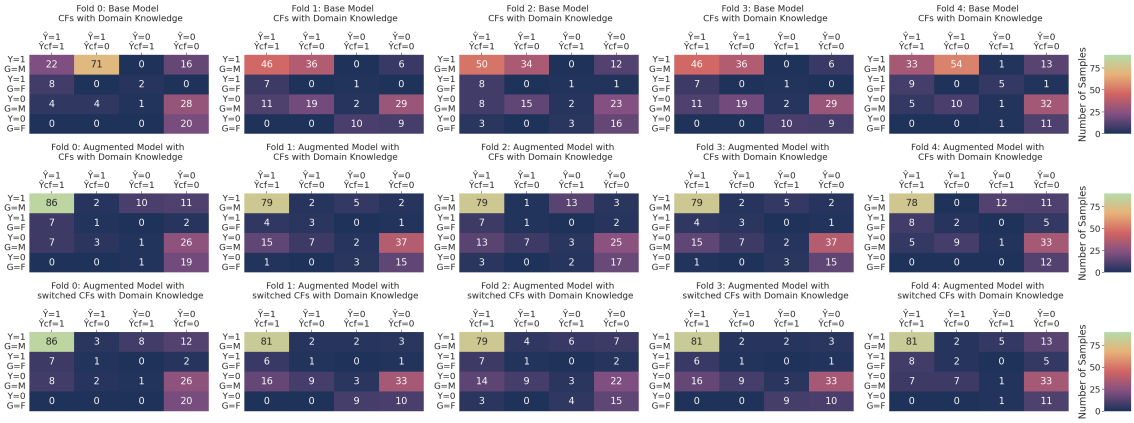
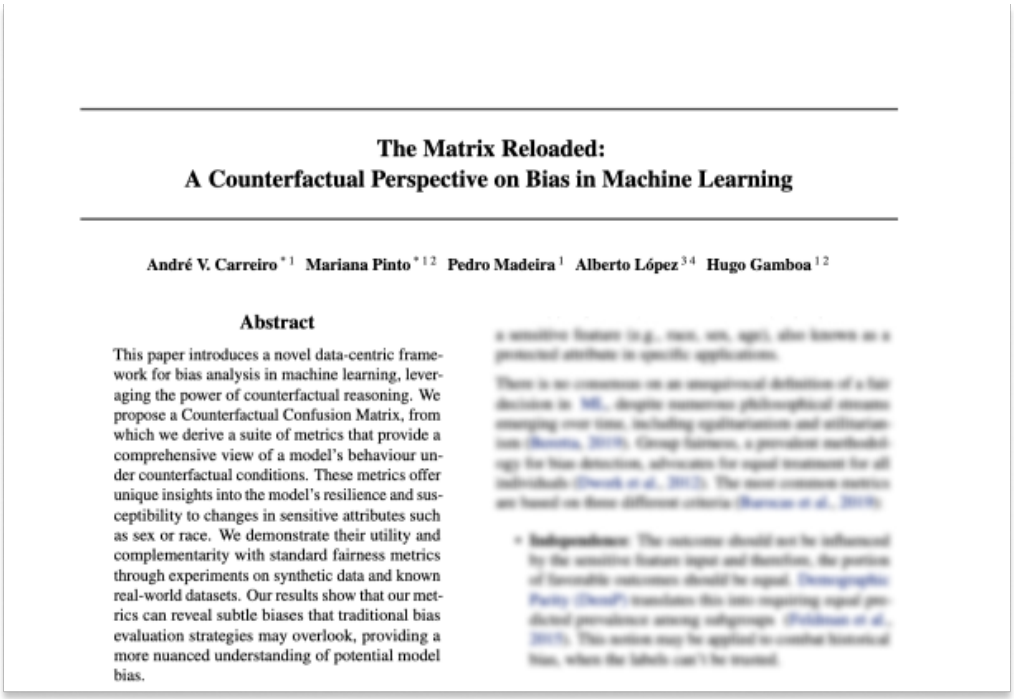


Figure D.4: ECCM for each fold of the trained model to assess bias for the sensitive feature 'sex' in the Heart Disease dataset, with respect to Section 5.2.

I

Adult Dataset

This experience is adapted from the paper accepted with displayed poster at the ICML's "DMLR Workshop: Data-centric ML Research"[15]. This paper has been entitled 'The Matrix Reloaded: A Counterfactual Perspective on Bias in Machine Learning', and its abstract can be found in Figure I.1.



The Matrix Reloaded:
A Counterfactual Perspective on Bias in Machine Learning

André V. Carreiro [*1] Mariana Pinto [*1 2] Pedro Madeira [1] Alberto López [3 4] Hugo Gamboa [1 2]

Abstract

This paper introduces a novel data-centric framework for bias analysis in machine learning, leveraging the power of counterfactual reasoning. We propose a Counterfactual Confusion Matrix, from which we derive a suite of metrics that provide a comprehensive view of a model's behaviour under counterfactual conditions. These metrics offer unique insights into the model's resilience and susceptibility to changes in sensitive attributes such as sex or race. We demonstrate their utility and complementarity with standard fairness metrics through experiments on synthetic data and known real-world datasets. Our results show that our metrics can reveal subtle biases that traditional bias evaluation strategies may overlook, providing a more nuanced understanding of potential model bias.

Figure I.1: Title Page of the paper accepted at ICML's "DMLR Workshop: Data-centric ML Research"[15].

This experiment uses the real-world datasets Adult Census Income [42], commonly used for bias analysis in ML. We evaluate bias disparities using both traditional and the newly proposed bias metrics derived from the ECCM, with *sex* as the sensitive feature.

To illustrate the proposed metrics' application, consider a scenario where a bank employs AI to evaluate loan applications using a model based on the Adult Census

Income dataset. The system decides on loan approvals with a binary outcome: grant (positive) or deny (negative) the loan, using a minimum income threshold of $ 50,000 as a primary criterion. To ensure fairness, it is crucial to prevent any sex-based discrimination in rejecting qualified applicants. Different fairness metrics could be employed although, in this case, separation metrics would likely be preferred.

We trained the model using the Light GBM algorithm [39], obtaining an ACC of 87.0% and a TPR of 66.1%. Despite the modest performance, we considered these results to suffice for demonstration purposes. Group fairness metrics yielded $9.0p.p.$ for EOpp and $7.1p.p.$ for PredEq, as detailed in Table I.1. These metrics, reflecting the differences in FNR and FPR respectively, indicate a minor bias against females, evidenced by a lower FNR for males and marginally higher FPR.

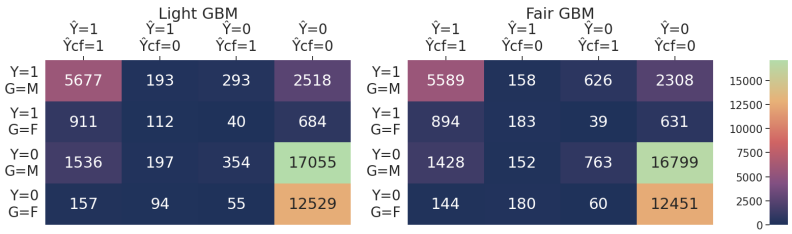


Figure I.2: ECCM for the models trained on the Adult Income Census data, considering the sensitive feature 'Sex'. Left: Light GBM. Right: Fair GBM.
Y - Ground Truth (0 for 'income < $50k$' and 1 for 'income > $50k$'); G - Group (M for 'male' and F for 'female'); $\hat{Y}$ - Prediction for the factual(original) sample; $\hat{Y}_{CF}$ - Prediction for the CF sample;

The counterfactual metrics also show slight biases ($CMCC > 90.0\%$), although in a different perspective. We highlight a higher NSR for females (16.2% vs. 5.1%), suggesting a higher likelihood of negative outcomes when flipping to males than vice versa. We note that the FPSR for females, at 37.5%, may indicate that over a third of the FPs for this subgroup become TNs after flipping to male, hinting at possible bias in this type of error. As a loose interpretation, we could infer that the model is biased towards approving loans for females who are less likely to repay, compared to males under similar conditions.

Revisiting our bank scenario under strict EOpp legislation, with a maximum threshold of 5%, we resorted to Fair GBM [21], a fairness-constrained algorithm derived from Light GBM, trained with identical hyperparameters. The resulting ECCMs are displayed in Figure I.2 and the corresponding metrics are summarized in Table I.1.

While EOpp improved, dropping from $9.0p.p.$ to $4.2p.p.$, our metrics showed increased counterfactual bias. The $NSR_{F\to M}$ rose from 16.2% to 25.9%, and the $FPSR_{F\to M}$ increased from 37.5% to 55.6%, indicating that the mitigation process worsened the counterfactual bias in favour of female instances.

Table I.1: Performance and Counterfactual Metrics for the Light GBM and for the Fair GBM model trained on the Adult Income Census data, considering the sensitive feature 'Sex'.

	Light GBM Model			Fair GBM Model		
ACC(%) ↑	**87.0**	93.3	83.7	87.0	93.2	83.8
TPR(%) ↑	**66.1**	**58.6**	**67.6**	65.4	**61.6**	**66.2**
TNR(%) ↑	**93.8**	98.0	90.9	94.0	97.5	91.7
CMCC(%) ↑	**90.5**	86.6	90.7	85.1	80.6	85.4
PSR(%) ↓	2.2	0.7	3.2	4.4	0.8	6.8
NSR(%) ↓	6.7	**16.2**	5.1	7.7	**25.9**	4.2
TNSR(%) ↓	1.4	0.4	2.0	2.7	0.5	4.4
FPSR(%) ↓	14.7	**37.5**	11.4	17.4	**55.6**	9.6
RMSCD ↓	0.054	0.059	0.052	0.073	0.071	0.074
JSCD ↓	0.079	0.130	0.062	0.097	0.148	0.083
	Total	$F_{\rightarrow}M$	$M_{\rightarrow}F$	**Total**	$F_{\rightarrow}M$	$M_{\rightarrow}F$